# Primera línea del título del trabajo Segunda línea del título

### Trabajo Fin de Master

Autor:

# Autor del trabajo

Tutor:

Tutor 1 del trabajo Tutor 2 del trabajo

# Master en Matemáticas





Junio, 2015 Universidad de AlmerÃa

# Índice general

1	1.1.	dudas	
2	Intro	oducción 3	
3		ción del algoritmo5Extracción de datos	
4	Instr	rucciones 7	
	4.1.	Texto matemático	
		Sin referencias posteriores	
		Con referencias posteriores	
	4.2.	Entornos de definiciones, teoremas,	
		Entornos predefinidos	
		Nuevos entornos	
	4.3.	Gráficos	
	4.4.	Cuadros	
	4.5.	Referencias bibliográficas	(
5	fgf	1	
	5.1.	tonto	1
		el que lo lea	
Ri	hlingr	rafía 1	7

# Abstract in English

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet

vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

# Resumen en español

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vi-

tae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

1

# **Objetivos**

Los objetivos que nos proponemos obtener en este trabajo son los siguientes:

- 1. Creación de un algoritmo, haciendo uso de un entorno de desarrollo Java como es Intellij IDEA, el cual nos permitirá la creación de un nuevo modelo mediante redes bayesianas y con la que podremos asignar nuevos valores a los datos faltantes que encontraremos en nuestro dataset.
- 2. Comparar los resultados obtenidos mediante un grupo de control que se ha creado previamente.
- 3. Estudiar diferentes modelos para una mejora de los resultados.

#### 1.1 dudas

¿Cual es nuestro problema? ¿Como podemos solucionarlo? ¿A que solución hemos llegado?

- Punto tres objetivo es por si sale mal el modelo y dar posibles soluciones (merece la pena?)

¿Se intercala parte teorica y parte práctica?

# Introducción

En esta sección, trataremos de explicar, de manera escueta, como está la situación actual de una de las ramas más influyentes de las matemáticas como es el "análisis de Ddatos".

El análisis de datos es un proceso en el cual se examinan un conjunto de datos para extraer conclusiones sobre la información que contienen dichos datos. Con el paso de los años, estas conclusiones se han ido mejorando con el avance de las nuevas tecnologías. Hoy en día las técnicas y tecnologías de análisis se aplican considerablemente en el mundo comercial ya que gracias a ellas se pueden tomar mejores decisiones tanto en el ámbito comercial como en el ámbito social. Estas nuevas técnicas son llevadas a cabo por "Data Scientist". Un "Data Scientist" es aquella persona que posee las habilidades y técnicas necesarias para resolver problemas complejos y además, tenga las aptitudes necesarias para poder gestionarla. De este modo, se han unificado dos mundos que hasta este momento se encontraban separados como es el de la gestión de datos y el análisis de datos.

En este trabajo, nos introduciremos en la piel de un "Data Scientist" con lo que para ello, deberemos de seguir los siguientes pasos:

- 1. **Extracción de datos**: Seremos capaces de poder extraer la información independientemente del origen (*csv* , *txt* , *arff* . . . ) y del volumen de esta (*Big Datat* o *Small Data*)
- 2. **Limpieza de datos**: Tendremos que ser capaces de eliminar toda aquella información que nos puedan ocasionar un perjuicio en el resultado final ó una distorsión de la información.
- 3. **Procesamiento de datos**: Tendremos que ser capaces de poder aplicar diferentes métodos estadísticos para obtener una respuesta lo más óptima posible. Dentro de estos métodos podremos hacer uso de los modeles de regresión, pruebas de hipótesis, inferencia estadística...
- 4. **Diseño**: Tendremos que ser capaces de poder diseñar o experimentar nuevos test en caso de que lo veamos necesario.
- 5. **Visualización**: Tendremos que ser capaces de poder mostrar esta información lo más nítida e inteligible posible para que pueda estar al alcance del mayor número posible de sujetos.

Comenzaremos partiendo de una base de datos la cual contiene información acerca de las viviendas (tipo de vivienda, número de metros cuadrados construidos, número de baños...) que posee en propiedad el Grupo Cooperativo Cajamar (GCC) y que es comercializado a través de la operadora Haya Real Estate.

# Creación del algoritmo

En este capítulo se expondrá de forma detallada el proceso seguido para poder obtener el principal objetivo del presente trabajo.

### 3.1 Pre-procesamiento

Como dijo *Dhanurjay Patil*, ex-anlista de datos de Google, el 80% de un proyecto consiste en la limpieza de datos. Es por ello, que gran parte de este trabajo se lo dedicaremos al pre-procesamiento de los datos.

Como hemos dicho al final del capitulo anterior, hemos seleccionado una base de datos en la que encontramos múltiples variables acerca de las características y morfología de las viviendas que tienen en propiedad el GCC.

Haciendo uso de una sentencia SQL realizaremos una llamada a la tabla que contiene toda la información de la que disponemos de los inmuebles. En total, encontramos 237 variables de las cuales nos quedaremos con 7 variables. Estas variables son:

- 1. Tipo vivienda.
- 2. Número de baños.
- 3. Número de habitaciones.
- 4. Número de ascensore.
- 5. Número de metros cuadrados construidos.
- 6. Número de metros cuadrados útiles.
- 7. Importe de tasación.

El motivo de la selección de esta variables es debido a que presenta un alto grado de datos faltantes y un alto grado de datos erróneos por culpa de una mala introducción por parte del operario. En total, nuestra base de datos constan de 27.945 registros. A continuación procederemos a guardar toda esta información en un libro de Excel (.csv) para así poder realizar una limpieza de los datos.

#### Tipo vivienda

Según nuestro dataset, podemos encontrar los siguientes tipos de viviendas.

### Instrucciones

#### 4.1 Texto matemático

#### Sin referencias posteriores

Para escribir texto dentro de una línea lo hacemos utilizando el símbolo del dólar.

Por ejemplo, la ecuación  $x^2 - 2x + 1 = 0$  tiene una raiz doble en x = 1. Las fracciones se ajustan al tamaño de la línea como, por ejemplo,  $\frac{1}{2}$ .

Sin embargo, si queremos que la ecuación aparezca en la línea siguiente y centrada, pero no nos interesa que aparezca numerada para posteriores referencias lo hacemos asi:

La ecuación 
$$2x^2 - 3x + 1 = 0$$
 tiene como raíces  $x = 1$ ,  $x = \frac{1}{2}$ .

Otro par de ecuaciones:

$$\frac{(\mu-1)^n S_n(x)}{\Gamma(n-x)} = \sum_{j=0}^n \frac{(\mu-1)^j b_{n,j}}{n^j} \frac{(\mu-1)^{n-j} n^j Q_{n-j}(x)}{\Gamma(n-x)}.$$

$$\eta = \frac{1}{2} \left( 1 + \frac{\lambda (1 - \mu)^2}{\mu} + \mu + (1 - \mu) \sqrt{\left( 1 + \frac{\lambda (1 - \mu)}{\mu} \right)^2 + 4\lambda} \right).$$

### Con referencias posteriores

Para poder referenciar una expresión matemática en el texto, se utiliza el entorno equation de la siguiente forma

$$x^2 + 2x + 1 = 0. (4.1)$$

La ecuación (4.1) tiene una raíz doble en el punto x = -1.

Existen varios entornos para poner varias ecuaciones alineadas. Un documento interesante se puede ver en el documento elaborado por Herbert Voß <sup>1</sup>.

Uno de los más usuales es el entorno align, que produce

$$x + 2y + 4z = 3, (4.2)$$

$$x - 5y - 2z = 0, (4.3)$$

$$x + 3y + 7z = 2.$$

Como se puede observar, la tercera ecuación no está numerada. Si se quiere que no se numere ninguna, se utiliza el entorno align\*.

Para referenciar una de esas ecuaciones, se hace de la misma forma. Ecuación (4.3) o ecuación 4.3 —si no se quiere que aparezca entre paréntesis—.

<sup>&</sup>lt;sup>1</sup>http://osl.ugr.es/CTAN/info/math/voss/mathmode/Mathmode.pdf

### 4.2 Entornos de definiciones, teoremas,...

#### Entornos predefinidos

Para escribir una definición, teorema, corolario,... y posteriormente referenciarlo lo haremos de la siguiente forma:

**Definición 4.1.** Sea X una muestra aleatoria extraída de una población modelada por la familia de distribuciones  $\mathcal{F}_{\theta}$ , entonces considereramos que un **estadístico** T(X) es una función medible de las variables que forman la muestra.

La definición 4.1 se referencia en el texto tal y como vemos en este ejemplo.

De la misma forma se han definido los entornos teorema, lema, proposicion, corolario (jerárquicamente dependiente de un teorema) y corolariop (jerárquicamente dependiente de una proposición), así como se ha definido un entorno demo para las demostraciones.

Veamos un ejemplo:

**Teorema 4.1** (Lehmann y Scheffé). Bajo las condiciones del teorema de Blackwell-Rao, si T es suficiente y completo, entonces  $\mathsf{E}(\hat{\mathbf{g}} \mid T)$  es el único UMVUE.

#### Demostración:

Sean  $\hat{g}_1,\hat{g}_2\in\mathcal{U}_{g(\theta)}$ , entonces según el Teorema de Blackwell-Rao sabemos que  $\mathrm{E}(\hat{g}_1\mid T)$  y  $\mathrm{E}(\hat{g}_2\mid T)$  también pertenecen a la clase  $\mathcal{U}_{g(\theta)}$ . Por la linealidad de la esperanza, tenemos que

$$\mathsf{E}_{\theta}(\mathsf{E}(\hat{g}_1 \mid T) - \mathsf{E}(\hat{g}_2 \mid T)) = 0, \ \forall \ \theta \in \Theta.$$

Puesto que T es completo, cualquier función de T con esperanza nula tiene que ser necesariamente nula, por lo que se verifica que

$$E(\hat{g}_1 \mid T) - E(\hat{g}_2 \mid T) = 0$$
,

y, por lo tanto,  $\mathrm{E}(\hat{g}_1 \mid T) = \mathrm{E}(\hat{g}_2 \mid T)$ , es decir, la esperanza condicionada dado T de cualquier estimador  $\hat{g} \in \mathcal{U}_{g(\theta)}$  es siempre la misma y, en consecuencia que  $\mathrm{E}(\hat{g} \mid T)$  es el UMVUE ya que, si aplicamos el teorema de Blackwell-Rao, tenemos que cualesquiera que sean los estadísticos insesgados de los que partimos, siempre tenemos que aquel que reduce su varianza siempre es el mismo, por lo que éste necesariamente tiene que ser el único UMVUE.

**Corolario 4.1.1.** Sea una muestra  $X_1, X_2, ..., X_n$  procedente de una familia de distribuciones  $\{f(x;\theta), \theta \in \Theta\}$  del tipo exponencial uniparamétrica dada por

$$f(\mathbf{x}; \theta) = h(\mathbf{x}) \exp\{\psi(\theta) T(\mathbf{x}) + D(\theta)\}.$$

Si existe un estimador  $\hat{g} \in \mathcal{U}(\theta)$ , entonces existe un único UMVUE dado por  $E_{\theta}(\hat{g} \mid T)$ .

#### Demostración:

Es evidente, ya que entonces el estadístico  $T(X) = T(X_1, X_2, ..., X_n)$  es suficiente y completo, y basta aplicar el teorema de Lehmann-Scheffé.

8

#### Nuevos entornos

Si se quiere definir un entorno nuevo hay que incluir en el fichero de estilo \newtheorem{nombre\_del\_entorno}{Leyenda\_que\_aparece}[De\_quien\_depende]

### 4.3 Gráficos

Si compilamos con PDFLATEX podremos incluir ficheros gráficos en formato jpg, png o pdf. ¡Ojo!, no se pueden incluir directamente ficheros PostScript.

Para referenciarlos adecuadamente utilizaremos el entorno figure.



Figura 4.1: Logotipo de la Facultad de Ciencias Experimentales

De esta forma podremos referenciar la Figura 4.1 en cualquier parte del texto. Es recomendable hacerlo así pues LATEX colocará la figura donde le quepa —siguiendo un orden de preferencia que se puede cambiar— y se debe evitar expresiones como «según la figura siguiente» o similares.

#### 4.4 Cuadros

Al igual que las figuras, utilizaremos el entorno flotante table para incluir los cuadros —nombre correcto en castellano en lugar del habitualmente visto tabla—. Si se quiere cambiar la palabra Cuadro por Tabla se escribe en el preámbulo del documento:

\renewcommand{\tablename}{Tabla}

Un ejemplo es el siguiente:

$x_i$	$n_i$	$f_i$	$N_i$
0	10		
1	20		
2	5		
3	15		
	50		

Cuadro 4.1: Frecuencias observadas

Rellena los huecos que faltan en el Cuadro 4.1.

Equipo		
Portero	De Gea	
	Mario	
Defensas	Etxeita	
Defelisas	Nacho	
	Azpilicueta	
	Thiago	
Controcampistas	San José	
Centrocampistas	Isco	
	Cesc	
Delanteros	Alcácer	
Detaille108	Nolito	

Cuadro 4.2: Alineación de España

Un cuadro más elaborado se puede construir utilizando el paquete multirow. En el Cuadro 4.2 podemos ver el equipo inicial que jugó en el partido Ucrania-España clasificatorio para la Eurocopa y que acabó con el resultado de 0 a 1.

### 4.5 Referencias bibliográficas

Para ir citando las referencias a la bibliografía que se expone al final de texto se utiliza el comando cite. Así se puede encontrar toda las información sobre TEX en [4]. La bibliografía se ordena por orden alfabético y se cita como sigue:

- Los libros con el formato de [1].
- Los artículos como aparece en [3].
- Los artículos en preimpresión como [2].
- Las páginas web tal y como se puede ver en [4].

# Bibliografía

- [1] M. Abramowitz, I.A. Stegun, *Pocketbook of Mathematical Functions*, Verlag Harri Deutsch, 1984.
- [2] D. Dominici, Mehler-Heine type formulas for Charlier and Meixner polynomials, ar-Xiv:1406.6193v1.
- [3] S.F. Khwaja, A.B. Olde Daalhuis, *Uniform asymptotic approximations for the Meixner–Sobolev polynomials*, Anal. Appl., **10** (2012), 345–361.
- [4] Página web de los usuarios de TEX: http://www.tug.org.