Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Análisis y predicción de datos de entrada en urgencias relativos a problemas respiratorios en la ciudad de Valencia

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: Javier Castaño Sánchez

Tutor: Cèsar Ferri Ramírez

2015-2016

Resumen

En el sector de la Sanidad pública, los recursos que se destinan deben ser gestionados de la manera más eficaz posible. Si bien, como cualquier otro servicio público, en el de la Sanidad interviene un factor sumamente importante, como es: la Salud.

En las urgencias hospitalarias el tiempo de respuesta y la utilización de estos recursos (materiales, personal sanitario, administrativos) pueden contribuir significativamente en la mejora y atención de los pacientes ingresados.

Una información importante en las urgencias, es conocer el número de ingresos que se van a producir a fin de poder preparar y gestionar los recursos necesarios para dar la atención necesaria. Las técnicas de Machine learning y minería de datos, junto con el uso de las bases de datos permiten reconocer patrones y aplicar métodos estadísticos para realizar predicciones que ayudan a aportar nueva información para anticipar los recursos necesarios y tomar las medidas adecuadas.

El estudio realizado en este trabajo final de master se centra en la predicción de entradas en urgencias de pacientes con problemas respiratorios o asmáticos causados por factores medioambientales, climatológicos y/o externos del entorno (contaminación).

Contar con predicciones precisas sobre el volumen de entrada de casos de urgencia, ayudaría a los servicios sanitarios a administrar mejor sus recursos, y de esta manera, mejorar la atención a los pacientes con estos cuadros diagnósticos.

Palabras clave: asma, urgencias, aprendizaje automático, minería de datos, bases de datos, predicción.

Abstract

In the healt public systems, resources allocated should be managed as efficiently as possible. Although, as any other public service, within the health public there is a very important factor: Health.

In emergencies, an information important to know in advance is the estimated number of income patients that is going to arrive to hospitals. This predictions are useful for the management of hospitals are ready to prepare and assign resources necessary to give the required attention .

Techniques of machine learning and mining of data, together with the use of them databases allow recognize patterns and apply methods statistical for perform predictions that help to provide new information to anticipate them resources necessary and take the measures appropriate.

In this final work of master focuses on the prediction of entries in emergencies patients with asthma or respiratory problems caused by environmental, climatic or external factors of the environment (pollution).

This prediction, help the hospital emergency to have the means and resources necessary to assist patients with these diagnoses pictures with a higher quality and attention.

Keywords: Asthma, emergency, machine learning, data mining, databases, prediction.

Índice

1. Introducción

"Ya no estamos en la era de la información. Estamos en la era de la gestión de la información." **(Chris Hardwick, actor).**

1.1 OBJETIVO

El objetivo de ese trabajo es determinar qué cantidad de pacientes recibirán los hospitales con episodios de urgencias relacionados con asma mediante el uso de técnicas de minería de datos y machine learning, dentro de los cuales se encuentran en el paradigma del Big Data.
Dicho objetivo conlleva desarrollar el conocimiento sobre minería de datos y el tratamiento de los mismos con la ayuda del lenguaje de programación estadístico R. Mediante la exploración de los datos y un conjunto de técnicas podemos llevar un análisis que nos permitirá conocer la información y aplicar técnicas de machine learning sobre los modelos de datos obtenidos para obtener una predicción sobre la variable u objetivo deseado.

1.2 MOTIVACIÓN

Dada la diversidad de información y datos abiertos que disponemos, nos centramos en la utilidad que podemos obtener de los mismos en relación donde habitualmente nos encontramos, Valencia. Una ciudad donde cada vez más se promueve el uso de los datos abiertos y el desarrollo del concepto de "Smart City"[1].

Sin duda en cualquier ciudad el sector sanitario es fundamental en la mejora y calidad de vida del ciudadano, es por ello, que cada vez se estudian nuevos procedimientos y mejoras, para tratar de que la calidad de los servicios públicos de este sector cumpla con las expectativas y necesidades demandadas por los usuarios y profesionales.
El estudio de este trabajo, se centra aplicar técnicas de machine learning y minería de datos para realizar un modelo que ayude a predecir el número de entradas en urgencias de pacientes con

[1] **Smart City**: Traducido como *(Ciudad Inteligente)*, se refiere a un tipo de desarrollo urbano basado en la sostenibilidad que es capaz de responder adecuadamente a las necesidades básicas de instituciones, empresas, y de los propios habitantes, tanto en el plano económico, como en los aspectos operativos, sociales y ambientales.

problemas asmáticos y/o respiratorios causados por los factores medioambientales como el polen, o contaminantes con su relación con el clima (temperaturas) que se producen en la ciudad.

La minería de datos está siendo cada vez más relevante en las gestiones y recursos en el área de la Sanidad. La utilidad que tiene poder gestionar estos recursos es sumamente importante y debido a ello se han realizado diversos estudios para poder facilitar estas tareas.

Podemos encontrar diversos estudios, como el que se encuentra en libro publicado: "**HealthCare Data Mining: Predicting Hospital Length of Stay**" (Ali Azari, 2012) que mediante técnicas de minería de datos, evalúan y clasifican diferentes grupos de datos utilizando varios clasificadores para predecir la estancia de los pacientes en urgencias.

Otro estudio más que se encuentra en el ámbito de la Sanidad es: "**Application of Data Mining Techniques to Healthcare Data**" (Obenshain, 2004), donde aplica la minería de datos a la detección temprana de infecciones nosocomiales[2] centrándose en la investigación de pacientes de alto riesgo detectando e identificando nuevos patrones de infección.

El fin de ese trabajo es determinar qué cantidad de pacientes recibirán los hospitales con la patología de asma o dificultades respiratorias en emergencias de manera que exista personal médico especializado para atender de manera rápida y adecuada los ingresos que se producen.

[2] **Infección nosocomial**: En el ámbito médico se denomina infección nosocomial (Del latín nosocomĭum, hospital de enfermos) o infección intrahospitalaria a la infección contraída por pacientes ingresados en un recinto de atención a la salud (no sólo hospitales).

1.3 Concepto de Big Data

Big data es la tendencia tecnológica que viene en aumento en los últimos años. Esta tendencia engloba el big data como un concepto que permite a través de grandes volúmenes de datos realizar predicciones y proporcionar una información certera. Detrás de estas predicciones se esconden técnicas estadísticas que trabajan con la cantidad de datos, siendo mayor el volumen de estos, mejores resultados y más precisos son los valores que se predicen.

Hoy en día, la cantidad de información crece exponencialmente, siendo así difícil tratar con ella cuando no está organizada (estructurada) y es por ello que podemos encontrarnos con diferentes tipos de datos, estructurados, semiestructurados y no estructurados. Todos estos datos abundan en nuestro entorno, de manera que poder tratar con ellos y obtener la información que se requiere puede ser una valiosa herramienta.

Tratar con grandes volúmenes de datos y con una alta heterogeneidad en sus tipos de datos no es trivial y es por ello que es necesario utilizar avanzadas herramientas informáticas para tratar con la información, almacenarla, y aplicar minería de datos junto con técnicas de machine learning para poder conseguir obtener la información que se requiere, esto es lo que resume el concepto de Big Data.

2. Estado del arte

2.1 Extracción Automática de conocimiento desde Bases de Datos

"Scientia potentia est"
(**Thomas Hobbes, filósofo**)

La frase: ***"El conocimiento es poder"*** es un dicho conocido o popular. Esta frase, donde se les atribuye a diferentes autores, y viene a significar que cuanto más conocimiento se posee mayor es el poder que se puede alcanzar sobre algo o alguien.

El proceso de extracción automática de conocimiento desde bases de datos (**KDD**)[3] tiene como fin alcanzar ese conocimiento, donde como se indica, se extrae de una o varias bases de datos**.** El KDD consta de una secuencia de fases o etapas donde al finalizar se logra el objetivo perseguido, obtener el conocimiento o información que se desea.

El proceso tiene como una de sus fases más características la minería de datos (data mining). Data mining es un campo multidisciplinar englobado dentro de la ciencia de la computación que busca patrones en grandes cantidades de datos a través de métodos como la inteligencia artificial, machine learning, estadística y sistemas de bases de datos. Gracias a esta fase, podemos "minar" los datos, creando nuevo conocimiento que sirva para el desarrollo de nuevas técnicas, como el Machine Learning, donde una máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema.

Dentro de todo el contexto de la extracción automática del conocimiento, la disciplina que permite poder realizar análisis sobre las variables, varianza, aplicar modelos de regresión, etc. y poder llegar a conclusiones sobre la información obtenida o desarrollada es la "**Estadística**". Esta disciplina hoy en día es aplicada fácilmente mediante herramientas de software y lenguajes de

[3] **KDD:** Proceso de extracción del conocimiento desde bases de datos, conocido por sus siglas en inglés *(Knowledge Discovery from Databases)*

programación, como el Lenguaje R, diseñados para ayudar a la investigación y dar soporte a diversas áreas, una de ellas: la minería de datos.

2.1.1 BASES DE DATOS: LA INFORMACIÓN

Con la aparición de Internet, Redes sociales y más tarde con el '**IoT**[4]', el volumen de información que se genera crece de una manera exponencial, tanto es así que empresas como Google, Yahoo!, Amazon, etc. tuvieron importantes problemas para seguir realizando sus negocios. Aunque actualmente estos problemas fueron resueltos, en su momento, estos eran producidos por la gran cantidad de datos cuyo procesamiento era cada vez más difícil de realizar, la heterogeneidad de estos, dificultaba las tareas de inserción, consulta o procesamiento de la información y dificultaba poder dar una respuesta rápida.

Ese ritmo de crecimiento ha hecho que Internet o la World Wide Web, se haya convertido en una de las mayores bases de datos o repositorio en la actualidad, tal es así, que la importancia de extraer información válida y útil se ha convertido hoy en día en un factor clave en el mundo empresarial y es por eso que cobra especial importancia las técnicas de minera de datos para poder conseguirlo.

2.1.1.1 BASES DE DATOS

Los tipos de datos que podemos encontrar pueden tener diferentes naturalezas, así como es en el caso de tipos de datos estructurados que se pueden encontrar en bases de datos relaciones, también podemos encontrar otros tipos, como son: **espaciales, temporales, textuales** y **multimedia**, y también datos **no-estructurados** que proceden de internet (páginas web o documentos on-line).

Hemos mencionado las bases de datos relacionales, que son aquellas que contienen tipos de datos estructurados. Los datos son almacenados en tablas que se relacionan entre sí. Cada tabla contiene una cantidad filas o tuplas con diversos datos (columnas o atributos) donde cada fila se identifica mediante su clave primaria.

[4] **IoT**: De sus siglas en inglés (Internet of Things). es un concepto que se refiere a la interconexión digital de objetos cotidianos con internet.

Existen otras bases de datos, como son:

1- **Bases de datos espaciales**: contienen información relacionada con el espacio físico, como datos geográficos, imágenes médicas, redes de transporte.
2- **Bases de datos temporales**: contienen información relacionada con el tiempo, donde se observa la importancia de la evolución en instantes temporales o a lo largo de un intervalo considerable de tiempo un acontecimiento o dato.
3- **Bases de datos documentales**: donde se guarda una relación de índices o descriptores de documentos para encontrar documentos.
4- **Bases de datos multimedia**: Contienen un gran volumen de datos, ya que sus objetos son de tipo video, audio o imágenes.

2.1.2 DATA MINING

"De una manera simplista pero ambiciosa, podríamos decir que el objetivo de la minería de datos es convertir datos en conocimiento." **(J.Hernández Orallo, M. Ramírez Quintana y C.Ferri Ramírez, 2004)**

El data mining o minería de datos, es la fase dentro del **KDD** más característica cuyo objetivo se centra en construir un modelo basado en los datos recopilados de las **BBDD**[5] para producir nuevo conocimiento. Este modelo se construye basándose en los patrones y relaciones que existen en los datos y que pueden usarse para realizar predicciones, o para comprender la información que aportan los datos.

2.1.2.1 TAREAS DE MINERÍA DE DATOS

Dentro de la minería de datos podemos encontrarnos con diferentes tareas en las que se podemos clasificarlas en dos tipos:

- **Predictivas:** Estas tareas pueden ser de clasificación y de regresión.

[5] **BBDD**: De las siglas de, Bases de Datos.

- **Descriptivas:** Estas tareas son de agrupamiento, reglas de asociación, secuenciales y las correlaciones**.**

2.1.2.2 MINERÍA DE DATOS: DEFINICIÓN, CONTEXTO, OBJETIVO

Para explicar qué es la minería de datos y el concepto que engloba, hemos recurrido a varios autores bibliográficos, en los cuales podemos resaltar tres puntos: *su definición, donde situarlo y cuál es su tarea fundamental.*

Como una de las muchas definiciones que existen, incluimos la que realiza M. Pérez Marqués, en su libro: "*Minería de datos a través de ejemplos*", donde añade la siguiente definición:

> "De un modo sencillo podemos definir la minería de datos como un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos. Se trata de analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos."
>
> *(Pérez Marqués, 2014)*

Otro de los autores que hemos incluido, donde explican el contexto en el cual se encuentra la minería de datos, son C. Pérez López y D. Santín González, de su libro: "*Minería de datos. Técnicas y herramientas*.", describiéndolo como una etapa dentro de un proceso y las fases que contiene:

> *"La minería de datos es sólo una etapa del proceso de extracción de conocimiento a partir de los datos (KDD). Este proceso consta de varias fases como la preparación de datos (selección, limpieza, y transformación), su exploración y auditoria, minería de datos propiamente dicha (desarrollo de modelos y análisis de datos), evaluación, difusión y utilización de modelos (output)."*
>
> *(César Pérez López, Daniel Santín González, 2007)*

Figura 1. Proceso de extracción del conocimiento (KDD) - Secuencia de fases

Y por último para terminar el concepto de "Minería de datos", en el libro: "Introducción a la minería de datos" por J. Hernández Orallo, M. Ramírez Quintana y C. Ferri Ramírez, definen cual es el cometido que debe llevar:

> *"…la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización"*
>
> *(J.Hernández Orallo, M. Ramírez Quintana y C.Ferri Ramírez, 2004)*

Por tanto, después de haber mencionado algunas citas bibliográficas podemos resumir de una manera simple y fácil la minería de datos como se representa en la ***Figura 2***.

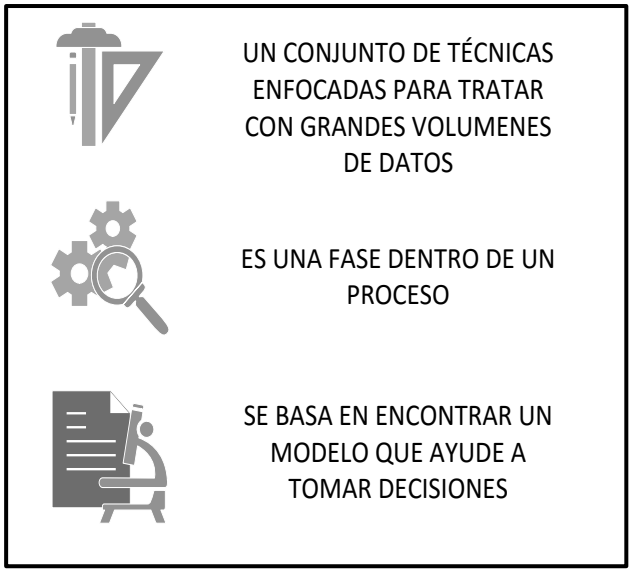


Figura 2. Concepto minería de datos

2.1.3 MACHINE LEARNING

Machine Learning o aprendizaje automático, es el área de la inteligencia artificial que mediante algoritmos o programas, y apoyándose en el campo de la estadística, una máquina o computadora es capaz de aprender, a través de un entrenamiento, un modelo y aplicarlo para desarrollar tareas o resolver problemas que difícilmente podríamos las personas resolver.

El uso en donde se aplica Machine Learning es amplio, ya que cada vez más se va extendiendo a diferentes ámbitos, como son:

- **Meteorológica**: realizando las predicciones del tiempo,
- **Industriales**: encontrando errores en los procesos de fabricación.
- **Estudios de bioingeniería y otras ciencias:** Predecir si un compuesto químico provoca cáncer.
- **Medicina:** Identificación de patologías y diagnósticos de enfermedades.

2.1.4 LENGUAJE R

R es el lenguaje estadístico basado en otro lenguaje llamado S. R es el lenguaje más popular en la comunidad de investigación y estadística convirtiéndose además en una poderosa herramienta en el campo de la minería de datos.

La utilización del Lenguaje R, viene facilitada con R Studio, un **IDE**[6] creado para tal uso. R Studio, funciona por línea de comandos en un terminal. Una de las cosas que lo hace realmente útil es la facilidad y sencillez para poder utilizar y descargar los paquetes desarrollados por los programadores y comunidad que mantiene el lenguaje R activo, dentro de los cuales se encuentran numerosas funciones que ayudan y facilitan al usuario muchas de las labores dentro de la estadística o manejo de datos.

Una de las mayores utilidades que dispone el Lenguaje R, es el paquete gráfico, en el cual se puede visualizar de manera gráfica y de diferentes formas los datos a estudiar. Uno de los paquetes gráficos más usados y del cual se han realizado la mayoría de gráficos presentados en este documento ha sido con la librería (**ggplot**).

[6] **IDE:** Sus siglas en ingles IDE (Integrated Development Enviroment) o Entorno de desarrollo integrado.

2.2 TRABAJO RELACIONADO

Los trabajos relacionados en la investigación dentro de la sanidad utilizando técnicas de minería de datos y machine learning son varios. Uno de los más recientes, y relacionado con la patología del ASMA y que parte como base en este estudio, es el que se realizó en la Universidad de Arizona (Tucson) por (**S. Ram, 2015**).

2.2.1 PREDECIR INGRESOS EN URGENCIAS A TRAVÉS DE TWITTER

El estudio que hicieron recogió como una de las fuentes de datos, los **tuits**[7] de los usuarios, donde observaron durante tres meses las publicaciones que hablaban sobre la patología del asma y la relación con los ingresos en urgencias del Hospital de Dallas.

Además de tener como datos, las publicaciones y los ingresos en urgencias, también tomo como datos, los registros médicos electrónicos, la calidad del aire (obtenida de los sensores ambientales próximos al hospital) y los mensajes de twitter, donde analizaron palabras claves como (‘asma’, ‘jadeo’, etc.). Mediante la información de los registros médicos, pudieron relacionar los tuits con los códigos postales de los pacientes que acudieron al hospital.

El estudio determino que a medida que empeoraba la calidad del aire o aumentaba las publicaciones de twitter relacionados con las palabras clave establecidas, se producían mayores ingresos en urgencias.

La predicción que realizaron determino en un **75%** el número de pacientes que ingresaban en urgencias, de manera que se podían establecer las medidas de prevención a nivel de recursos personales y materiales para atender a estos pacientes. (**2015**)

[7] **Tuit**: Mensaje digital que se envía a través de la red social Twitter® y que no puede rebasar un número limitado de caracteres.

3. Extracción y exploración de datos

Como hemos citado anteriormente la minería de datos es una etapa que se encuentra en el *proceso de extracción de conocimiento a partir de datos* (KDD). La primera fase dentro de este proceso es la *preparación de los datos* donde seleccionaremos, limpiaremos y transformaremos los datos. A continuación, una vez realizada esta fase procederemos al análisis exploratorio y gráfico de los datos.

Con la ayuda de software estadístico podemos abordar esta etapa aplicando las técnicas gráficas para poder estudiar los datos y examinar la información que nos aporta. En este trabajo hemos utilizado R Studio, que nos permite poder utilizar el lenguaje R y utilizar las librerías gráficas para el examen de los conjuntos de datos que vamos a tratar.

3.1 Caso de estudio: Hospital Universitario y Politécnico "La Fe"

Este estudio se sitúa dentro del contexto geográfico y demográfico de una ciudad, que es Valencia, aunque no tenga una relevancia considerada, es importante situarnos para conocer mejor el entorno sobre el que el estudio se centra, ya que considerando que este estudio se fija en uno de los servicios sanitarios de la ciudad, como pilar fundamental, se considera que es también importante conocer los factores que intervienen a groso modo, como el clima que tiene o la cantidad de hospitales que existen actualmente.

Igualmente importante es conocer cómo funcionan los servicios de urgencia y la relevancia que tiene este estudio dentro de ellos.

3.1.1 Valencia

Actualmente la ciudad de Valencia tiene una población en su núcleo urbano de 786.189 habitantes (INE 2015), es la tercera ciudad más poblada detrás de Madrid y Barcelona. La ciudad cuenta con un clima mediterráneo suave durante los inviernos y caluroso y seco durante los veranos, la temperatura media anual es de 18,4ºC.

La ciudad cuenta con diferentes Hospitales, como son:

Centros Hospitalarios de la ciudad de Valencia		
Centro Hospitalario	**Tipo**	**Nº de Camas**
Hospital Pare Jofre	Público	125
Clínica fontana	Privado no benéfico	7
Hospital 9 de Octubre	Privado no benéfico	300
F.I. Valenciano de oncología	Otro – privado-benéfico	160
Hospital Valencia al mar	Privado no benéfico	70
Clínica Casa de la Salud	Privado benéfico (Iglesia)	192
Consorcio Hospital General Universitario de valencia	Público	592
Hospital La Malvarrosa	Público	47
Hospital Clínico Universitario	Público	587
Hospital Arnau de Vilanova	Público	302
Hospital Universitario Doctor Peset	Público	539
Hospital Universitario y Politécnico La Fe	Público	1440
Clínico Quirón de Valencia S.A.	Privado no benéfico	79
Clínica Virgen del Consuelo	Privado no benéfico	156

Tabla 1. Hospitales en Valencia

Podemos observar que Valencia dispone de un gran servicio sanitario, aunque bien no todos estos servicios son públicos, sí lo son en su mayoría. El estudio que se realiza se centra en los ingresos de urgencias de los pacientes con diagnóstico asmático de uno de estos hospitales como es el **Hospital Universitario y Politécnico La Fe.**

3.1.2 SERVICIOS: URGENCIAS

Los hospitales utilizan en sus urgencias un sistema de triado para determinar la prioridad de cada paciente. En la Comunidad Valenciana, se emplea el sistema de triado **MTS**[8] que consiste en una metodología para la clasificación y prioridad de las atenciones que realizan los profesionales sanitarios a los pacientes que ingresan. Dentro de esta metodología, se

[8] **MTS (Manchester):** Basado en el sistema del mismo nombre del Reino Unido. A partir de 51 motivos de consulta y a través de unas preguntas dirigidas en un diagrama. Es decir según la respuesta si/no, se produce la clasificación, con 5 niveles de gravedad.

encuentra como 1 nivel (prioridad máxima), la vía respiratoria, en la que entra como un discriminador general de clasificación.

3.1.2.1 SISTEMA DE TRIADO MTS

Número	Nombre	Color	Tiempo Máximo (min.)
1	Atención inmediata	**Rojo**	0
2	Muy urgente	**Naranja**	10
3	Urgente	**Amarillo**	60
4	Normal	**Verde**	120
5	No urgente	**Azul**	240

Tabla 2. Manchester System Triage

La importancia que tiene este estudio sobre la predicción de pacientes en urgencias con diagnóstico asmático viene dada también por este sistema de triaje en el cual estos pacientes requieren de una atención inmediata por parte del personal sanitario pudiendo ser un contratiempo en la organización y atención a otros pacientes dentro del módulo de Urgencias del hospital.

Un artículo de una revista científica sobre el Análisis del Sistema Sanitario en Navarra (Año 2010) expone que:

> *"El usuario demanda de este servicio una respuesta rápida y satisfactoria y el gestor intenta proporcionarla de la manera más organizada y eficiente posible. En este escenario, los profesionales de la medicina de urgencias y emergencias, que son los encargados de interpretar este complejo equilibrio, se encuentran inmersos en un marco de acción complejo. Sin posibilidades de control sobre el acceso del primero, y con los medios proporcionalmente insuficientes para hacerle frente que le facilita el segundo, el resultado neto de este equilibrio es a menudo el retraso en la dispensación del servicio, cuando no la saturación del sistema "*
>
> *(Urgencias y emergencias: al servicio del ciudadano, 2010)*

3.2 ASMA

El Asma, está catalogado como una enfermedad respiratoria crónica, la **OMS**[9] la define de esta manera:

> *"El asma es una enfermedad crónica que se caracteriza por ataques recurrentes de **disnea**[10] y **sibilancias**[11], que varían en severidad y frecuencia de una persona a otra".*

La gravedad de esta enfermedad viene dada cuando las personas que la padecen presentan un ataque de asma, ya que el revestimiento de sus bronquios se inflama, lo que hace que las vías respiratorias se obstruyan y el flujo de aire que entra sea mucho menor.

Aunque no es una enfermedad letal, en comparación con otras enfermedades crónicas, se estima que la tasa de mortalidad por asma en 2005 fue de 255.000 personas, según datos de la OMS. Estudios más frecuentes, como el realizado en otros países como en Villa Clara, Cuba, afirman:

> *"En 18 años se han realizado 16340 autopsias, corresponden a asma bronquial como causa básica de muerte, 41: siete fallecieron por status asmático, 16 por sepsis respiratoria y 18 por muerte súbita, entre las que se encontró relación con la medicación con broncodilatadores en aerosoles. El número de muerte por asma bronquial es bajo como expresión de la adecuada atención médica a todos los niveles; la prevención de la muerte súbita por esta enfermedad debe apoyarse en la divulgación de los riesgos y los beneficios del tratamiento con broncodilatadores. (Estudio de la mortalidad por asma bronquial, 2011)*

3.2.1 ¿QUÉ FACTORES INFLUYEN?

Una alergia es un proceso en el que una sustancia que aparentemente puede ser inocua para la mayoría de personas para otras puede ser intolerante o dañina. Estas sustancias se conocen como alérgenos y hay de diferentes tipos, para el estudio en el que nos centramos, los relevante son los Alérgenos del aire (**Neumoalergenos**).

---

[9] **OMS:** *Organización Mundial de la Salud. La Constitución de la OMS entró en vigor el 7 de abril de 1948, fecha que conmemoramos cada año mediante el Día Mundial de la Salud.*

[10] **Disnea**: Ahogo o dificultad en la respiración.

[11] **Sibilancia:** La sibilancia es un ruido inspiratorio o espiratorio agudo que aparece en el árbol bronquial como consecuencia de una estenosis (estrechamiento de un orificio o conducto corporal).

Este tipo de alérgeno da lugar a enfermedades en órganos que se exponen al aire, como los ojos (Conjuntivitis), nariz (Rinitis, Polipos y sinusitis) o bronquios (Asma).

Los factores que pueden intervenir a acrecentar este tipo de alergia pueden ser: ácaros del polvo, polen, mohos, productos contaminantes y la caspa de animales (la cual está formada por diminutas escamas o partículas que se desprenden del pelo, las plumas o la piel) de cualquier animal doméstico.

3.2.1.1 POLEN

El polen es un alérgeno relevante en los problemas respiratorios relacionados con el asma. Según datos del Ayuntamiento de Valencia:

> "El polen, como alérgeno ocupa un segundo lugar, en orden de importancia, en la etiología de problemas alérgicos, después de los ácaros, en nuestro ambiente, al contrario de lo que ocurre en otras latitudes tanto de nuestra Comunidad Valencia como de otras zonas geográficas de España".

Los beneficios de que Valencia se encuentre en una zona costera conllevan que la humedad reduce la importancia del riesgo por alergia al polen en la ciudad, aunque según un informe del propio Ayto. de Valencia sigue existiendo alergias al Polen en la ciudad, aunque en menor medida que en zonas más secas.

3.2.1.2 CLIMATOLOGÍA

Como decíamos anteriormente, el clima no es un factor que directamente influya en los problemas alérgicos de asma, pero tal y como comentábamos en el punto anterior, Valencia se beneficia de la humedad, aunque si consideramos los periodos del año en que las temperaturas son más cálidas, el factor de riesgo puede aumentar en la medida que el Polen aumenta.

Comentamos que el clima no afecta directamente, pero según los estudios, las personas con problemas asmáticos suelen verse indirectamente más afectadas en periodos de baja temperatura y humedad. La temperatura media anual en Valencia se encuentra a 18,4ºC.

3.2.1.3 CONTAMINACIÓN ATMOSFÉRICA

La contaminación atmosférica se presenta como una concentración creciente de aire y partículas de materia contaminantes. El aumento de estas concentraciones o niveles disminuye la calidad y pureza del aire y crea factores de riesgo para la salud.

Estos niveles están relacionados y derivados por la emisión de gases de vehículos de motor y fábricas.
Si profundizamos en los tipos de contaminantes tóxicos que se pueden encontrar, como principales encontramos:

- dióxido de nitrógeno (NO2)
- dióxido de azufre (SO2)
- ozono (O3)
- partículas de diámetro inferior o igual a 10 µm (PM10).

De estos contaminantes, el nivel de Ozono a un nivel bajo cerca del suelo y en una concentración alta, puede causar inflamación y dañar el revestimiento de los pulmones, lo que hace que las personas con asma tengan mayores dificultades para respirar.
Cerca del suelo, es perjudicial, porque está formado por reacciones químicas entre rayos del sol y gases orgánicos, y por óxidos de nitrógeno emitidos por coches, centrales eléctricas, calderas industriales, refinerías, plantas químicas, etc..

También el Dióxido de nitrógeno provocado por los vehículos a motor, entre otros, aumenta la incidencia del asma y el riesgo de muerte por neumopatías.

El Dióxido sulfúrico o de azufre causa también enfermedades respiratorias en especial niños y ancianos, agrava las enfermedades cardiacas y pulmonares, en especial en las personas con asma. Y por último las partículas de sulfato (formadas cuando el SO2 reacciona con otros compuestos químicos del aire) se acumulan en los pulmones y aumentan los síntomas y las enfermedades respiratorias, la dificultad respiratoria e incluso el riesgo de muerte prematura. (Comité de Salud y Medio ambiente Soc. Europ Enfermedades Respiratorias)

3.3 Selección, Limpieza y Transformación

Esta fase tiene como punto primordial extraer la información que sirva como entrada para las sucesivas fases. Es por ello, que se debe extraer conocimiento valido y útil a raíz de la información que tratamos.

Para el tratamiento de los datos, hemos preferido trabajar con formatos **CSV**[12] para poder realizar la lectura y extracción. Dicho esto, algunos conjuntos de datos se han tenido que transformar a este formato, apoyándonos de una hoja de cálculos de un paquete ofimático, podemos insertar y transformar fácilmente estos conjuntos de datos al formato deseado.
Mostraremos los conjuntos de datos con los que hemos trabajado y su estructura. Dentro de la exploración de datos, realizaremos las operaciones de limpieza (DATA CLEANING), eliminado aquellos datos que no resultan relevantes o útiles y transformaremos los conjuntos de datos, añadiendo atributos para poder obtener la información necesaria.

El primer conjunto de datos que vamos a mostrar, es el más relevante, ya que está relacionado con las ***urgencias hospitalarias.*** Este conjunto de datos, facilitado por la unidad informática del Hospital Universitario y Politécnico "La Fe", contiene la información de que aquellos pacientes que han sido ingresados desde el 2008 en urgencias con diagnostico relacionado con el código **493.0**[13].

[12] **CSV:** *Formato CSV (del inglés comma-separated values) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal: Argentina, México, Brasil...) y las filas por saltos de línea.*

[13] **493.0:** *Código que hace referencia al diagnóstico: ASMA. Este valor se encuentra en los campos ciePrincipalCodigo y cieObjetivoCodigo*

3.3.1 URGENCIAS HOSPITALARIAS

COLUMNA	TIPO	DESCRIPCIÓN
dataingr	Fecha	Fecha de ingreso
horaingr	Hora	Hora de ingreso
Numerohc	entero	Numero de Historia clínica
Numicu	entero	Número del episodio
Fechanac	Fecha	Fecha de nacimiento del paciente
Sexo	entero	**1:** Hombre **2:** Mujer
ciePrincipalCodigo	entero	Código del diagnóstico principal
ciePrincipalDescripcion	texto	Descripción del diagnóstico principal
cieObjetivoCodigo	entero	Código del diagnóstico secundario
cieObjetivoDescripcion	texto	Descripción del diagnóstico secundario
tipoDiagObjetivo	entero	**1**: Los diagnósticos principal y objetivo coinciden. **2**: El ingreso es a causa del diagnóstico principal pero tiene relación con el diagnóstico objetivo.

Tabla 3. Formato de los datos de ingresos en Urgencias

Examinando el conjunto de datos, hemos tenido que realizar varias operaciones de selección y limpieza, para obtener la información útil:

1) Hemos añadido una columna para calcular la "**edad**" *(aunque en principio no es un dato relevante, sirve como conocimiento de la población que vamos a estudiar y determinar que rango de edades han sido las que más han sido afectadas.)*
2) Dado que tenemos varios diagnósticos, hemos observado que el conjunto de datos tiene ingresos de pacientes cuyo motivo principal no es a causa de asma, ya que estos pacientes son ingresados por cualquier otro motivo, pero muchos de ellos su diagnóstico secundario es de tipo asmático ya que se relaciona con el diagnostico principal.
3) Observamos que el número de ingresos varía drásticamente desde comienzos del 2009 hasta mediados de 2011. Se comprueba que esto es a causa del traslado del antiguo Hospital "La Fe" ubicado en Campanar, al bulevar sur donde se sitúa actualmente. Por

tanto, se decide sesgar los datos a partir de la fecha de 2011 para estudiar los ingresos que se realizan habitualmente en la nueva ubicación.

Extraemos un nuevo conjunto de datos de aquellos pacientes con diagnóstico principal relacionado con asma, eliminado aquellos donde han sido ingresados por otra causa y obteniendo los datos a partir de mediados del 2011, a partir de estos datos obtenemos la frecuencia de ingresos por semanas, meses y año.

	anyo	mes	semana	Total
1	2011	6	22	17
2	2011	6	23	9
3	2011	6	24	9
4	2011	6	25	9
5	2011	6	26	4

Tabla 4. Datos de urgencias aplicando las transformaciones

3.3.2 TEMPERATURAS

Los datos sobre temperatura, muestran los registros indicando temperaturas máximas y mínimas desde 2010 hasta Mayo de 2016.

COLUMNA	TIPO	DESCRIPCION
Date	Fecha/Hora	Fecha y hora del registro
TMax	Decimal	Temperatura máxima
TMin	Decimal	Temperatura mínima

Tabla 5. Formato de los datos climatológicos: Temperaturas

Lo primero que hemos tenido que realizar, es obtener en columnas separadas las semanas, meses y años de cada registro, a fin de poder agrupar las temperaturas medias de cada semana por meses y año.

Una vez realizado esto, podemos ver qué semanas han sido más calurosas y frías, pero para poder realizar una estimación de la temperatura media, hemos considerado obtener la media de estos dos valores de modo que sirva de indicador sin tener que irnos a las temperaturas extremas de frio y calor.

La transformación de los datos mantiene la siguiente estructura:

	anyo	mes	semana	Total
1	2010	1	0	11.83
2	2010	1	1	8.40
3	2010	1	2	12.24
4	2010	1	3	12.75
5	2010	1	4	10.24
6	2010	2	5	11.70
7	2010	2	6	9.09
8	2010	2	7	10.49

Tabla 6. Datos de temperaturas aplicando las transformaciones

3.3.3 POLEN

Los datos polínicos provienen de tres estaciones donde obtienen los niveles para varios tipos de polen. Considerando que una de las estaciones proviene de la estación de "**Xátiva**" y el estudio se centra en la ciudad de Valencia, eliminamos esta estación quedándonos únicamente con la estación del "**Hospital Clínico**" y la estación de "**Campanar**".

Estaciones de nivel de polen en Valencia:

Mapa 1. Estación de H. Clínico

Mapa 2. Estación de La Fe (Campanar)

COLUMNA	TIPO	COLUMNA	TIPO
Fecha	Fecha/Hora	**Artemisia**	Entero
Betula	Entero	**Castanea**	Entero
Chenopodiaceae.Amarantaceae	Entero	**Cupressaceae.Taxaceae**	Entero
Morus	Entero	**Olea**	Entero
Palmae	Entero	**Pinus**	Entero
Plantago	Entero	**Platanus**	Entero
Poaceae	Entero	**Populus**	Entero
Typhaceae	Entero	**Ulmus**	Entero
Urtica.membranaceae	Entero	**Otros**	Entero

Tabla 7. Conjunto de datos de Polen (conjunto reducido, existen 64 tipos de polen)

Para poder trabajar más fácilmente hemos agrupado en un solo conjunto los datos y añadido una columna que sirva como discriminador, que será la "**Estación**". Empleamos el mismo procedimiento que anteriormente, obtenemos la semana, mes y año de cada registro para agregarlos en tres columnas. Dado que el tipo de polen no es determinante para el estudio, sino la cantidad global que se alcanza, añadimos por último, una columna, "**Total**", donde indica la suma de cada tipo polínico de cada registro.

Después de aplicar esta transformación, podemos obtener la media de las dos estaciones fácilmente. El resultado de la transformación quedaría como se muestra en la siguiente tabla:

	anyo	mes	semana	Total
1	2009	1	0	6.50
2	2009	1	1	2.21
3	2009	1	2	16.36
4	2009	1	3	16.07
5	2009	1	4	5.08
6	2009	2	4	8.00
7	2009	2	5	14.93
8	2009	2	6	69.57
9	2009	2	7	88.71

Tabla 8. Datos de polen aplicando las transformaciones

3.3.4 CONTAMINACIÓN

La **RVVCCA**[14] dispone de una red de estaciones para poder llevar un control de los niveles de calidad de aire en la comunidad valenciana. En este estudio, se muestran las estaciones ubicadas en la ciudad de Valencia ya que son especialmente las que nos interesan. Podemos situarlas dentro de la ciudad en el siguiente mapa:

Mapa 3. Estaciones de contaminación (Valencia)

Como podemos observar, una de las estaciones, queda fuera del núcleo urbano, la estación de "Valencia-Albufera", por lo que, para obtener unos valores medios más precisos descartamos esta estación de los datos, al igual que la estación de "Conselleria" que no contiene valores de contaminantes.

Por cada estación encontramos una estructura similar a la mostrada en la siguiente tabla:

COLUMNA	TIPO
Fecha	Fecha
NOx	Entero
O3	Entero
Veloc.	Decimal
Direc.	Entero
PM2.5	Entero
PM1	Entero
SO2	Entero
CO	Decimal
NO	Entero
NO2	Entero
PM10	Entero

Tabla 9. Estructura datos contaminación

[14] **RVVCCA:** *Red Valenciana de Vigilancia y Control de Contaminación Atmosférica. Es el organismo competente para la evaluación y gestión de la calidad del aire ambiente en la Comunidad Valenciana.*

Según la **OMS** las partículas más perjudiciales para la salud son las de 10 micrones de diámetro o menos (≤**PM10**[15]), como las **PM2.5**[16], ya que estas pueden penetrar y alojarse en el interior profundo de los pulmones. Las partículas de 1 micrón de diámetro (**PM1**[17]), son perjudiciales en la salud, pero afectan más al corriente sanguíneo que al respiratorio, por lo que suelen afectar más a mujeres embarazadas o a personas con problemas cardiacos, este tipo de partículas es descartado del conjunto de datos.

Sobre los contaminantes a tener cuenta, en los relacionados al asma o problemas respiratorios, la OMS nos indica lo siguiente:

> "El ozono (O3) es un importante factor de mortalidad y morbilidad por asma, mientras que el dióxido de nitrógeno (NO2) y el dióxido de azufre (SO2) pueden tener influencia en el asma, los síntomas bronquiales, las alveolitis y la insuficiencia respiratoria." *((OMS), Organización Mundial de la Salud)*

Por tanto para la transformación de los datos, hemos tenido que estudiar qué contaminantes y partículas de materia son relevantes para nuestro caso, como son el O3 (Ozono), SO2 (Dióxido de Azufre) y NO2 (Dióxido de nitrógeno), también hemos incluido las partículas de materia PM2.5 y PM10.

En estos datos, el esfuerzo de transformación ha sido mayor, ya que además de tener que extraer las columnas de los valores que nos interesan, la cantidad de datos por estaciones por cada año hace que el trabajo haya sido más laborioso, es por ello que vamos a mencionar los pasos que hemos llevado.

1. Eliminación de columnas no relevantes en cada uno de los datos de cada estación.
2. Igualar columnas (en aquellas que carezcan de algún valor, añadimos una columna con valores nulos)

[15] **PM10**: *Partículas que pasan a través del cabezal de tamaño selectivo, para un diámetro aerodinámico de 10 µm con una eficiencia de corte del 50 %. (Partículas respirables)*
[16] **PM2.5:** Partículas que pasan a través del cabezal de tamaño selectivo, para un diámetro aerodinámico de 2,5 µm con una eficiencia de corte del 50 %. (Partículas finas)
[17] **PM1:** *Partículas sub micrónicas (muy pequeñas, que solo afectan a embarazadas o problemas cardiacos al introducirse en la corriente sanguínea)*

3. Añadimos una columna que identifique la "Estación".
4. Fusionamos los datos de cada estación en un solo conjunto de datos.
5. Aplicamos el formato de fecha adecuado.
6. Añadimos la suma total de contaminantes y de partículas en dos columnas "**ConTotal**" y "**PMTotal**" por cada registro.

Después de realizar este proceso agrupamos los datos por fecha (año, mes y semana) y obtenemos de cada registro como indicador la media para la contaminación total y las partículas de materia.

Puesto que necesitamos un valor como indicador general de contaminación (incluyendo las PM), consideramos que la suma de ambas variables puede servirnos como indicador del total de contaminación en el ambiente por cada registro, por ello el conjunto de datos se formará con los datos del año, mes y semana y el indicador global de contaminación que llamaremos "**Total**". La transformación queda de la siguiente forma:

	anyo	mes	semana	Total
1	2010	1	0	94.64
2	2010	1	1	77.94
3	2010	1	2	92.00
4	2010	1	3	105.83
5	2010	1	4	98.86
6	2010	2	5	104.83
7	2010	2	6	102.29
8	2010	2	7	95.80
9	2010	2	8	101.40

Tabla 10. Datos de contaminación aplicando las transformaciones

3.4 ANÁLISIS EXPLORATORIO Y GRÁFICO DE LOS DATOS

3.4.1 URGENCIAS HOSPITALARIAS:

Como decíamos anteriormente, se observa una clara diferencia en los primeros años respecto mediados del 2011. Este cambio se debe al traslado del Hospital, ubicado en la zona de Campanar a la nueva ubicación donde el cambio de población se redujo considerablemente.

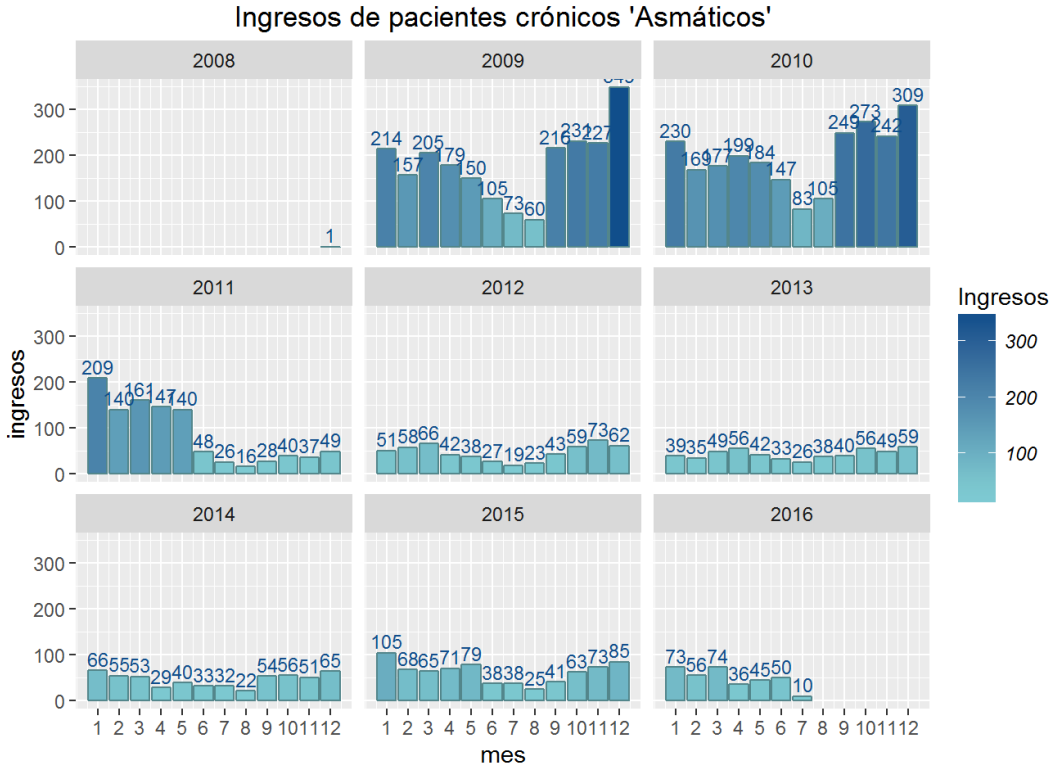


Figura 3. Ingresos desde 2009

Observamos como el número de ingresos cambia significativamente, pasando de tener ingresos elevados de más de 250, a aproximadamente una media entorno a los 40 / 50 ingresos al mes.

Por tanto, después de las transformaciones, y eliminando los valores anteriores a Junio de 2011, podemos observar gráficamente con mayor detalle el número de ingresos.

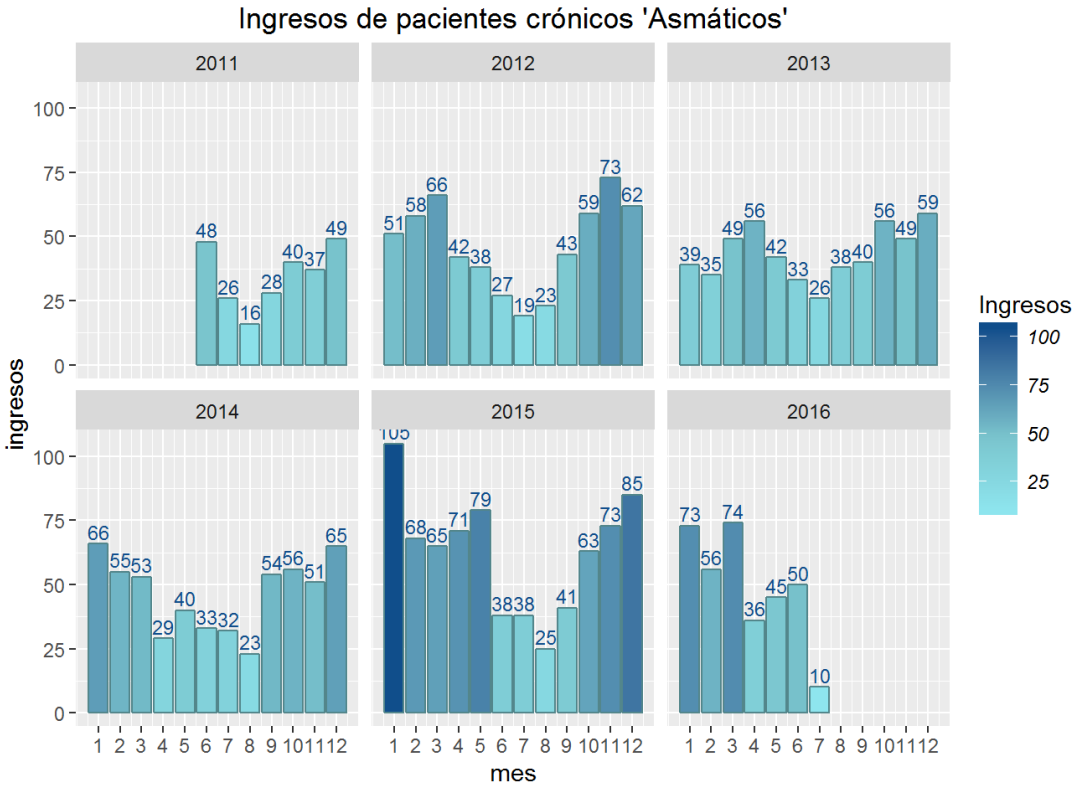


Figura 4. Ingresos desde Junio 2011

Concluimos, que entre los primeros meses del año, y los últimos hay una mayor frecuencia de ingresos. Los motivos pueden ser, a causa de bajas temperaturas, donde suele afectar a los pacientes asmáticos, los meses donde existe mayor polinización, como suele ser en la estación de primavera (marzo, abril y mayo), y en la estación de Otoño, donde según **SEICAP**[18], los cambios de humedad y temperatura generan un aumento de los casos de ataque de asma en urgencias.

[18] **SEICAP**: Sociedad Española de Inmunología Clínica y Alergia Pediátrica.

Como información adicional podemos observar los rangos de edades que más frecuencia han tenido, a fin de conocer la población en la que se aplica el estudio:

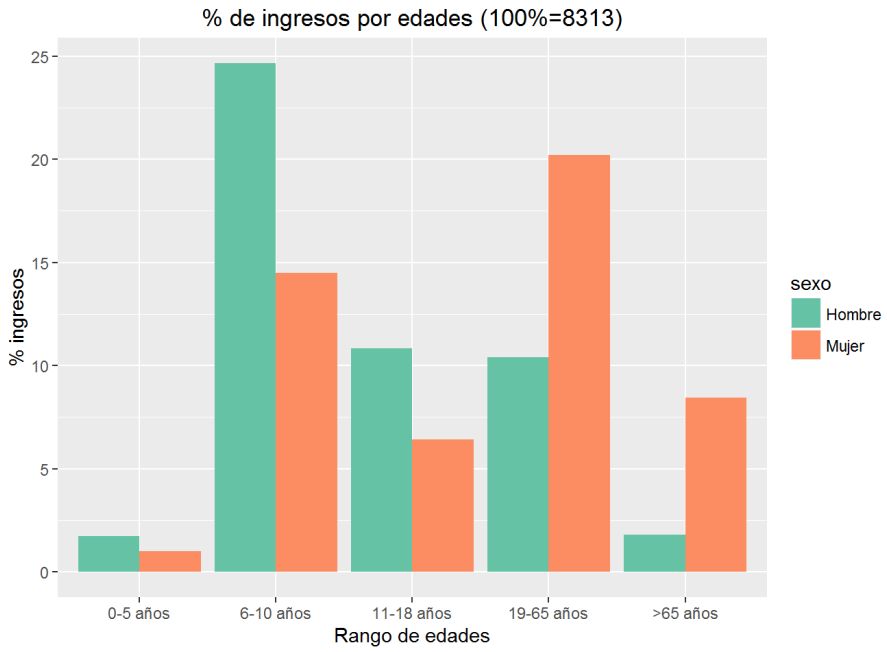


Figura 5. Ingresos por rango de edades

Podemos observar como las edades más tempranas, de entre 6-10 años, son las que más padecen los ataque por asma, particularmente el doble de hombres que mujeres, lo que cambia inversamente en la edad adulta donde son las mujeres las que sufren con mayor frecuencia los problemas asmáticos.

3.4.2 TEMPERATURAS:

Dentro del análisis exploratorio de la temperatura, en lo que respecta al estudio, se trata de un factor independiente, por lo que haciendo una exploración gráfica no podemos determinar a qué se deben los cambios que observamos.

Aun así, esta exploración nos sirve para conocer los valores cuantitativos que ha tomado la temperatura en los últimos años y determinar si existe una relación con los ingresos de los pacientes en urgencias.

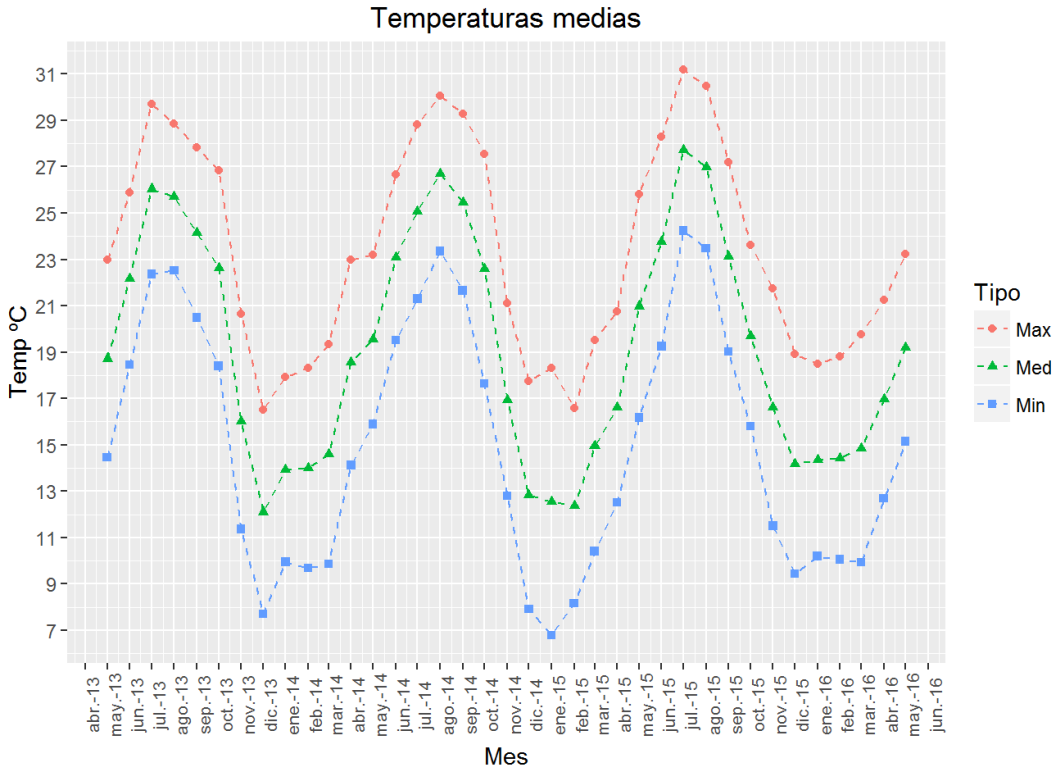


Figura 6. Temperaturas medias

Observamos en este gráfico como cada año, las medias, tanto la máxima como la mínima han ido aumentando ligeramente, lo que nos lleva a pensar que este aumento pueda afectar al entorno, como por ejemplo, que se adelante la época de polinización.

3.4.3 POLEN

En la exploración gráfica de los datos del polen, observamos como en la estación H. Clínico, al ser en una zona céntrica y con mucha vegetación alrededor, los niveles de polen son muchos más elevados en comparación con los niveles de la zona de Campanar.

Se aprecia en ambos casos que en el comienzo de la primavera (mes de Marzo), el aumento es significativo, además si lo comparamos anualmente vemos cada año un nivel de polen creciente en los meses de Marzo, Abril y Mayo.

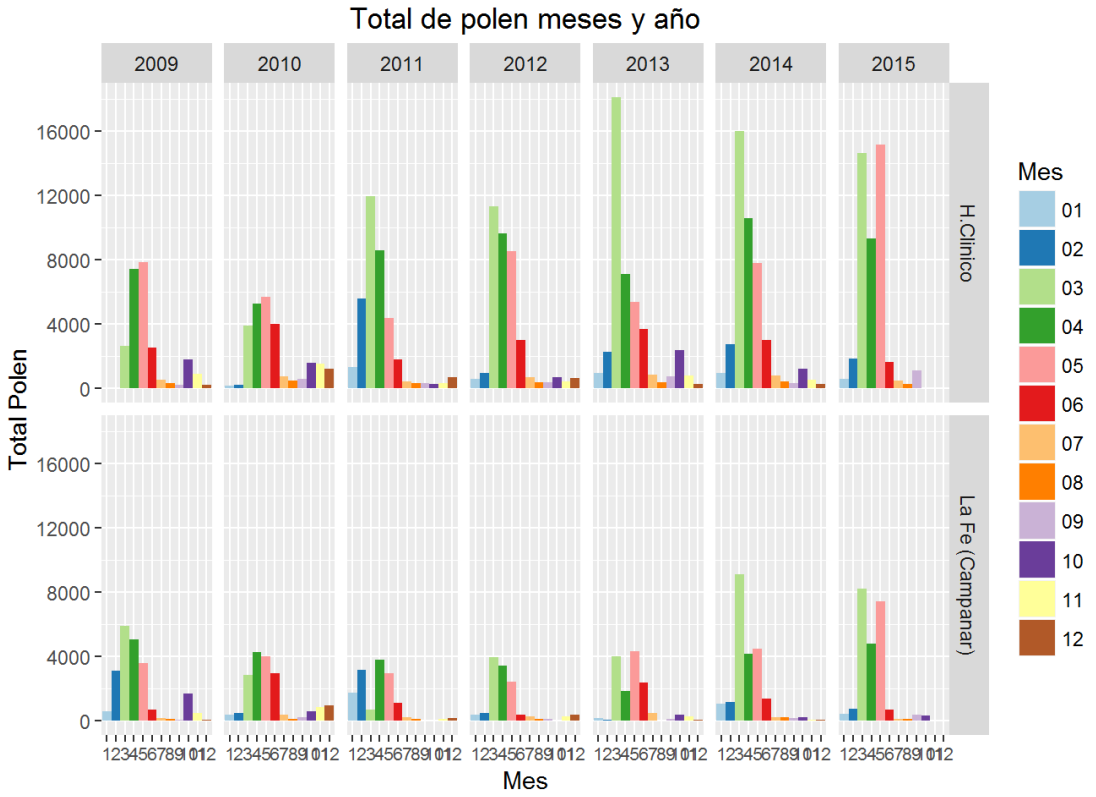


Figura 7. Niveles de polen

Si aplicamos la media de los valores de las dos estaciones, podemos observar el aumento del polen en los últimos años con mayor claridad.

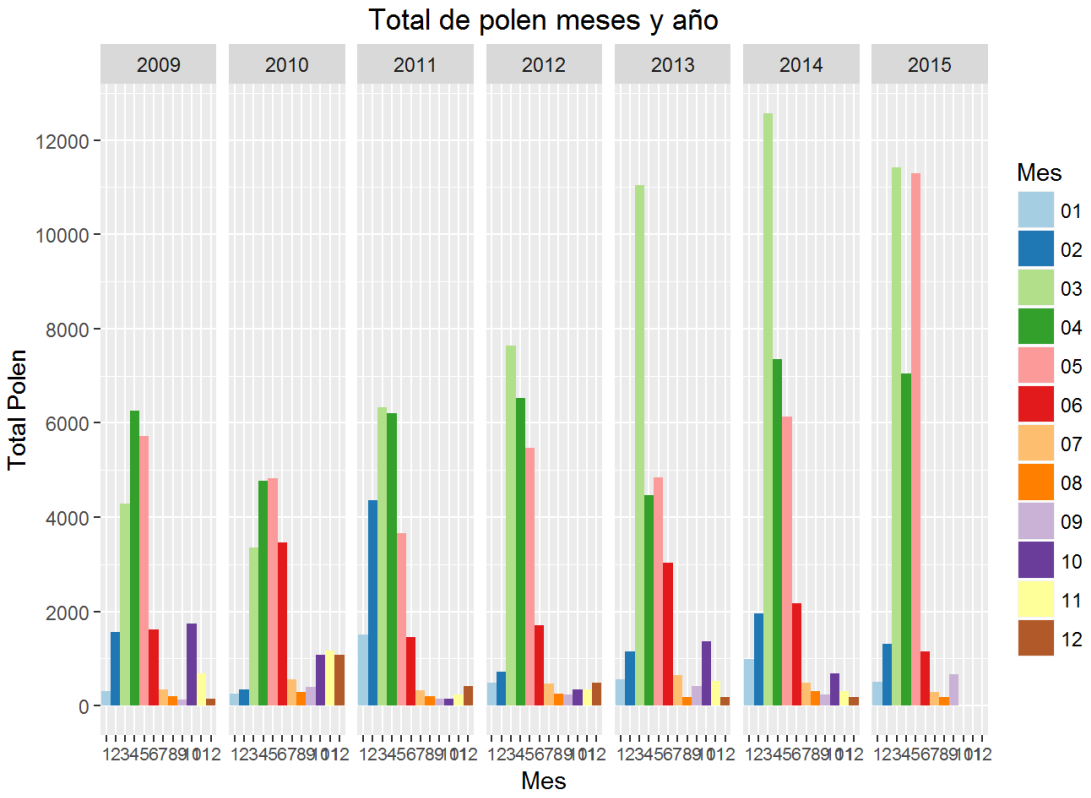


Figura 8. Niveles de polen - Media de estaciones

3.4.4 CONTAMINACIÓN

Como podemos observar en la gráfica siguiente, la evolución de los contaminantes principales nos muestra que en los últimos años, desde 2011 a 2015, los valores del Ozono han sido mayores entre el 2º y 3º trimestre del año (desde Marzo a Agosto) cuya media no muestra tendencia al cambio, esto no ocurre con el dióxido de nitrógeno, que su media sí en el último año mostraba una ligera tendencia creciente en su nivel, algo que no ocurre lo mismo con el Dióxido de nitrógeno, que muestra una línea decreciente en los últimos años.

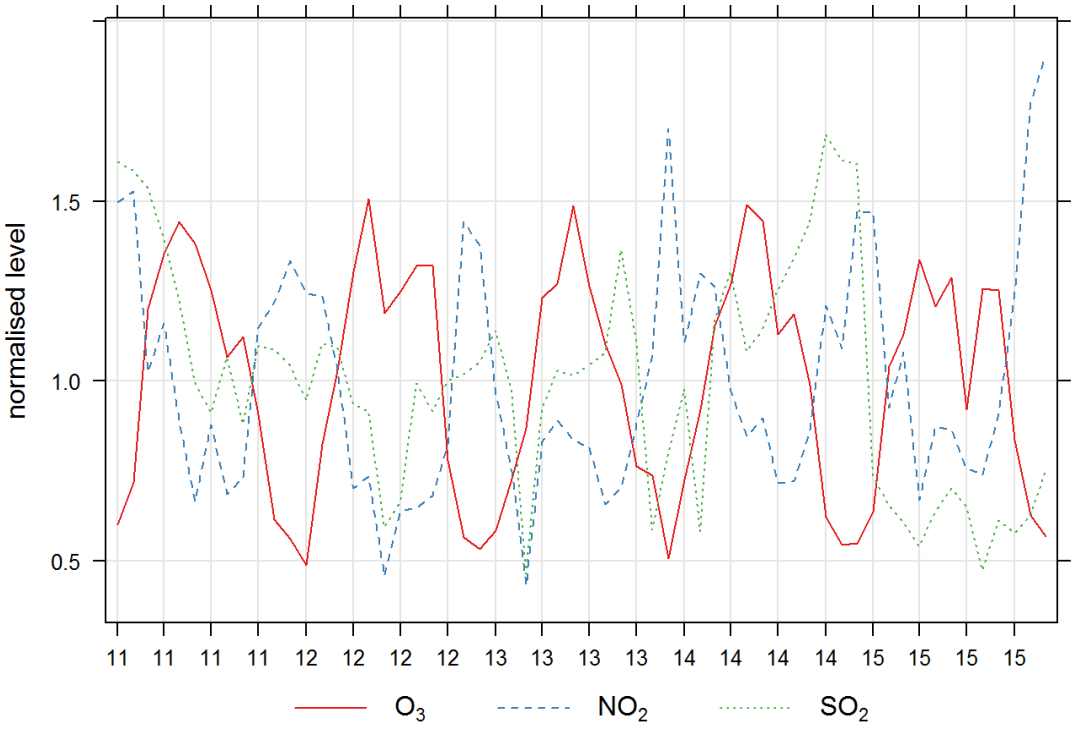


Figura 9. Evolución de los contaminantes en los últimos años

En la siguiente gráfica observamos la media de cada mes de los últimos años y la evolución durante la semana, donde hay una concentración creciente del ozono desde principios de año hasta mediados.

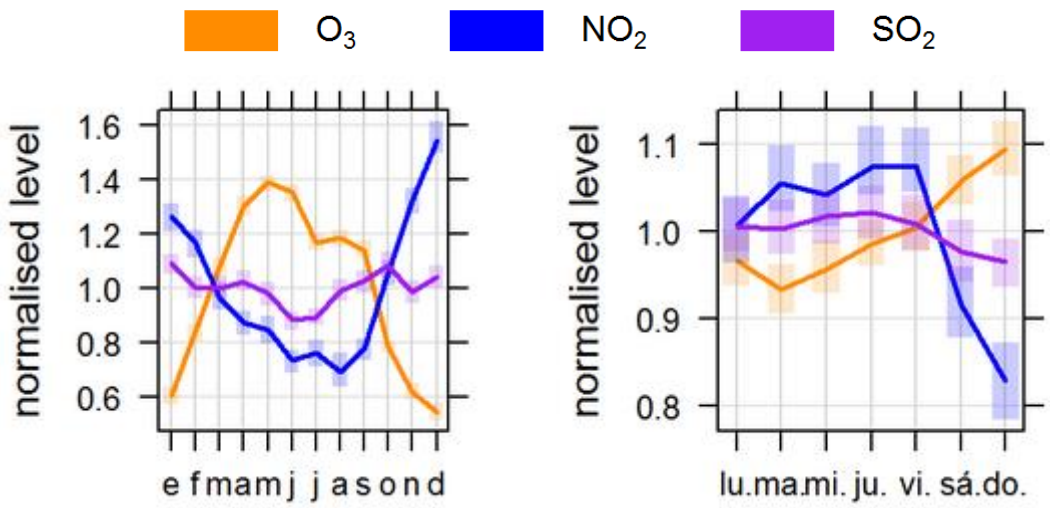


Figura 10. Evolución mensual y por semana

Si observamos en detalle los años: 2011, 2013 y 2015 podemos ver la evolución de estos contaminantes:

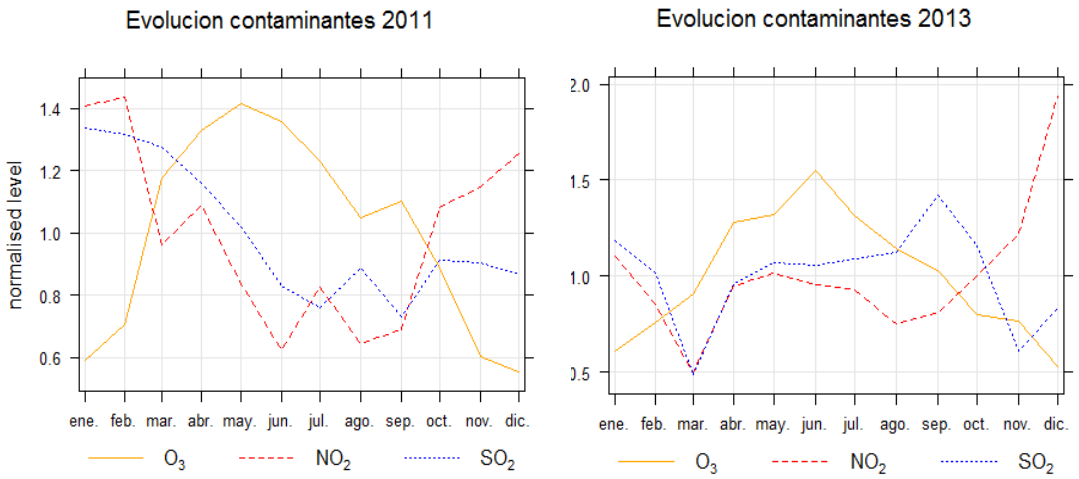


Figura 11. Evolución de contaminantes de 2011 y 2013

Como podemos observar entre 2011 y 2013 hubo un cambio en los niveles donde vemos que aumentaron destacando sobretodo el dióxido de nitrógeno. Si observamos la gráfica de la evolución en 2015, la tendencia se mantiene aunque ha bajado ligeramente el Ozono, presentado más fluctuación en los meses de verano y el dióxido de azufre aumenta alcanzado su mayor nivel.

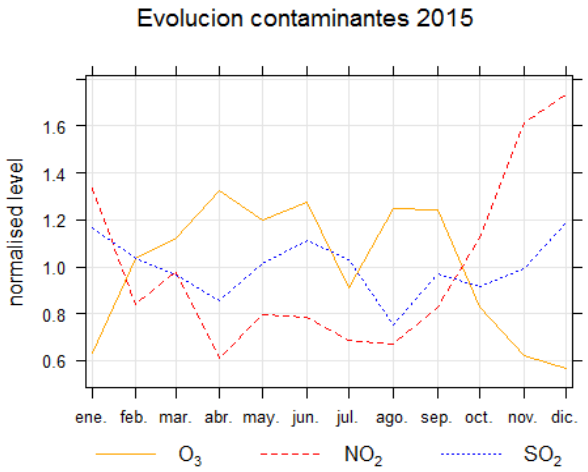


Figura 12. Evolución contaminantes año 2015

Parte importante del estudio de la calidad del aire y la relación con problemas respiratorios viene derivado tambien de las particulas de materia (PM) que se encuentran. Las gráficas que se muestran, enseñan la evolución de las concentraciones de estas particulas presentes en los ultimos años:

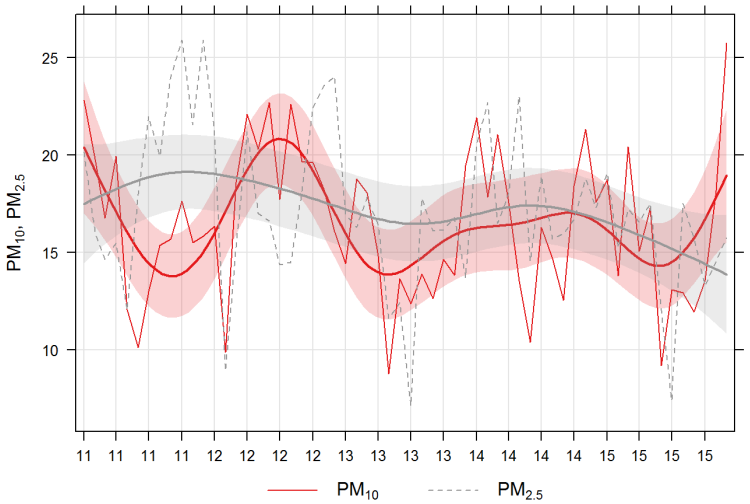


Figura 13. Evolución de las partículas (PM)

Tal y como se observa, podemos apreciar que durante los años las PM10 ha ido oscilando, alcanzando cotas superiores a **20µ/m³** durante algunos periodos.

Observando en la ultima gráfica, podemos ver que la media mensual de los ultimos años hace ver que donde se realiza mayor concentración de µ/m^3 que suele ser en los meses de febrero a abril.

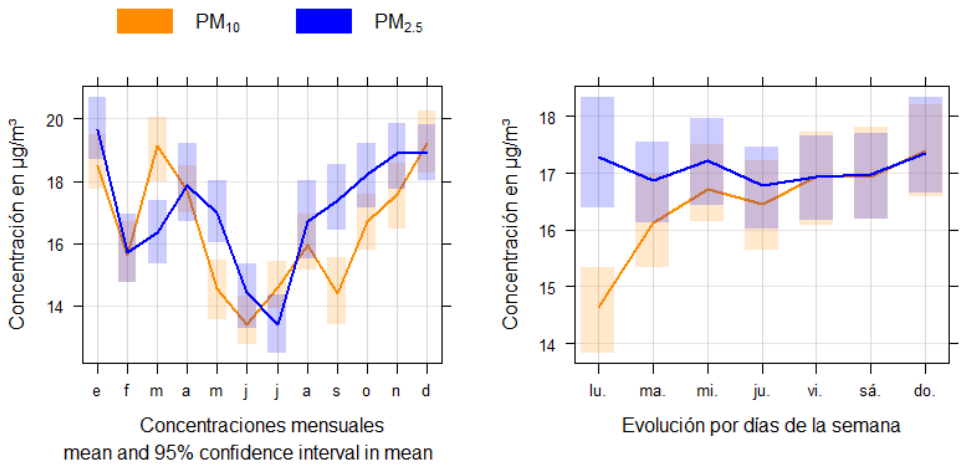


Figura 14. Evolución mensual y por semana

Si mostramos las tendencias del total de contaminación y del total de partículas de materia en los últimos años, podemos una evolución creciente en la que destaca notablemente el aumento de PM:

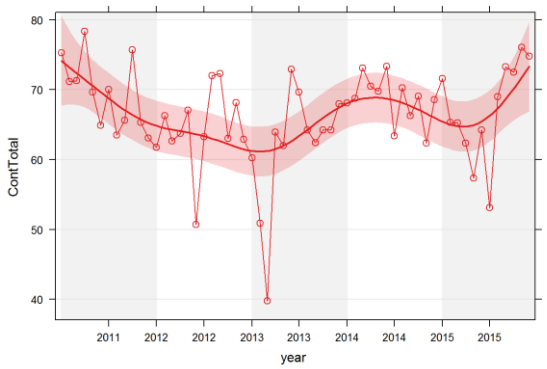


Figura 15. Evolución de la concentración de los contaminantes primarios.

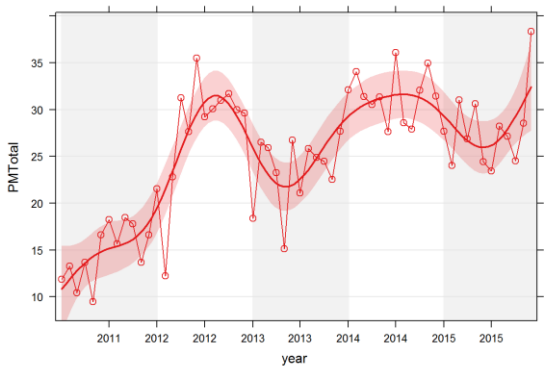


Figura 16. Evolución de la concentración de PM

4. Predicción de entrada de urgencias

4.1 METODOLOGÍA

La metodología que hemos aplicado, se basa en construir un modelo de datos para poder aplicar técnicas de aprendizaje automático dejando que la máquina realice con los subconjuntos del modelo creado, un entrenamiento y posteriormente el test para comprobar el resultado.

Este modelo de datos, consiste en obtener un conjunto de variables independientes y la variable dependiente de estas, como en nuestro caso la variable dependiente trataría de los episodios de urgencias en una semana futura (**EU_1S_fut**) y como las variables independientes serían: Polen (**POL**), Contaminación (**CONT**), Temperatura (**TEMP**), Episodios Urgencias (**EU**).

Estas predicciones se pueden realizar aplicando diferentes modelos de regresión que explicaremos más adelante. En nuestro caso hemos utilizado cuatro:

1- Modelo de referencia
2- Regresión Lineal
3- K-Nearest Neighbors
4- Random Forest.

Para obtener una medida de la calidad de la predicción, debemos comprobar los márgenes de error que se producen en cada método de regresión aplicado.

4.2 MACHINE LEARNING

Ya anteriormente habíamos comentado sobre el concepto de Machine Learning. En este apartado, explicaremos algunas técnicas de **Machine Learning**, donde realizaremos tareas predictivas aplicando diferentes modelos de regresión de las que obtendremos diferentes predicciones.

Añadiendo una definición global sobre lo que es Machine Learning podemos incluir esta:

> *"En ciencias de la computación el aprendizaje automático o aprendizaje de máquinas (del inglés, "Machine Learning") es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos." (Wikipedia, 2016)*

En nuestro estudio utilizamos el aprendizaje automático (Machine Learning) para entrenar a la computadora o máquina que realiza los algoritmos, con un modelo datos dado, donde dependiendo de la técnica de regresión que apliquemos consigue elaborar una fórmula que ayuda a realizar futuras predicciones. Esta fórmula se contrasta posteriormente con unos datos finales, y finalmente verificaremos la exactitud o margen de error de la predicción realizada.

4.2.1 CONSTRUCCIÓN DEL MODELO

El aprendizaje automático requiere de un modelo donde pueda aprender a desarrollar la predicción o encuentre la fórmula que resuelva el objetivo deseado.

En nuestro caso, pretendemos que el aprendizaje automático encuentre en el modelo de datos dado, las predicciones de los ingresos de pacientes asmáticos en urgencias que se producirán en una semana posterior.

En el apartado anterior, explicábamos la extracción y transformación de los datos con los que vamos a trabajar. Del resultado de esa fase, podemos construir el modelo que queremos para poder entrenar la máquina.

El modelo trata de reunir las variables independientes y la variable dependiente. Dado que se pretende hacer una predicción en el tiempo, el modelo debe contar con las fechas en las que se dataron el valor de las variables, en nuestro caso: Año, Mes y Semana.

Para ello, hemos construido el modelo obteniendo los valores de cada año, mes y semana de cada variable. Una variable que añadimos, después haber hecho la fase de exploración y observar el comportamiento de los datos, es la "**Estacion**", esta variable hace referencia a una de las cuatro estaciones del año, por lo que, hemos añadido un valor numérico, dado que solo podemos trabajar con variables cuantitativas, de forma que los valores más altos (4 y 3 pertenecen a las estaciones de Invierno y Primavera respectivamente que son los que registran mayores ingresos).

El modelo que hemos construido queda de la siguiente manera:

| Anyo | Mes | Semana | Estacion | TEMP | POL | CONT | EU| **EU_1S_fut** |

Como vemos, la variable dependiente y que se quiere predecir, es **EU_1S_fut**, que es la cantidad de episodios de urgencias que ocurrirán en una semana posterior.

Una vez construido el modelo, las siguientes fases son, hacer la partición de los datos, de forma que una parte sea para el entrenamiento (Train) del modelo y la otra para la comprobación (Test) que explicaremos a continuación.

Al obtener los dos subconjuntos realizamos los experimentos con los diferentes métodos de regresión y comparamos los resultados.

4.2.2 TRAIN Y TEST

La manera en que aplicamos el aprendizaje automático, consiste en dividir el conjunto de datos que tenemos en dos subconjuntos, uno de ellos servirá para el entrenamiento y otro al cual se aplicará la formula obtenida de dicho entrenamiento para obtener las predicciones. En nuestro caso, nuestro conjunto de datos tiene una fecha de inicio de mediados de 2011 hasta finales de 2015, por lo que, vamos a dividir el conjunto de la siguiente manera creando los dos subconjuntos siguientes:

- **TRAIN**: Desde Junio de 2011 hasta Abril de 2014 (inclusive).
- **TEST**: Desde Mayo de 2014 hasta Diciembre de 2015 (inclusive)

El porcentaje de los subconjuntos quedaría así:

Figura 17. Porcentaje subsets (TRAIN y TEST)

TRAIN servirá para que el modelo de regresión que utilicemos aplique una formula con los valores que ha estimado. Una vez obtenida la formula, esta se aplicará sobre **TEST**, de manera que se puede comparar los resultados de las predicciones con los valores reales.

4.2.3 MODELOS DE REGRESIÓN

Como explicábamos, una vez hecho la separación de los datos en dos subconjuntos (Train y Test), Machine learning aplica las técnicas de la estadística como base fundamental para realizar las predicciones u clasificaciones de forma que obtenga el resultado esperado. En este estudio aplicamos tres modelos de regresión que a continuación vamos a detallar en qué consisten cada uno de ellos.

4.2.3.1 REGRESIÓN LINEAL

Este modelo de regresión es uno de los más sencillos y utilizados en la estadística. En la regresión lineal existe una variable dependiente Y (variable a predecir) que cambia su valor en base a otras variables, variables independientes Xs. La relación entre la variable dependiente y las variables independientes deben estar estrechamente ligada. La fórmula que describe este comportamiento es la siguiente:

$$y = a + \beta x$$

En esta fórmula, la letra Y indica que es la variable dependiente y la variable X la independiente. El coeficiente **β** indica por cuanto aumentará Y por el valor que disponga X. De esta manera valores negativos de X indicarán un incremento negativo o de lo contrario positivo. El coeficiente **α** es el valor donde comienza a crecer la variable Y.

En nuestro caso, la variable Y serán los episodios de urgencia que ingresarán en la semana siguiente o en una semana futura, a la que llamaremos (**EU_1S_fut**).

Las variables Xs serán todas las demás variables, como el POLEN, CONTAMINACION, TEMPERATURA, EU así como sus valores en las tres semanas anteriores.

$$y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_{5\ldots} + \beta_i X_i$$

4.2.3.1.1 MATRIZ DE CORRELACION

Si estudiamos la matriz de correlación, como su nombre indica, nos proporciona una matriz de todas las variables y la relación que existe entre ellas. El grado de relación oscila entre el valor -1 hasta 1, siendo -1 y 1 los valores máximos donde indica el mayor grado de relación que bien puede ser de forma negativa (-1) o de forma positiva (1). Cuanto más cercano se encuentre el valor 0 menor relación existe entre ambas variables, siendo 0 como el valor que indica que no existe relación alguna.

	Estacion	TEMP	POL	CONT	EU	EU_1S_fut
Estacion	1.00	-0.84	0.18	0.33	0.54	0.46
TEMP	-0.84	1.00	0.03	-0.35	-0.54	-0.49

POL	0.18	0.03	1.00	0.17	0.05	0.10
CONT	0.33	-0.35	0.17	1.00	0.08	0.30
EU	0.54	-0.54	0.05	0.08	1.00	0.35
EU_1S_fut	**0.46**	**-0.49**	**0.10**	**0.30**	**0.35**	1.00

Tabla 11. Matriz de correlación

Como podemos ver en la matriz de correlación, la variable dependiente (**EU_1S_fut**) no tiene una fuerte relación ninguna de las variables, aunque sí que guarda una relación algo más estrecha con la TEMPERATURA, la ESTACION y la CONTAMINACION.

Si observamos la variable EU, vemos como la relación entre las demás variables es menos significativa aproximándose más aún a 0, salvo por la TEMPERATURA y ESTACION.

El valor negativo de la TEMP indica que cuanto menor sea este valor más aumentarán los episodios de urgencias en la siguiente semana (**EU_1S_fut**)

4.2.3.2 K-NEAREST NEIGHBORS (K-VECINOS MÁS CERCANOS)

A diferencia del método de regresión lineal, el algoritmo de K-Nearest Neighbors (K-NN) no compara todas las observaciones del conjunto de datos sino que realiza una clasificación de ellas y las compara, por lo que es utilizado tanto para la clasificación como para la predicción. Este algoritmo consiste en ubicar el dato a predecir o clasificar comparándolo con aquellos datos que más se le asemejen o se acerquen.

En nuestro caso la variable a predecir es **EU_1S_fut**. El algoritmo trata de buscar en los N casos más cercanos de las variables que más se le parezcan o se le acerquen. Por ejemplo, para el caso de N=10, buscará los 10 valores más próximos al valor a predecir y obtendrá una media de ellos. Otro ejemplo, es N=1, donde buscará el valor más próximo o cercano asignándole el valor de este.

La complejidad del algoritmo vendrá determinado por el coeficiente K, donde si es K = 1, calculará la media de las N observaciones más cercanas y si es K=N calculará 1 sola media, ya que N es el total de toda la muestra.

Nosotros podemos configurar el valor de N la cantidad de observaciones cercanas debe compararse la variable a predecir.

4.2.3.3 RANDOM FOREST

El algoritmo de Random Forest mejora la precisión en la clasificación particionando el espacio en arboles de decisión construido por observaciones y variables aleatorias.

El proceso que realiza trata de seleccionar individuos al azar (usando muestreo con reemplazo) creando diferentes conjuntos de datos. Por cada conjunto de datos, construye un árbol de decisión, donde una entrada se introduce en el nodo superior y, hacia abajo, a medida que atraviesa el árbol de los datos se acumulan en conjuntos más y más pequeños, consiguiendo crecer el árbol y crear diferentes arboles con variables distintas.

En la construcción de los árboles, se eligen las variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad. Por último, las predicciones de los nuevos datos las realizará usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.

Para nuestro caso, el algoritmo construirá un número determinado de Árboles, donde la muestra será el conjunto de datos de entrenamiento (train) con un reemplazo para crear un subconjunto de los datos. El subconjunto será el 66% del conjunto total.

Para cada nodo, obtendrá un valor con el número de variables independientes seleccionadas al azar. Mediante una función objetiva se escoge el valor de predicción de **EU_1S_fut** que proporciona la mejor división, y se emplea para realizar la división binaria en ese nodo. El siguiente nodo, vuelve a repetir la operación, escogiendo un valor con otro número de variables independientes y repite el mismo paso.

4.3 EVALUACIÓN DEL MODELO DE REGRESIÓN: MAE Y MSE

Para conocer la validez o precisión de las predicciones o resultados, uno de los cálculos que se suelen utilizar son los errores **MAE** y **MSE.**

En cualquier experimento que se realice y su número de ensayos, podemos encontrarnos con valores diferentes, es decir, en un ensayo podemos encontrarnos unos valores y en el siguiente otros diferentes.

En el caso de estudio que nos ocupa, la medición que realizamos por semana, incluye diferentes variables y cada una de ellas un valor, lo que diferencia una semana de otra, es decir, para la semana 23 de 2013, por ejemplo, las variables tendrán unos valores y si realizamos la predicción para esa semana será diferente a la de la semana siguiente, por lo que en cada predicción, tendremos unos valores diferentes. Entonces, ¿cómo averiguamos el margen de error en estos cálculos? Para estimar en regresión la calidad de un modelo se suele calcular la diferencia entre las predicciones del modelo y los valores reales que se tienen.

Una de los cálculos que se realizan es **MAE**, y consiste en calcular la diferencia entre los valores predecidos y los valores reales, obteniendo el valor absoluto de la diferencia. Por supuesto, volviendo a lo comentado anteriormente, si tenemos N observaciones, y en cada observación valores distintos en las variables, cada predicción tendrá un valor. Para saber el valor total del error que tiene el modelo, haríamos la suma de cada error y la dividiríamos por el número de predicciones / observaciones, o lo que es lo mismo, obtendríamos la media de todos los errores de las predicciones realizadas. Veamos la fórmula matemática que refleja esto:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(predicción(i) - valor\ real(i))|$$

El cálculo hecho con **MAE** es útil ya que trata de manera igual todas las diferencias, y aquellas diferencias que sean poco significativas no tendrán relevancia en el cálculo, o dicho de otra manera, en una distribución gaussiana se centra más en la mediana de los datos. Pero, ¿Y si queremos obtener un error donde queramos profundizar en errores más extremos?, entonces otro calcula para para medir el margen de error es **MSE**.

El **MSE** es similar al **MAE**, pero su cálculo se realiza elevando al cuadrado la diferencia entre los valores predecidos y reales y obteniendo la media de los cálculos de las N observaciones. Es decir, calculamos el cuadrado de la diferencia de ambos valores (predecidos y reales) de cada observación, sumamos todos los valores y dividimos por el número de observaciones (**N**). La fórmula quedaría así:

$$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(predicción(i) - valor\ real(i)\right)^2$$

En ocasiones, la suma de los cuadrados puede distorsionar la magnitud real de los errores, y para ello una variación de la formula **MSE**, es el cálculo de la raíz cuadrada en su valor final (**RMSE**). Veamos la fórmula:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(predicción(i) - valor\ real(i)\right)^2}$$

Teniendo una variedad amplia de medidas que podemos escoger, además de las presentadas, escoger la medida adecuada para una determinada situación o experimento no es trivial.

4.4 EXPERIMENTOS

Los experimentos realizados se realizan probando diferentes modelos de regresión. En cualquier predicción, existe siempre un margen de error, que ayuda a conocer cuánto de fiable son los valores que se predicen.

En este apartado expondremos los resultados de cada modelo aplicado y los márgenes de errores que han obtenido. De esta forma podremos comparar la exactitud de las predicciones y que modelo es el más adecuado.

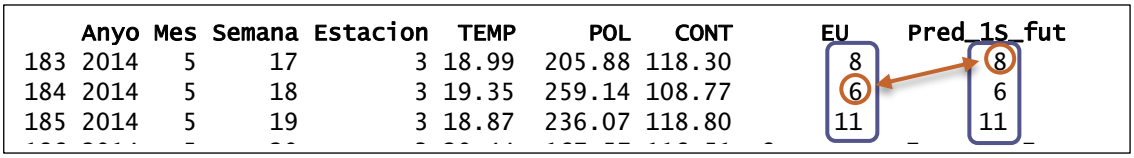
Para realizar estos experimentos, hemos ido calculando sobre el modelo inicial y probado incluir nuevas variables, como son hasta la 3 semana anterior, para realmente observar cómo se comporta el modelo y cuanto afecta a las predicciones estas nuevas variables.

Los márgenes de error que hemos utilizado y calculado son, **MAE** y **MSE**.

4.4.1 MODELO DE REFERENCIA

Para aplicar un modelo base y compararlo respecto a los otros modelos de regresión y sus resultados, hemos tomado un modelo de referencia.

Este modelo de referencia consiste en aplicar el modelo de datos que hemos cogido pero tomando como valor de la predicción de la semana siguiente, los ingresos de la semana en curso (**EU**). Por ejemplo, vamos a suponer que en la próxima semana, los ingresos serán el mismo valor que tiene (**EU**) de la semana en curso:



	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	8
184	2014	5	18	3	19.35	259.14	108.77	6	6
185	2014	5	19	3	18.87	236.07	118.80	11	11

De esta manera si hacemos el cálculo de los errores será:

MAE =(|V. Real (**EU** (**semana siguiente**) – Predicción (**Pred_1S_fut** (**semana en curso**))|)

MSE =(V. Real (**EU** (**semana siguiente**) – Predicción (**Pred_1S_fut** (**semana en curso**)))^2

Vamos a ver los resultados aplicando el modelo de referencia que hemos propuesto:

	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	8	8
184	2014	5	18	3	19.35	259.14	108.77	6	6	6
185	2014	5	19	3	18.87	236.07	118.80	11	11	11
186	2014	5	20	3	20.44	167.57	116.51	8	8	8
187	2014	5	21	3	19.88	111.67	109.13	7	7	7
188	2014	6	21	1	19.45	193.00	127.00	3	3	3
189	2014	6	22	1	21.62	135.86	104.34	8	8	8
190	2014	6	23	1	24.68	67.93	107.94	5	5	5
191	2014	6	24	1	23.10	36.07	108.66	5	5	5

Tabla 12. Resultado Mod.Referencia

Si observamos los errores que tenemos con este modelo de referencia tenemos:

MAE	MSE
5.45	47.45

4.4.2 Regresión Lineal

Como habíamos comentado en el anterior apartado sobre la regresión lineal, la matriz de correlación no presenta una fuerte relación entre las variables independientes y la variable a predecir. Esto se hace presente cuando realizamos las predicciones con todas las variables, como son la temperatura, contaminación, polen y episodios de urgencias.

Realizando regresión lineal sobre el modelo de datos, tenemos:

	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	6	8
184	2014	5	18	3	19.35	259.14	108.77	6	11	7
185	2014	5	19	3	18.87	236.07	118.80	11	8	8
186	2014	5	20	3	20.44	167.57	116.51	8	7	7
187	2014	5	21	3	19.88	111.67	109.13	7	3	7
188	2014	6	21	1	19.45	193.00	127.00	3	8	7
189	2014	6	22	1	21.62	135.86	104.34	8	5	7
190	2014	6	23	1	24.68	67.93	107.94	5	5	5
191	2014	6	24	1	23.10	36.07	108.66	5	10	6
192	2014	6	25	1	23.44	36.57	96.97	10	2	6
193	2014	6	26	1	23.05	46.50	65.40	2	6	6
194	2014	7	26	1	24.98	12.75	108.67	6	4	5
195	2014	7	27	1	23.36	25.29	99.17	4	5	6
196	2014	7	28	1	26.18	14.79	85.14	5	12	5

Tabla 13. Resultado Reg.Lineal

Como vemos en el resultado la predicción no es muy acertada, si bien observamos los errores que existen entre las predicciones y los valores reales vemos que estos mejoran ligeramente respecto al modelo de referencia.

MAE	MSE
5	44.65

Si incluimos en la predicción una variable más, modificando el modelo, añadiendo la semana anterior el error que obtenemos es mayor:

	Anyo	Mes	Semana	Estacion	TEMP_1S_Ant	TEMP	POL_1S_Ant	POL	CONT_1S_Ant	CONT	EU_1S_Ant	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.65	18.99	213.33	205.88	104.07	118.30	1	8	6	8
184	2014	5	18	3	18.99	19.35	205.88	259.14	118.30	108.77	8	6	11	8
185	2014	5	19	3	19.35	18.87	259.14	236.07	108.77	118.80	6	11	8	7
186	2014	5	20	3	18.87	20.44	236.07	167.57	118.80	116.51	11	8	7	5
187	2014	5	21	3	20.44	19.88	167.57	111.67	116.51	109.13	8	7	3	6
188	2014	6	21	1	19.88	19.45	111.67	193.00	109.13	127.00	7	3	8	8
189	2014	6	22	1	19.45	21.62	193.00	135.86	127.00	104.34	3	8	5	6
190	2014	6	23	1	21.62	24.68	135.86	67.93	104.34	107.94	8	5	5	4
191	2014	6	24	1	24.68	23.10	67.93	36.07	107.94	108.66	5	5	10	6

Tabla 14. Resultado Reg.Lineal + 1 vble.

MAE	MSE
5.44	52.8

Si incluimos más variables como son las tres semanas anteriores de cada variable, el error aumenta:

MAE	MSE
6.36	72.75

Como podemos observar los errores son elevados, y cuanto más variables incluimos mayor es el error en la predicción. Vemos en el último caso como tenemos un error mayor de 6 de media como diferencia entre los valores reales y predecidos y si elevamos esta diferencia al cuadrado y obtenemos la media el error es mucho más elevado.

4.4.3 K-NEAREST NEIGHBORS

En este modelo un factor que afecta es el valor de N, que son las N observaciones más cercanas a las variables independientes y por tanto realiza una predicción obteniendo la media de los valores más cercanos o del valor más próximo.
Realizando varias pruebas con el modelo, vamos a mostrar la diferencia en las predicciones y los errores que hay cambiando el valor de N.

En el primer caso, y considerando las variables sin contar con ninguna semana anterior, pondremos **N = 1**, que significa que el modelo buscará la observación más próxima a las variables independientes y por tanto buscará el valor de **EU_1S_fut** para incluir como predicción.

Si aplicamos este modelo sobre los datos el resultado es:
Para **N = 1**:

	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	6	10
184	2014	5	18	3	19.35	259.14	108.77	6	11	10
185	2014	5	19	3	18.87	236.07	118.80	11	8	1
186	2014	5	20	3	20.44	167.57	116.51	8	7	1
187	2014	5	21	3	19.88	111.67	109.13	7	3	6
188	2014	6	21	1	19.45	193.00	127.00	3	8	10
189	2014	6	22	1	21.62	135.86	104.34	8	5	1
190	2014	6	23	1	24.68	67.93	107.94	5	5	10
191	2014	6	24	1	23.10	36.07	108.66	5	10	10
192	2014	6	25	1	23.44	36.57	96.97	10	2	1
193	2014	6	26	1	23.05	46.50	65.40	2	6	8
194	2014	7	26	1	24.98	12.75	108.67	6	4	10

Tabla 15. Resultados K-NN (N=1)

Como vemos la predicción no es muy exacta:

MAE	MSE
6.53	71.93

Esto significa que encontrar una observación que sea aproxime al valor que se quiere predecir tiene un margen de error alto. Si incluimos más observaciones cercanas al valor que queremos predecir **EU_1S_fut**, podemos comparar los resultados.

Realizando la misma prueba sobre el modelo de datos con más observaciones, tenemos:

Para **N = 10**:

	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	6	7
184	2014	5	18	3	19.35	259.14	108.77	6	11	7
185	2014	5	19	3	18.87	236.07	118.80	11	8	7
186	2014	5	20	3	20.44	167.57	116.51	8	7	6
187	2014	5	21	3	19.88	111.67	109.13	7	3	7
188	2014	6	21	1	19.45	193.00	127.00	3	8	7
189	2014	6	22	1	21.62	135.86	104.34	8	5	7
190	2014	6	23	1	24.68	67.93	107.94	5	5	6
191	2014	6	24	1	23.10	36.07	108.66	5	10	6
192	2014	6	25	1	23.44	36.57	96.97	10	2	7
193	2014	6	26	1	23.05	46.50	65.40	2	6	5
194	2014	7	26	1	24.98	12.75	108.67	6	4	7

Tabla 16. Resultados K-NN (N=10)

Aplicando más observaciones, disminuye el error, debido a que de entre todas las observaciones realiza una media, lo que hace que se aproxime más a la observación. Comparándolo con la anterior prueba (**N=1**), en el que error, daba un valor mayor de 6, el margen es menor pero sigue existiendo un error elevado, como podemos ver en los MAE y MSE, y no podríamos considerarla como una buena predicción:

MAE	MSE
5.26	46.99

Si añadimos más variables en el modelo de datos, como las semanas anteriores, el error es ligeramente menor, aunque sigue siendo un margen de error considerado, por ejemplo si incluimos una semana anterior:

MAE	MSE
5.26	47.05

Prácticamente el error no ha cambiado relativamente, veamos que sucede si incluimos hasta la tercera semana anterior:

MAE	MSE
5.38	49.57

Como podemos ver, este modelo de regresión no es tampoco indicado para hacer una buena predicción sobre la variable **EU_1S_fut.**

4.4.4 RANDOM FORESTS

En el modelo de regresión Random Forest vamos a aplicar las mismas pruebas que en los anteriores modelos y comparar sus diferencias.

El resultado del modelo aplicado a los datos es:

	Anyo	Mes	Semana	Estacion	TEMP	POL	CONT	EU	EU_1S_fut	Pred_1S_fut
183	2014	5	17	3	18.99	205.88	118.30	8	6	5
184	2014	5	18	3	19.35	259.14	108.77	6	11	6
185	2014	5	19	3	18.87	236.07	118.80	11	8	5
186	2014	5	20	3	20.44	167.57	116.51	8	7	6
187	2014	5	21	3	19.88	111.67	109.13	7	3	6
188	2014	6	21	1	19.45	193.00	127.00	3	8	6
189	2014	6	22	1	21.62	135.86	104.34	8	5	7
190	2014	6	23	1	24.68	67.93	107.94	5	5	6
191	2014	6	24	1	23.10	36.07	108.66	5	10	6
192	2014	6	25	1	23.44	36.57	96.97	10	2	7
193	2014	6	26	1	23.05	46.50	65.40	2	6	5
194	2014	7	26	1	24.98	12.75	108.67	6	4	6

Tabla 17. Resultados R.Forest

El error que nos muestra es el siguiente, muy similar:

MAE	MSE
4.8	38.64

Podemos apreciar como el error que nos da entre las predicciones y los valores reales es sin duda mucho menor que en los anteriores modelos.

Si añadimos más variables al modelo, veremos que añadiendo una semana anterior como variable el error en la predicción es mayor:

MAE	MSE
5.07	43.01

Aun así, el error sigue siendo un poco mejor que en los anteriores casos. Y si añadimos todas las variables incluyendo hasta la 3 semana anterior, el margen de error que obtenemos es:

MAE	MSE
5.52	51.35

Sin duda, el modelo de regresión Random Forest tampoco realiza una predicción fiable, ya que sus márgenes de error son también altos.

4.5 COMPARACIÓN DE RESULTADOS

Para llegar a una conclusión sobre las pruebas realizadas en los modelos, comparamos los errores y sus resultados.

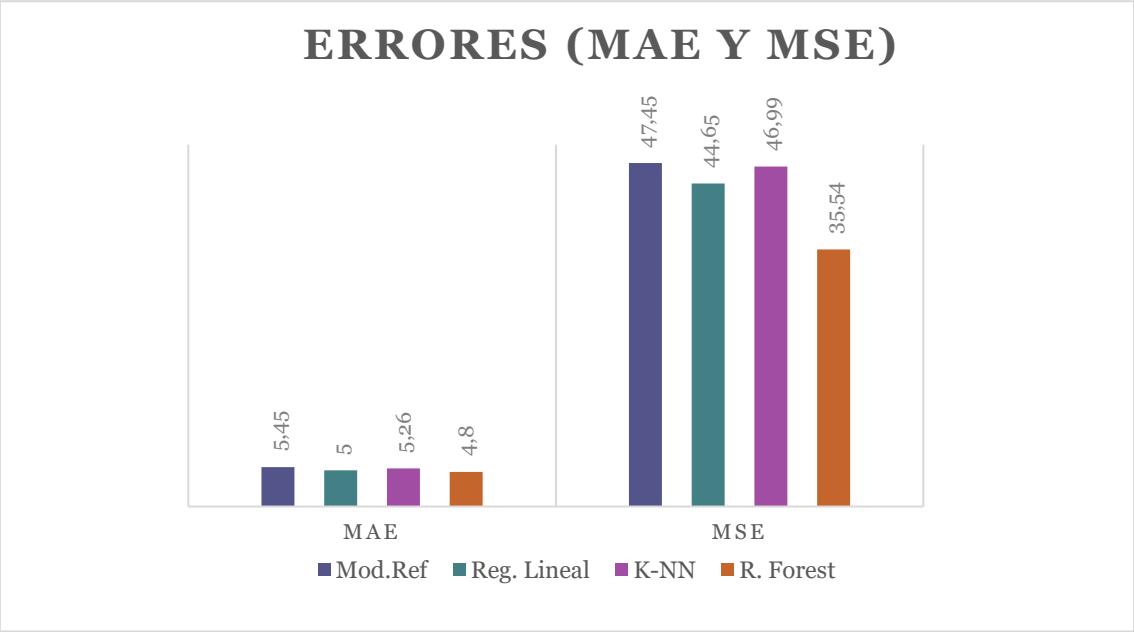


Figura 18. Comparación de errores entre modelos

Comparamos los márgenes de error de los modelos añadiendo nuevas variables (el modelo de referencia no aplica en este caso ya que tiene siempre las mismas predicciones)

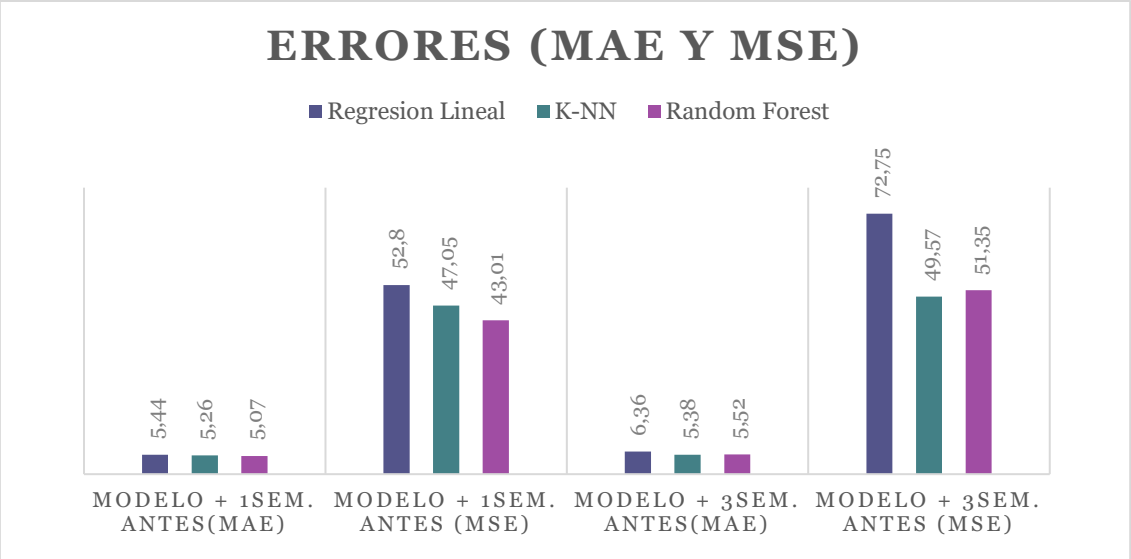


Figura 19. Comparación de errores con varias variables

Observamos la gráfica en las predicciones con el modelo de datos, vemos que las predicciones en todos los modelos de regresión se alejan bastante de los valores reales que debían predecir.

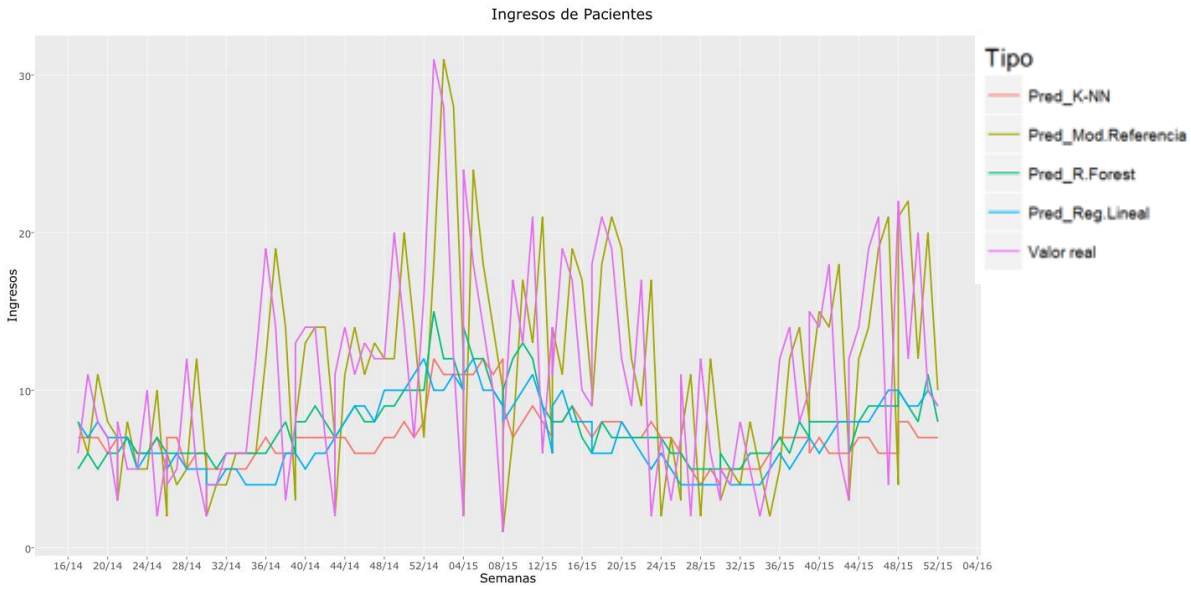


Figura 20. Comparación de predicciones entre modelos

En general no hay ningún modelo de los aplicados que pueda hacer una predicción fiable ya que en todos los errores suelen ser muy similar, con un valor MAE alrededor de 5, aunque, si bien podemos pensar que de los tres, si observamos los errores MAE, aunque prácticamente no presentan mucha diferencia entre los modelos, Random Forest tiene menos error. El error cuadrático es menor en el modelo Random Forest con mayor diferencia.

4.6 CONCLUSIÓN

Después de realizar este estudio, hemos visto y comparado varios modelos de regresión de los cuales no podemos asegurar con una gran fiabilidad las predicciones realizadas de ninguno de ellos.

Como hemos visto, los modelos de regresión aplicados mejoran respecto al modelo de referencia, siendo el modelo Random Forest quien mejor resultados y menores errores muestra. En el resto de modelos, los errores tan elevados y sus predicciones son peores ya que, como veíamos la correlación no es muy fuerte entre las variables.

Si pensamos en las variables que incluye el modelo, sabemos que si incluimos un histórico como son las semanas anteriores de cada una de sus variables, incrementa el margen de error lo cual, vemos que la correlación entre variables es muy baja y al añadir más variables las predicciones son cada vez peores.

Como conclusión, hemos podido saber que la predicción sobre ASMA no es trivial, ya que existen, con probabilidad más variables que no hayamos tenido en cuenta. El ASMA en cada individuo se manifiesta por diferentes razones y en diferentes grados, por lo que lleva a pensar que existen dos tipos de factores, a grandes rasgos, que son: factores externos, como los estudiados en este trabajo y factores internos, como las características individuales de cada paciente, como la edad, peso, fumador, etc.

Se espera que este estudio sirva como punto de partida para posteriores estudios donde, con lo recogido en este trabajo, puedan desarrollarse nuevas soluciones que mejoren la asistencia recibida a estos pacientes.

5. Bibliografía

(OMS), Organización Mundial de la Salud. Organización Mundial de la Salud (OMS). [En línea] http://www.who.int/mediacentre/factsheets/fs313/es/.

Ali Azari, Vandana P.Janeja, Alex Mohseni. 2012. *Healthcare Data Mining: Predicting Hospital Length of Stay (PHLOS).* Baltimore, USA : Library & Information Science Abstracts (LISA), 2012.

Arévalo, Edwin. 2013. Random Forest. [En línea] 5 de 2013. http://randomforest2013.blogspot.com.es/2013/05/randomforest-definicion-random-forests.html.

César Pérez López, Daniel Santín González. 2007. *Minería de datos, Técnicas y herramientas.* Madrid : Thomson Editores Spain Paraninfo, S.A., 2007.

Comité de Salud y Medio ambiente Soc. Europ Enfermedades Respiratorias. *La contaminacion del aire y los pulmones.* [PDF]

2015. El poder del Big Data: ¿Puede Twitter ayudar a predecir un aumento de las visitas a Urgencias? [En línea] 22 de 04 de 2015. http://prnoticias.com/salud/20140610-big-data-salud-urgencias-hospital?tmpl=component&print=1.

El triaje: herramienta fundamental en urgencias y emergencias. **W. Soler, M.Gómez Muñoz, E.Bragulat, A.Álvarez. 2010.** 2010, Anales del Sistema Sanitario de Navarra, pág. 14.

Estudio de la mortalidad por asma bronquial. **Claudia Roche Albemas, Kenia González Valcárcel, Lumey Hernández Niebla, Raisel García Pérez. 2011.** 3, Santa Clara, Villa Clara, Cuba. : s.n., 2011, Vol. 5.

J.Hernández Orallo, M. Ramírez Quintana y C.Ferri Ramírez. 2004. *Introducción a la minería de datos.* Madrid : Pearson educación, S.A., 2004. 978-84-205-4091-7.

Jaynal, Abedin y Kumar Das, Kishor. 2014. *Data Manipulation with R.* Birmingham : Packt Publishing Lt, 2014.

Lantz, Brett. 2015. *Machine Learning with R.* Birmingham : Packt Publishing Ltd, 2015.

Obenshain, Mary K. 2004. *Application of Data Mining Techniques to Healthcare Data.* 2004. 690.

Pérez Marqués, María. 2014. *Minería de datos a través de ejemplos.* Madrid : RC Libros, 2014.

S. Ram, W. Zhang, M. Williams y Y. Pengetnze. 2015. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE Xplore.* [En línea] Julio de 2015. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7045443.

Santana, Emmanuel. 2014. Ejemplo de random forest. [En línea] 11 de 2014. http://apuntes-r.blogspot.com.es/2014/11/ejemplo-de-random-forest.html.

Urgencias y emergencias: al servicio del ciudadano. **J. Sesma, O.Miró. 2010.** 2010, Analisis del Sistema Sanitario en Navarra, Vols. Vol. 33, Suplemento 1.

Wikipedia. 2016. Aprendizaje_automático. [En línea] 01 de 08 de 2016. https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico.

6. Índice de tablas y figuras