ESCUELA

Universidad Rey Juan Carlos

Titulo del Trabajo Fin de Máster

Trabajo Fin de Máster

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

AUTOR: Abel de Andrés Gómez

TUTOR/ES: Nombre y Apellidos y

Nombre y Apellidos

2019

AGRADECIMIENTOS

Agrademos a...

RESUMEN

Extensión máxima de una página

SUMMARY

Extensión máxima de una página

Índice

Índice de figuras

Índice de cuadros

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.

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.
    - 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 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.

2. JUSTIFICACIÓN TEÓRICA

El la actualidad existen numerosos informes acerca del uso de la ciencia de datos y sus tecnicas en el ambito educativo.

En el artículo de prensa de Fernandes [1], 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.

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 el articulo anterior 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 genero.
7. El entorno en el aula.

En esta investigacion se utiliza la metodologia CRISP-DM (del ingles Cross Industry Standard Process for Data Mining) que es una metodologia frecuente en el desarrollo de proyectos de Data Mining.

Para realizar la parte de prediccion, utiliza las variables anteriormente comentadas, incluidas en dos conjuntos de datos. En el primer conjunto de datos, se almacenan los datos obtenidos antes de comenzar el comienzo del año escolar. El segundo conjunto de datos incluye las variables del primer conjunto de datos y alguna nueva que se ha obtenido despues del segundo mes del año escolar. Siendo algunas de estas variables nuevas las asignaturas, las notas y las ausencias. Estas dos ultimas variables son las que mayor importancia tienen en la revelacion de los resultados academicos finales.

3. 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. [3]. 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. 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). [6]
- 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.[6]

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.

Existe una serie de procesos o metodologías para llevar a cabo proyectos de ciencia de datos. Entre ellos podemos destacar los siguientes:

- Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases - KDD):

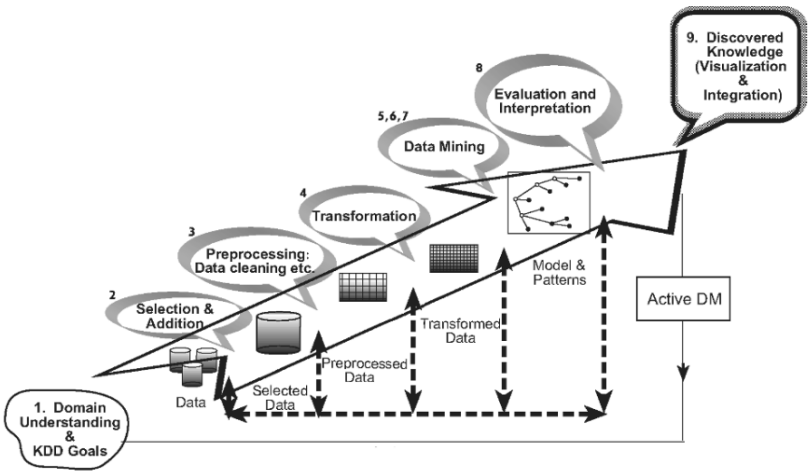

Fig. 1: asdad KDD

- SEMMA

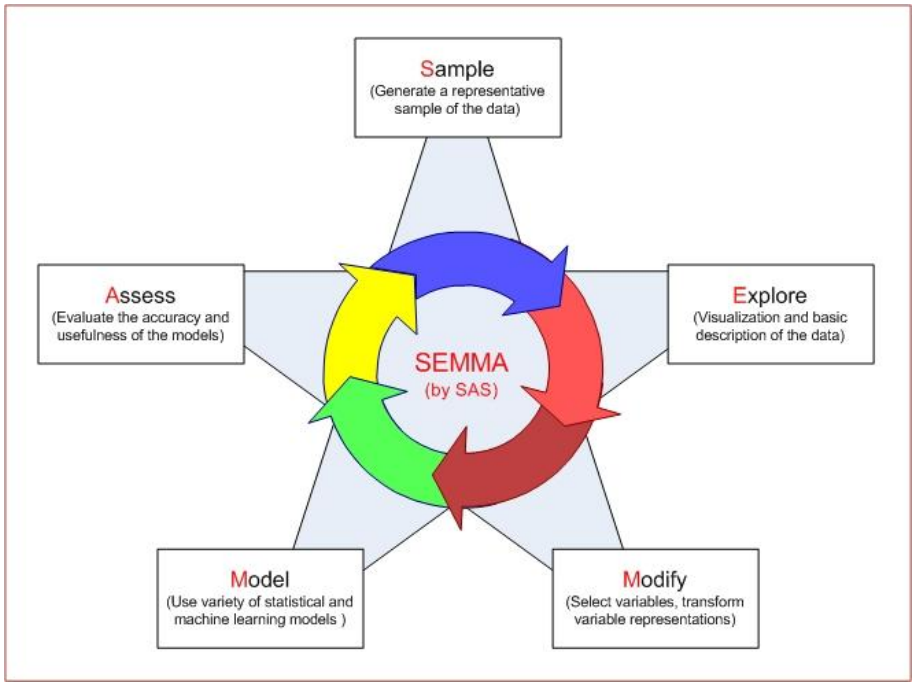

Fig. 2: Esquema SEMMA

- CRISP-DM

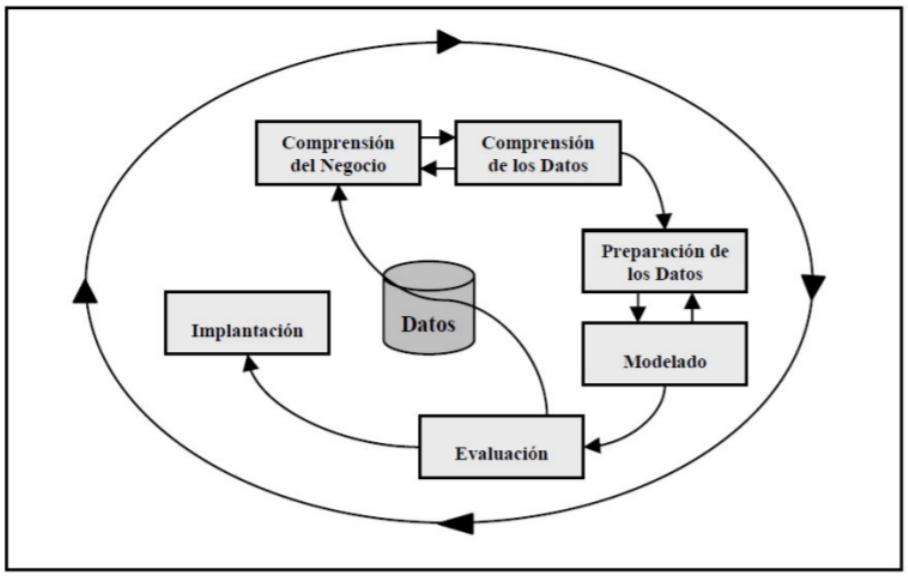

Fig. 3: Esquema SEMMA

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.

4.1. Proceso de Extracción, Extracción 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:

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.

4.2. 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.[4]. 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.[4]

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. [5]

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. [5]

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

4.2.1. Aprendizaje No Supervisado

1. Algoritmos de Clustering
2. Análisis de Componentes Principales
3. Descomposición en valores singulares
4. Analisis de componentes independientes
5. Stepwise Regression

4.3. 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. [3].

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 %. [2] 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. 4: Funcionamiento Arboles Decisión

2. Análisis de Componentes Principales

prueba

3. Descomposición en valores singulares
4. Analisis de componentes independientes
5. Stepwise Regression

4.3.1. Aprendizaje No Supervisado

1. Algoritmos de Clustering
2. Análisis de Componentes Principales
3. Descomposición en valores singulares
4. Analisis de componentes independientes
5. Stepwise Regression

5. ANÁLISIS DE RESULTADOS

6. CONCLUSIONES

7. SOBRE LAS REFERENCIAS

La bibliografaa o referencias deben aparecer siempre al final de la tesis, incluso en aquellos casos donde se hayan utilizado notas finales. La bibliografaa debe incluir los materiales utilizados, incluida la edician, para que la cita pueda ser facilmente verificada.

Citar dentro del texto:

Las fuentes consultadas se describen brevemente dentro del texto y estas citas cortas se amplaan en una lista de referencias final, en la que se ofrece la informacian bibliografica completa.

La cita dentro del texto es una referencia corta que permite identificar la publicacian de dande se ha extraado una frase o parafraseado una idea, e indica la localizacian precisa dentro de la publicacian fuente. Esta cita informa del apellido del autor, la fecha de publicacian y la pagina (o paginas) y se redacta de la forma que puede verse a travas de los siguientes ejemplos:

Cuando se citan las palabras exactas del autor deben presentarse entre comillas e indicarse, tras el apellido del autor y, entre parantesis, la fecha de publicacian de la obra citada, seguida de la/s pagina/s.

Si lo que se reproduce es la idea de un autor (no sus palabras exactas) no se pondran comillas y se indicara, entre parantesis, el apellido del autor seguido de la fecha de publicacian de la obra a la que se refiere.

No se puede eliminar una parte del texto citado sin seaalarse; debe indicarse siempre con puntos suspensivos entre corchetes [...]

Ejemplos de como citar una referencia en el texto son los siguientes [**?**] o [**?**, **?**, **?**].

Camo ordenar las referencias:

1. Las referencias bibliograficas deben presentarse ordenadas alfabaticamente por el apellido del autor, o del primer autor en caso de que sean varios.

2. Si un autor tiene varias obras se ordenaran por orden de aparician.

3. Si de un mismo autor existen varias referencias de un mismo aao se especificaran los aaos seguidos de una letra minascula y se ordenaran alfabaticamente.

4. Si son trabajos de un autor en colaboracian con otros autores, el orden vendra indicado por el apellido del segundo autor, independientemente del aao de publicacian. Las publicaciones individuales se colocan antes de las obras en colaboracian.

Camo citar un artaculo de revista

Un artaculo de revista, siguiendo las normas de la APA, se cita de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre o nombres. (Aao de publicacian). Tatulo del artaculo. Tatulo de la revista en cursiva, volumen de la

revista (namero del fascaculo entre parantesis), primera pagina- altima pagina del artaculo.

Camo citar una monografaa/libro

Las monografaas, siguiendo las normas de la APA, se citan de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre. (Aao de publicacian). Tatulo del libro en cursiva. Lugar de publicacian: Editorial. Opcionalmente podremos poner la mencian de edician, que ira entre parantesis a continuacian del tatulo; y, si fuera el caso el volumen que ira en cursiva.

Camo citar un capatulo de un libro

Los capatulos de los libros se citan de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre o nombres. (Aao). Tatulo del capatulo. En A. A. Apellido(s) Editor A, B. B. Apellido(s) Editor B, y C. Apellido(s) Editor C (Eds. o Comps. etc.), Tatulo del libro en cursiva (pp. xxx-xxx). Lugar de publicacian: Editorial.

Camo citar un acta de un congreso

Apellido(s), Iniciales del nombre o nombres. (Aao). Tatulo del trabajo. En A. A. Apellido(s) Editor A, B. B. Apellido(s) Editor B, y C. Apellido(s) Editor C (Eds. o Comps. etc.), Nombre de los proceedings en cursiva (pp. xxx-xxx). Lugar de publicacian: Editorial.

Camo citar tesis doctorales, trabajos fin de master o proyectos fin de carrera

Apellido(s), Nombre. (Aao). Tatulo de la obra en cursiva. (Tesis doctoral). Institucian a acadamica en la que se presenta. Lugar.

Camo citar un recurso de Internet

Los recursos disponibles en Internet pueden presentar una tipologaa muy variada: revistas, monografaas, portales, bases de datos... Por ello, es muy difacil dar una pauta general que sirva para cualquier tipo de recurso. Como manimo una referencia de Internet debe tener los siguientes datos:

1. Tatulo y autores del documento.
2. Fecha en que se consulta el documento.
3. Direccian (URL auniform resource locatora)

Veamos, a travas de distintos ejemplos, camo se citan especaficamente algunos tipos de recursos electranicos.

Monografaas: Se emplea la misma forma de cita que para las monografaas en versian impresa. Debe agregar la URL y la fecha en que se consulta el documento

Artaculos de revistas: Se emplea la misma forma de cita que para los artaculos de revista en versian impresa. Debe agregar la URL y la fecha en que se consulta el documento.

Artaculos de revistas electranicas que se encuentran en una base de datos: Se emplea la misma forma de cita que para los artaculos de revista en versian impresa, pero debe aaadirse el nombre de la base datos, la fecha en que se consulta el documento.

ANEXOS

BIBLIOGRAFÍA

Referencias

[1] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2018, 7 febrero). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, pp. 1–9.

[2] Manguart, A. (2017, 13 junio). Conceptos básicos del aprendizaje supervisado (para personas no técnicas) [Publicación en un blog]. Recuperado 15 enero, 2019, de https://medium.com/@manguart/machine-learning-conceptos-básicos-del-aprendizaje-supervisado-para-personas-no-técnicas-142bbb222140

[3] Marín, J. L. (2018, 5 abril). *Ciencia de datos, machine learning y deep learning* (Comunicado de prensa). Recuperado 17 enero, 2019, de https://datos.gob.es/es/noticia/ciencia-de-datos-machine-learning-y-deep-learning

[4] Marín, J. M. (s.f.). Estadística Descriptiva Univariante [Tema Universitario]. Recuperado 17 enero, 2019, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema2am.pdf

[5] Orellana, L. (2001). Introducción. In L. Orellana (Ed.), Estadística Descriptiva (pp. 1–64). Recuperado de http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf

[6] Recuero, P. (2017, 16 noviembre). Los 2 tipos de aprendizaje en Machine Learning: supervisado y no supervisado [Publicación en un blog]. Recuperado 17 enero, 2019, de https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html