E.T.S. de Ingeniería Industrial,
Informática y de Telecomunicación

Elaboración de un sistema inteligente para la predicción del número de visitas a urgencias en base a factores ambientales

Grado en Ingeniería Informática

Trabajo Fin de Grado

Alumno: Adrián Errea López

Director: Mikel Galar Idoate

Pamplona, 19 de marzo de 2018

30 1987-2017 upna

RESUMEN

Tras la digitalización de la información sanitaria en bases de datos se han llevado a cabo numerosos proyectos de interés sobre esta información debido a su gran valor e interés tanto público como para los especialistas sanitarios.

En esta ocasión, el doble objetivo del proyecto es trabajar sobre esos datos con el fin de intentar observar una correlación entre las temperaturas y las visitas a urgencias en personas de avanzada edad en Navarra. Por otra parte, realizar una predicción a nivel de paciente intentando predecir si un paciente irá a urgencias en base a variables demográficas y clínicas y realizar una predicción del número de visitas a urgencias para un día determinado en base a variables atmosféricas del día a día. Se han utilizado dos datasets: pacientes mayores o iguales de 65 años al inicio del estudio (2013) con zona de domicilio Pamplona y comarca y las visitas a urgencias durante cuatro años (2013-2016) en el Complejo Hospitalario de Navarra (CHN). Y por otra parte pacientes mayores o iguales de 16 años al inicio del estudio (2013) con zona de domicilio Pamplona y comarca y las visitas a urgencias durante el mismo periodo en el CHN (2013-2016).

PALABRAS CLAVE

Base de Datos, Predicción, Clasificación, Minería de Datos, Regresión, Salud, Urgencias, LSTM.

Contenido

1. Introducción

Actualmente se conocen una serie de factores tanto externos como internos que afectan a la salud de cualquier persona. La OMS recomienda llevar una dieta saludable, realizar ejercicio físico y tener unos hábitos beneficiosos y sostenibles en el tiempo. Está más que estudiado que estos factores que cada persona puede controlar de forma individual son de vital importancia a la hora de prevenir enfermedades cardiovasculares, respiratorias y en general a poder mantener alejadas las enfermedades en la mayor medida posible.

Sin embargo, hay una serie de factores ajenos a las personas que tienen una importancia considerable a la hora de analizar las causas de estas enfermedades. Entre ellas se encuentran la edad, el sexo, factores genéticos y factores ambientales como son la contaminación atmosférica y la temperatura ambiental, entre otras. Se han realizado multitud de estudios en base a este último factor con diferentes resultados. En 2016 se llevó a cabo un estudio en Hong Kong en donde se relacionaban las bajas temperaturas con los ingresos por enfermedad cardiovascular ([1]: Linwei Tian, et al*., 2016)*. De la misma forma otro estudio llevado a cabo en Cataluña, observó también un incremento con olas de frío en los ingresos cardiovasculares ([2]: Anna Ponjoan, et al., 2017). Además, hay estudios, como el realizado en Atlanta ([3]: Tianqi Chen, et al., 2017) que encuentran un aumento en el riesgo de ir a urgencias en olas de calor. Por tanto podemos obtener una premisa clara: las temperaturas extremas (tanto muy bajas como muy altas) tienen claras repercusiones en la salud de las personas. Estas repercusiones son acentuadas en personas de avanzada edad ya que su sistema inmunológico es más débil y además han estado expuestos a las temperaturas extremas durante un mayor número de años.

Para poder obtener una correlación más ajustada a la comunidad foral de Navarra se ha realizado un estudio para ver qué tipo de temperaturas afectan más a la población navarra. Hay que tener en cuenta que estos valores se han obtenido de diferentes estaciones meteorológicas repartidas a lo largo de todo el territorio y que a cada paciente del SNS-O se le ha asignado la estación más cercana a su domicilio. Como población para este estudio se han escogido dos conjuntos de pacientes: aquellas personas residentes en Pamplona y comarca con cupo en enero de 2012 y mayores o iguales de 65 años a enero de 2013. Los datos para de este conjunto van desde enero de 2013 hasta diciembre de 2016. El otro conjunto está compuesto por aquellas personas residentes en Pamplona y comarca con cupo en enero de 2012 y mayores o iguales de 16 años a enero de 2013. Los datos en este caso también van desde enero de 2013 hasta diciembre de 2016. Cabe resaltar que existen muchos factores secundarios que influyen en los resultados como por ejemplo el tiempo que pasa cada persona expuesta a dichas temperaturas, su nivel económico que influye en la temperatura del hogar, etc. pero que no pueden ser recogidas de forma sencilla.

Para poder llevar a cabo el estudio se ha utilizado una regresión de Poisson en base a Riesgos Relativos (RR). Se llegó a esta decisión debido a que en los estudios mencionados anteriormente del mismo carácter se llevaba a cabo esta metodología y parecía que arrojaba buenos resultados. Este apartado será explicado más adelante en la *Sección 5: Estudio estadístico*.

Por otra parte, esta influencia de la temperatura está relacionada directamente con las visitas a urgencias como se ha visto en los anteriores estudios y sería de gran valor, por tanto, el poder saber cuánta gente va a visitar el servicio de urgencias en base a la temperatura de cierto día. Con ello, se podría prever con bastante exactitud el número de gente que requiere el servicio cierto día en concreto y con ello asignar el personal y material clínico necesario. De esta forma, es más sencillo saber el volumen de trabajo cada día y el trabajo que va a suponer para el servicio, ahorrando costes y dando un mejor servicio al paciente.

Para poder dar una posible solución a la cuestión mencionada se han realizado dos tipos de predicciones:

I. A nivel de paciente: En este caso en base a variables demográficas de la persona (Edad, Sexo, TSI, GMA, Dependencia, etc.), a variables clínicas o de riesgo de la persona (Número de visitas a urgencias previas, Número de ingresos urgentes previos, Número de medicamentos dispensados en receta, etc.) y a variables atmosféricas (Temperaturas, Gripe, etc.) se tratará de predecir si un paciente en concreto va a ir a urgencias un día determinado.

II. A nivel de día: Mediante técnicas basadas en series temporales se trata de predecir el número de visitas a urgencias un día en concreto en base a variables a nivel de día: Ola de calor, ola de frío, festivo, epidemia de gripe, etc.

Se explicará con detalle en las *Secciones 6: Predicción a nivel de paciente y 7: Predicción a partir de series temporales* correspondientes así como la metodología utilizada.

2. Fuentes de datos

En este apartado se explicará de dónde se han obtenido los datos para el proyecto tanto para los datos clínicos (*Sección 2.1*) como para los datos meteorológicos (*Sección 2.2*).

2.1. Datos clínicos (Servicio Navarro de Salud-Osasunbidea)

El Servicio Navarro de Salud-Osasunbidea (SNS-O) posee diferentes bases de datos dónde se guarda información de cada sistema de información. Además el Servicio posee múltiples aplicaciones usadas por los diferentes clínicos y personal administrativo del Departamento de Salud. En el *Anexo A* se encuentra una estructuración del Departamento de Salud que puede ayudar a la comprensión de esta parte. Estas aplicaciones son:

I. ATENEA

   Es la principal herramienta utilizada por los profesionales de atención primaria (AP). Muestra la historia clínica, situación médica y datos del paciente para que el clínico tenga una vista general del paciente.

II. LAKORA-TIS

   Se gestionan todos los datos demográficos de las personas que son beneficiarias del Servicio Navarro de Salud. Entre estos datos se encuentran la fecha de nacimiento, residencia, cotización de farmacia, etc.

III. LAMIA

   Se utiliza para prescribir el tratamiento ambulatorio (no hospitalario) de los pacientes, tanto de atención primaria (AP) y de atención especializada (AE). De esta forma se dispone del historial farmacoterapéutico del paciente, independientemente del profesional que prescriba el tratamiento.

IV. HCI

   Es la herramienta utilizada por los profesionales de atención especializada (AE). Se documentan exploraciones realizadas al paciente, información de algún ámbito concreto, pruebas de laboratorio, radiologías, etc.

V. HIS-LEIRE

   Esta herramienta es utilizada para la gestión de la parte administrativa (no clínica) de toda la atención especializada (AE) y hospitalaria. Entre otros se tiene registro de urgencias hospitalarias, ingresos, hospital de día, citaciones, listas de espera, etc.

VI. FARHO

Es la herramienta para prescribir tratamientos farmacológicos durante un ingreso y actúa de forma separada a LAMIA.

Tras el proceso de digitalización de la información que se realizó para pasar toda la información de papel a digital, actualmente se está trabajando en tener toda la información recopilada e integrada de forma correcta. Como resultado se creó la base de datos poblacional que es desde dónde se ha extraído toda la información para el proyecto. Esta base de datos es una integración de las bases de datos individuales de las aplicaciones anteriormente mencionadas. La información contiene valores a partir del año 2012 y está siendo construida y mantenida por el equipo de trabajo del SSIAS (Servicio de Sistemas de Información del Área Sanitaria).

2.2. Datos meteorológicos (Gobierno de Navarra)

Para la extracción de las temperaturas ambientales se utilizaron los archivos que guarda el Gobierno de Navarra en el portal web “meteo.navarra.es”. Para los datos atmosféricos se extrajo la información del portal “gobiernoabierto.navarra.es” que también pertenece al Gobierno de Navarra.

También se contactó mediante correo electrónico con Tragsatec (Tecnologías y Servicios Agrarios, S.A.) para poder obtener algún valor sobre la presión atmosférica. Finalmente debido a que había una falta de datos considerable, no se llegaron a utilizar.

3. Extracción y limpieza de datos

En este apartado se presentará el proceso de extracción llevado a cabo para obtener información de valor con la que poder trabajar. En primer lugar, en la *Sección 3.1* se describirá el proceso llevado a cabo en cuanto a los datos ambientales tales como temperaturas *(Sección 3.1.1*) y gripe (*Sección 3.1.2*). Además en la *Sección 3.2*, se hablará de cómo se realizó la extracción de los datos clínicos y demográficos. Esto es, domicilios (*Sección 3.2.1*), pacientes (*Sección 3.2.2*), una pequeña explicación de la estructuración de las bases de datos en Salud (*Sección 3.2.3*), fuentes de datos clínicos (*Sección 3.2.4*) y la limpieza de datos llevada a cabo (*Sección 3.2.5*). Finalmente, se expondrá un esquema de la base de datos creada (*Sección 3.3*).

3.1.Datos ambientales

3.1.1. Temperaturas

Para poder realizar el estudio con las variables que afectan a una persona se pensó en obtener las temperaturas ambientales, humedad y temperaturas atmosféricas. En un principio se empezó la extracción de datos del portal web "gobiernoabierto.navarra.es" ya mencionado. Los datos que se guardan en este portal van desde el año 2011 a la actualidad. Sin embargo, hubo varios inconvenientes que hicieron decantar la balanza hacia la no inclusión de los datos de la humedad y las temperaturas atmosféricas:

I. Escasez de estaciones que midan estos datos (9 estaciones en todo el territorio): Leitza, Alsasua, Rotxapea, Plaza de la Cruz, Iturrama, Sangüesa, Olite, Funes y Tudela (Ver *Figura 3.1.a*)

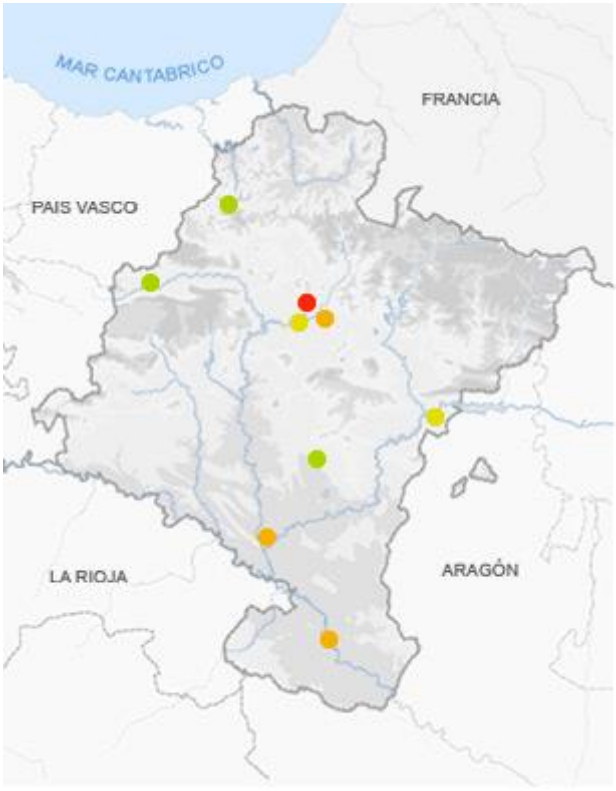


Figura 3.1.a: Estaciones que miden temperaturas ambientales y humedad. (Fuente: Situación de estaciones. En navarra.es [4])

II. Falta de consistencia de esos datos debido a la inclusión y eliminación de diferentes estaciones en los últimos años (En 2015 se instala la estación de Leitza, la estación de Arguedas (no aparece en la *Figura 3.1.a*) fue desmantelada en el año 2013 y la estación de Lesaka (no aparece en la *Figura 3.1.a*) fue desmantelada en el año 2013).

III. Falta de datos en las estaciones existentes. Multitud de datos de prácticamente todas las estaciones están incompletos o con valores erróneos. Esto puede ser debido al poco tiempo que llevan estas estaciones en funcionamiento y el mantenimiento que llevan. Por tanto se tomó la decisión de no usarlos en el proyecto.

En cuanto a las temperaturas ambientales, se extrajeron los datos de la página web meteo.navarra.es que pertenece al servicio meteorológico del Gobierno de Navarra. En este portal web se muestran las estaciones que están repartidas a lo largo de la comunidad foral.

Como se puede observar en la *Figura 3.1.b* existen dos tipos de estaciones en Navarra: Automática y Manuales.

Las automáticas miden: Temperatura máxima, Temperatura mínima, Temperatura media, Humedad relativa media, Humedad relativa máxima, Humedad relativa mínima, Precipitación, Radiación global, Velocidad media viento y Dirección del viento. Se guarda un registro por día.

Las manuales solo miden Temperatura máxima, Temperatura mínima y Precipitación. Se guarda también un registro por día.

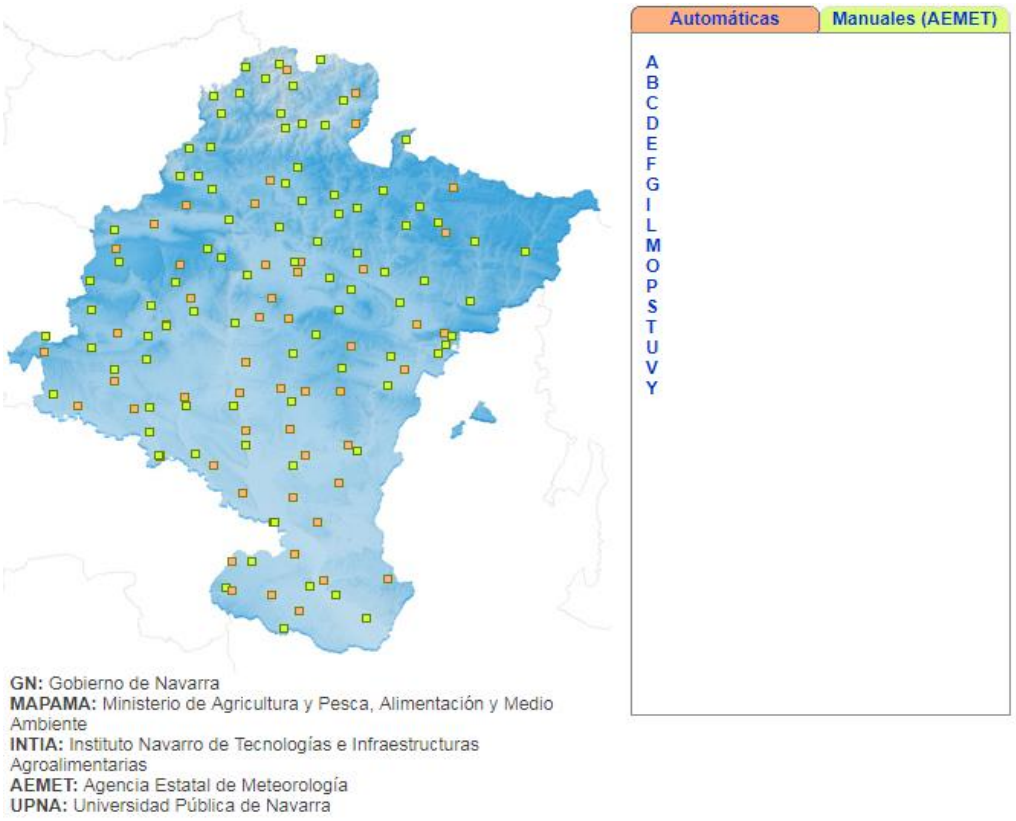


Figura 3.1.b: mapa con las estaciones automáticas y manuales recogidas en el portal "meteo.navarra.es". (Fuente: Mapa de estaciones. En meteo.navarra.es [5])

Cada estación meteorológica pertenece a un organismo y se identifican de la siguiente manera:

- GN: Gobierno de Navarra
- MAPAMA: Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente
- INTIA: Instituto Navarro de Tecnologías e Infraestructuras Agroalimentarias
- AEMET: Agencia Estatal de Meteorología
- UPNA: Universidad Pública de Navarra

Todas las estaciones manuales pertenecen al Gobierno de Navarra y AEMET. Cabe destacar que todas las estaciones, tanto automáticas como manuales, tienen guardados valores desde antes del año 2013 que es el año corte del estudio.

Se empezó a extraer datos de las estaciones automáticas ya que son más completas en cuanto a sus datos y además ofrecen una mayor seguridad en cuanto a los valores que recogen. De las 58 estaciones automáticas que ofrecen datos se descartaron en un principio 2: Iñarbegui (debido a que solo medía datos del año 2013) y Sartaguda GN (debido a que los datos estaban copiados de Sartaguda INTIA). Se pudo observar que había bastantes datos que faltaban debido a que la estación dejaba de medir y hasta que no se solucionaba el problema, no se registraban datos.

Por tanto se realizó un análisis de los datos de cada estación y de su localización. Con ello se pretendía observar si se disponía de datos suficientes en el periodo de tiempo del estudio y si la estación en cuestión era representativa de la población cercana. Tras el proceso se llegó a la representación observada en la *Figura 3.1.*c

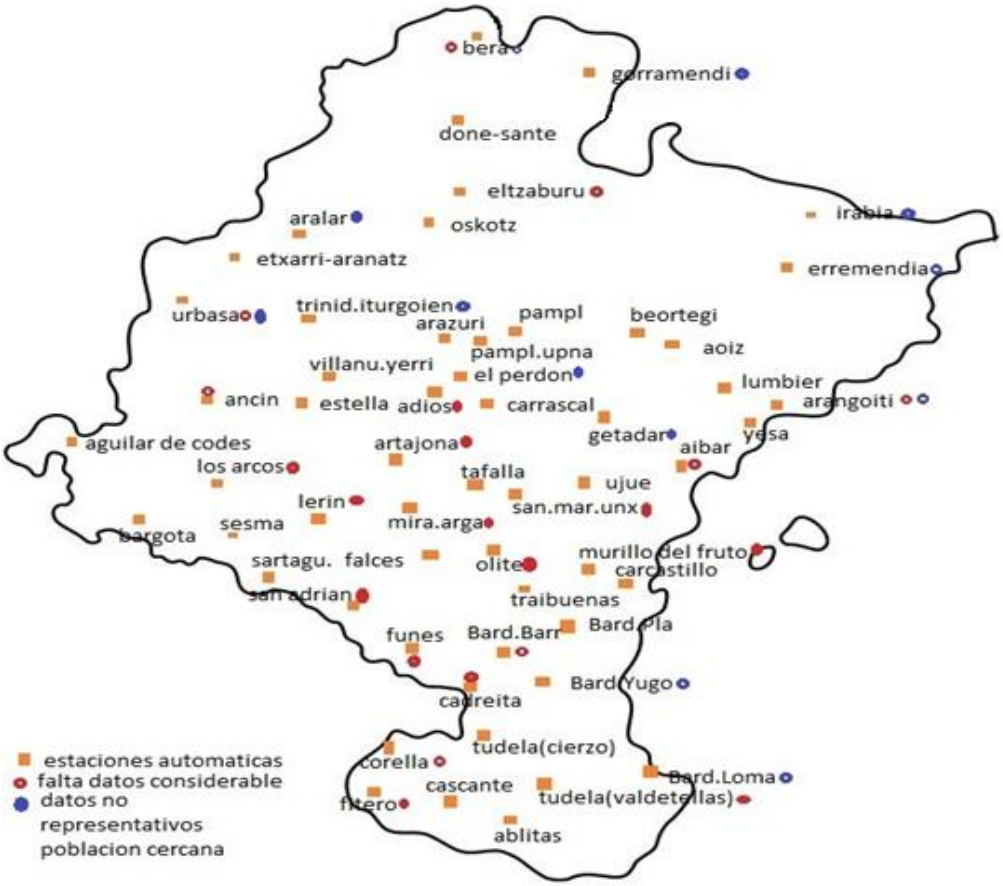


Figura 3.1.c: estaciones automáticas con sus inconvenientes detectados

Ciertas estaciones fueron descartadas por uno de los siguientes motivos:

1) Falta de datos considerable: La gran mayoría de días no había registro de la temperatura o había gran cantidad de días consecutivos sin registro.
2) Datos no representativos: Cuando una estación se encontraba a una altura que no correspondía con la población más cercana, no se tenía en cuenta debido a que no es una medición real de la temperatura en la población. Para ello, se comparaba la altura de la estación y la altura de la población más cercana y se decidía si era representativa o no.

Tras esta limpieza y extracción de datos el mapa de estaciones automáticas quedó como se observa en la *Figura 3.1.d*

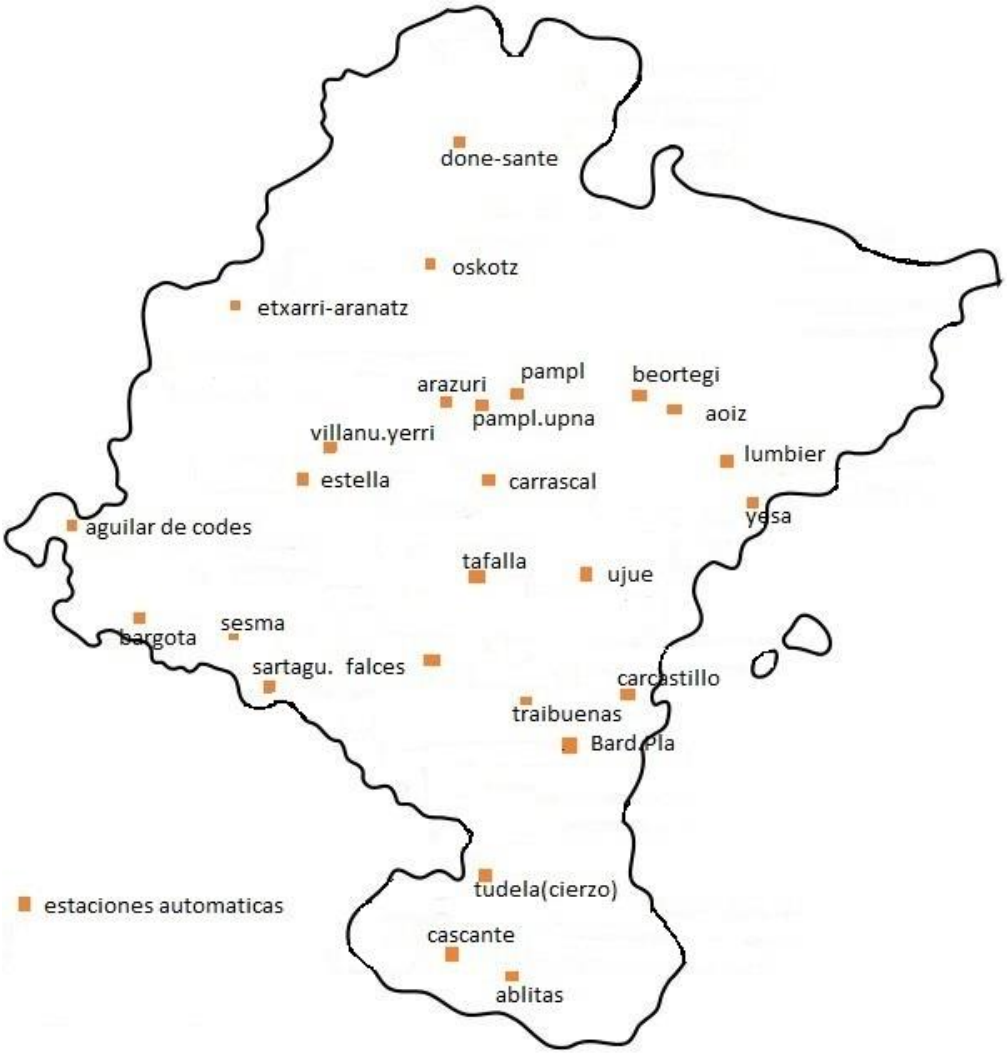


Figura 3.1.d: mapa con las estaciones automáticas seleccionadas

En este punto se observó que con los datos obtenidos había muchas poblaciones que no eran tratadas debido a que no había una estación automática cercana y muchas personas no tenían una temperatura real asignada. Todo ello sumado a la falta de información de algunas estaciones, hizo que fuera necesario extraer información de alguna otra forma. Por ello se pasó a realizar la extracción de datos de estaciones manuales.

Cabe recalcar que debido a estas pocas estaciones que medían la humedad se consideró guardar la temperatura máxima y la temperatura mínima de forma que se tuviera el mismo tipo de información en estaciones automáticas y en estaciones manuales.

Se empezó con la extracción de las estaciones manuales y se observó que había 91 estaciones disponibles que se quedaron en 83 debido a que se eliminaron algunas de ellas por diversos motivos:

-Alli-Larraun: No había datos almacenados

-Amaiur-Maya: No había datos almacenados

-Belate: No había datos almacenados

-Central Arrambide: Mismos valores que estación Cáseda

-Lekaroz MAN: Datos hasta el año 2006

-Lodosa: No había datos almacenados

-Noain MAN: No había datos almacenados

-Zuazu: No había datos almacenados del año 2015 ni 2016

Para este tipo de estaciones también existía el problema de falta de datos en varios días en distintas estaciones ya que el mantenimiento de estas se realiza de forma manual por lo que es todavía más complicado el paliar un fallo técnico rápidamente.

Para las estaciones restantes se aplicó el mismo análisis que a las estaciones automáticas (Falta de datos considerable y Datos no representativos) y se obtuvo el resultado de la *Figura 3.1.e*

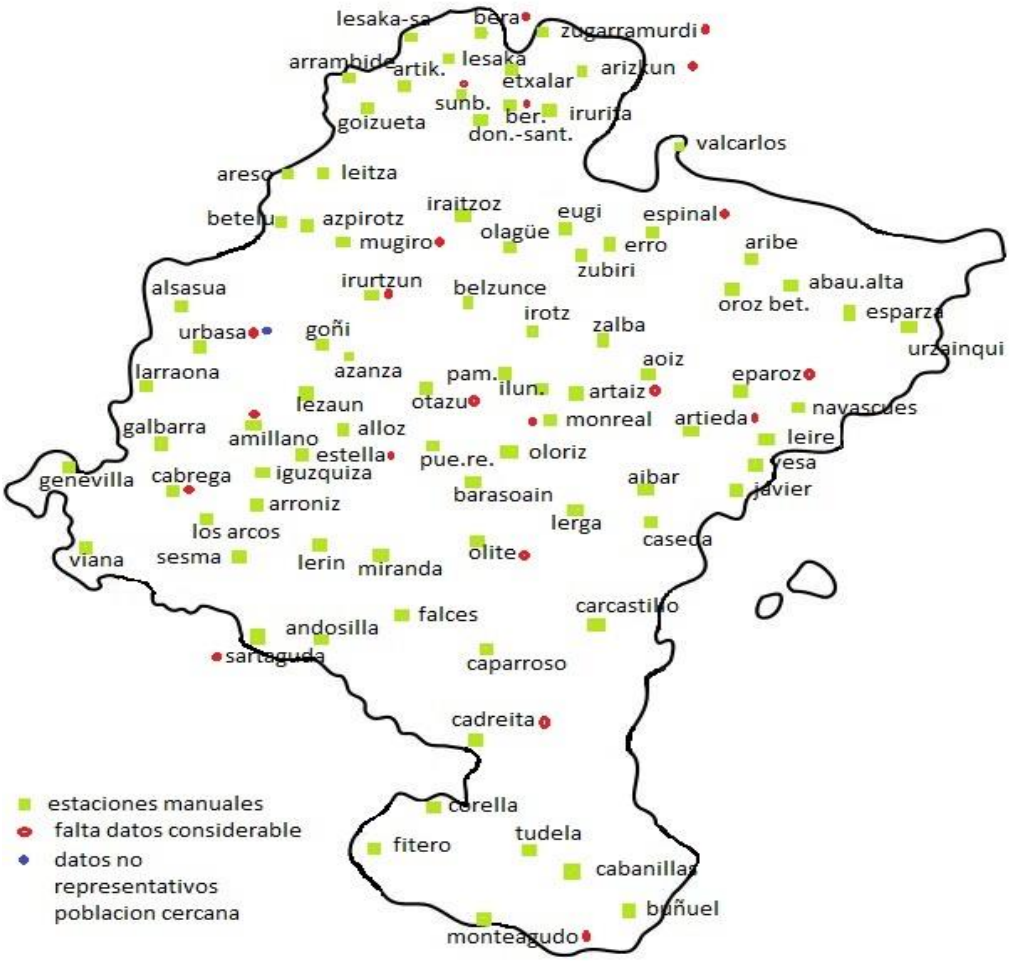


Figura 3.1.e: estaciones manuales con sus inconvenientes detectados

Por tanto se descartaron las estaciones de la misma forma anteriormente mencionada con el resultado que se puede observar en la *Figura 3.1.f*

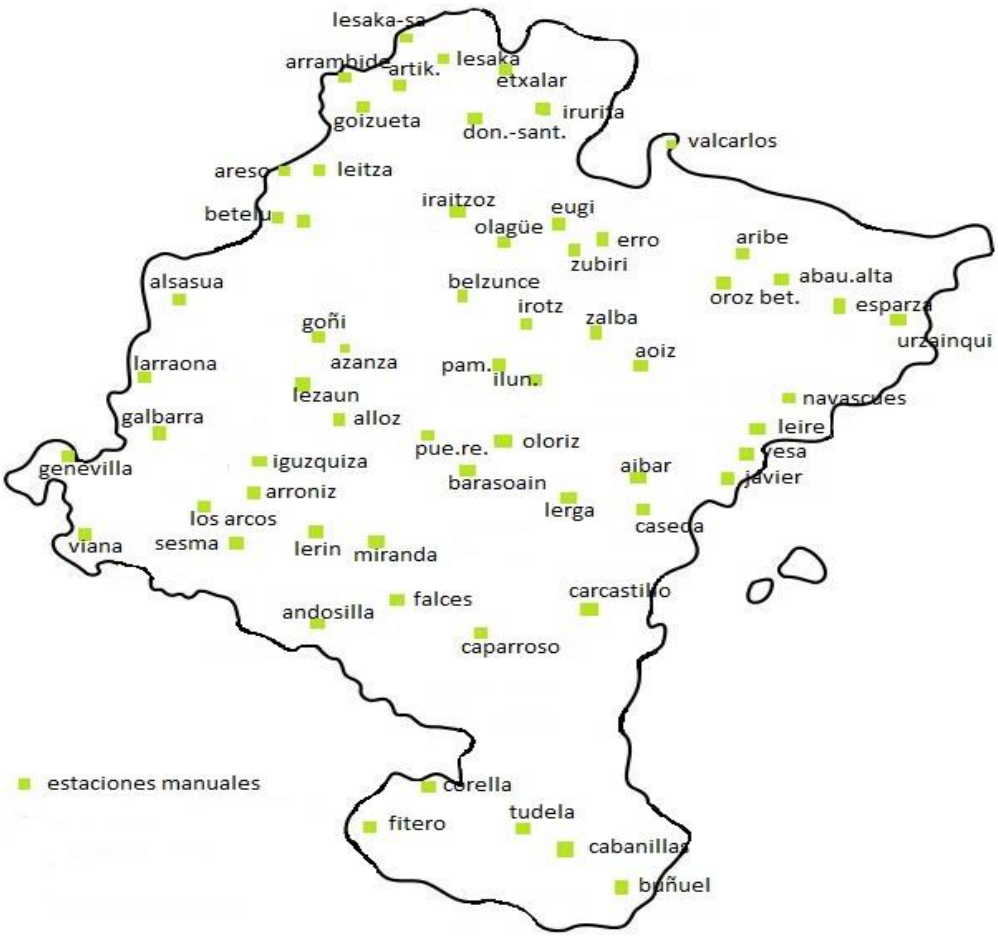


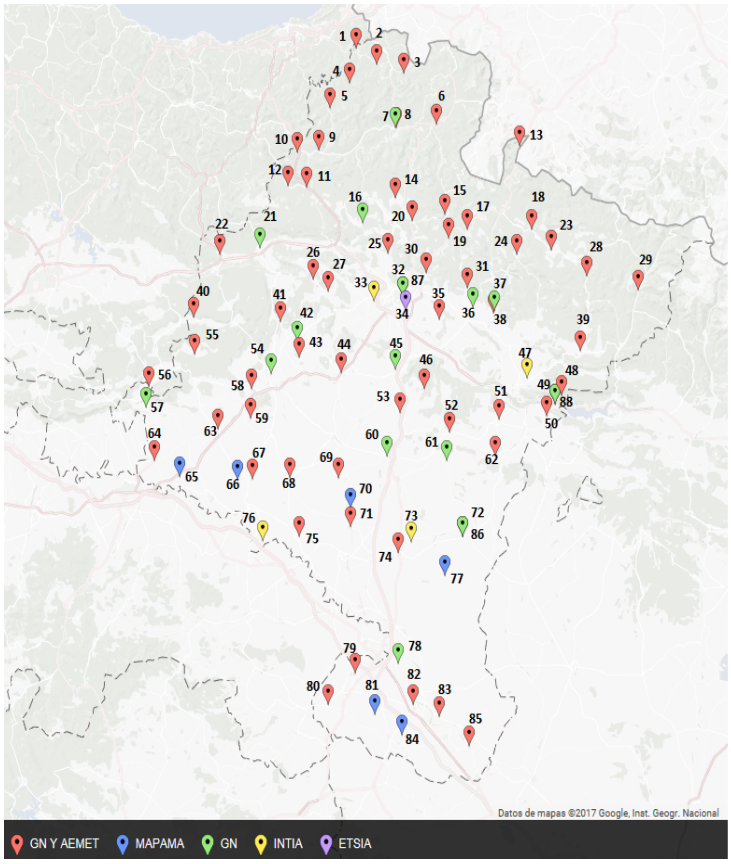
Figura 3.1.f: mapa con las estaciones manuales seleccionadas

Llegados a este punto las estaciones finalmente consideradas fueron:

-Estaciones automáticas (26): Ablitas MAPAMA, Aguilar de Codés GN, Aoiz, Arazuri INTIA, Bardenas (El Plano) MAPAMA, Bargota MAPAMA, Beortegi GN, Carcastillo (La Oliva) GN, Carrascal GN, Cascante MAPAMA, Doneztebe-Santesteban GN, Estella GN, Etxarri-Aranatz GN, Falces MAPAMA, Lumbier INTIA, Oskotz GN, Pamplona UPNA, Pamplona GN, Sartaguda INTIA, Sesma MAPAMA, Tafalla GN, Traibuenas INTIA, Tudela (Montes del Cierzo) GN, Ujué GN, Villanueva de Yerri GN y Yesa GN.

-Estaciones manuales (62): Abaurregaina-Abaurrea Alta, Aibar MAN, Alloz, Altsasu-Alsasua, Andosilla, Aoiz MAN, Areso, Aribe, Arróniz, Artikutza, Azanza, Azpirotz, Barasoain, Belzunce, Betelu, Buñuel, Cabanillas, Caparroso, Carcastillo (La Oliva) MAN, Cáseda, Corella MAN, Doneztebe-Santesteban MAN, Erro, Esparza de Salazar, Etxalar, Eugi, Falces MAN, Fitero MAN, Galbarra, Genevilla, Goizueta, Goñi, Igúzquiza, Ilundain MAN, Iraizotz, Irotz, Irurita (Baztán) MAN, Javier, Larraona, Leire, Leitza, Lerga, Lerín MAN, Lesaka, Lesaka-San Antón, Lezáun, Los Arcos, Luzaide-Valcarlos, Miranda, Navascués, Olagüe, Olóriz, Oroz Betelu, Pamplona MAN, Puente La Reina, Sesma MAN, Tudela MAN, Urzainqui, Viana, Yesa MAN, Zalba y Zubiri

Esta información se puede observar de forma conjunta en la *Figura 3.1.g*



1	Lesaka-San Antón
2	Lesaka
3	Etxalar
4	Artikutza
5	Goizueta
6	Irurita (Baztán) MAN
7	Donez.-Santest. GN
8	Donez.-Santest. MAN
9	Leitza
10	Areso
11	Azpirotz
12	Betelu
13	Luzaide-Valcarlos
14	Iraizotz
15	Eugi
16	Oskotz GN
17	Erro
18	Aribe
19	Zubiri
20	Olagüe
21	Etxarri-Aranatz GN
22	Altsasu-Alsasua
23	Abaurre.-Abaurr. Alta
24	Oroz Betelu
25	Belzunce
26	Goñi
27	Azanza
28	Esparza de Salazar
29	Urzainqui
30	Irotz
31	Zalba
32	Pamplona GN
33	Arazuri INTIA
34	Pamplona UPNA
35	Ilundain MAN
36	Beortegi GN
37	Aoiz
38	Aoiz MAN

39	Navascués
40	Larraona
41	Lezaun
42	Villanu. de Yerri
43	Alloz
44	Puente La Reina
45	Carrascal GN
46	Oloriz
47	Lumbier INTIA
48	Leire
49	Yesa GN
50	Javier
51	Aibar MAN
52	Lerga
53	Barasoain
54	Estella GN
55	Galbarra
56	Genevilla
57	Aguilar de Codes GN
58	Iguzquiza
59	Arroniz
60	Tafalla GN
61	Ujué GN
62	Caseda
63	Los Arcos
64	Viana
65	Bargota MAPAMA
66	Sesma MAPAMA
67	Sesma MAN
68	Lerin MAN
69	Miranda
70	Falces MAPAMA
71	Falces MAN
72	Carcastillo (La Oliva) GN
73	Traibuenas INTIA
74	Caparroso
75	Andosilla
76	Sartaguda INTIA

77	Bardenas (El Plano) MAPAMA
78	Tudela (Cierzo) GN
79	Corella MAN
80	Fitero MAN
81	Cascante MAPAMA
82	Tudela MAN
83	Cabanillas
84	Ablitas MAPAMA
85	Buñuel
86	Carcastillo (La Oliva) MAN
87	Pamplona MAN
88	Yesa MAN

Figura 3.1.g: mapa final con las estaciones escogidas para el proyecto

Con este resultado se conseguía una gran representación de todo el territorio con valores consistentes, bastante completos y representativos.

Llegados a este punto y a pesar de haber considerado un gran número de estaciones, todavía faltaba cierta cantidad de datos de diferentes días y diferentes estaciones. Este punto era bastante importante ya que para poder obtener resultados de calidad es imprescindible tener datos de calidad y tener varios días sin temperaturas no era lo idóneo. Por ello, se planteó la posibilidad de asignar a cada estación la estación más cercana en distancia para de esta forma poder obtener, en un día concreto, el valor que tenía esa estación cercana. Así, si en un día no teníamos una temperatura se asignaba a ese día el valor de ese día en su estación más próxima.

Se concluyó que era la solución más factible y el error no era muy grande ya que al poseer gran cantidad de estaciones los valores serían bastantes similares en estaciones cercanas. Para el cálculo de la distancia entre estaciones se utilizó el error cuadrático medio (d=$\sqrt{x^2+y^2}$), en donde "x "era la Latitud e "y" *la* Longitud de cada estación.

Tras la realización de este proceso, surgió la duda de que era posible que la estación más cercana para un día en concreto tampoco tuviera almacenado ese valor. En este caso, habría que plantearse si era una mejor opción el obtener el valor de la estación más cercana a la más cercana o si por otro lado sería más acertado obtener el valor

de la segunda estación más cercana a la estación en cuestión. Esta decisión no fue necesaria tomarla ya que tras aplicar el primer paso, no quedó ningún día en ninguna estación sin valor. Tampoco era de extrañar debido a que no había una gran cantidad de días sin valor.

Por último, se analizó la posible existencia de valores outliers que no fueran consecuentes con las temperaturas en Navarra debido a fallos en la medición, fallos en el almacenamiento, etc.

Tan solo se encontró un valor que se salía de lo normal (Temp. Max=76,3º en Cascante MAPAMA) y se le asignó el valor para ese día de su estación más cercana (Ablitas MAPAMA).

3.1.2. Gripe

Para poder obtener los datos sobre la gripe se contactó con Jesús Castilla (Instituto de Salud Pública y Laboral de Navarra) con el fin de obtener el número de casos de gripe por día en el periodo de estudio (2013-2016). Se quería almacenar dos variables para esta causa: si un día concreto estaba dentro del rango de días donde la epidemia de gripe muestra su mayor influencia y la probabilidad de que el día en concreto pertenezca a ese rango en base al número de casos. A esta información se le tuvo que añadir la información relativa a si era festivo o no debido a que está comprobado que en un día de fiesta la gente acude menos a urgencias hospitalarias y a urgencias extrahospitalarias. Para ello se realizó un gráfico (*Figura 3.1.h)* para ver esta correlación y se observó lo siguiente (año 2014):

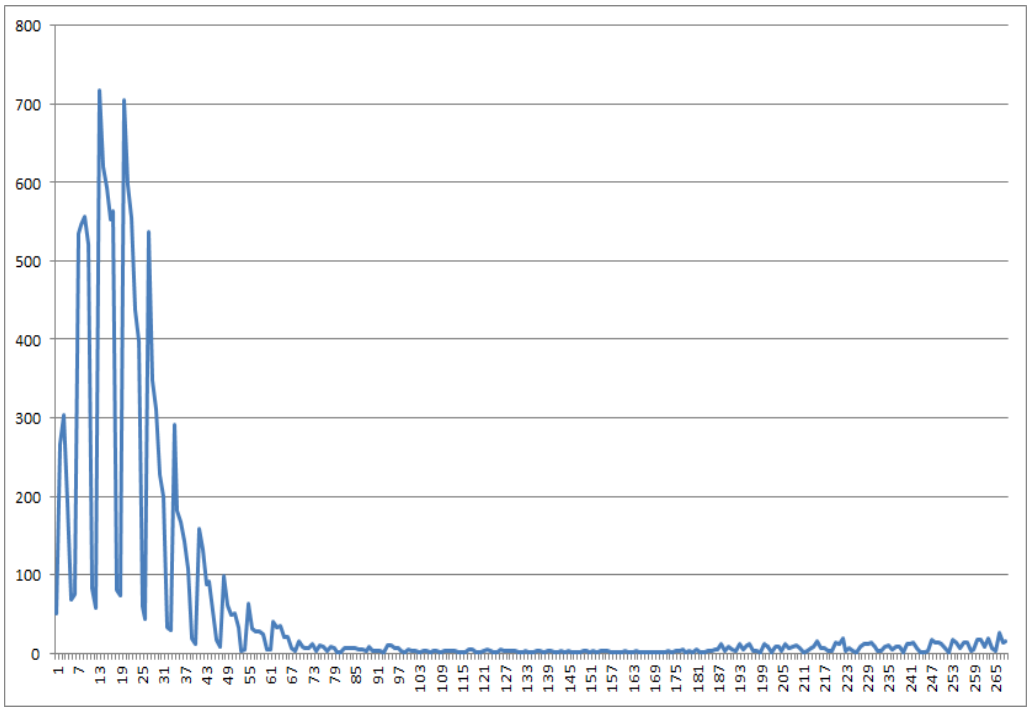


Figura 3.1.h: gráfico que muestra el número de casos de gripe recogidos a lo largo del año 2014

Se puede observar como en días consecutivos hay grandes subidas y bajadas de casos en épocas donde la epidemia está establecida. La razón es la ya mencionada bajada de casos de gripe cuando es festivo.

Por lo tanto, se llegó a la decisión de establecer el número de casos de gripe en un día festivo como el número de casos de gripe del día más cercano no festivo. Por ejemplo, para el sábado se le colocan los casos del viernes, para el domingo los del lunes, etc. De esta forma cometíamos un error en cuanto a los datos reales, pero nos servía para determinar de una forma más realista qué probabilidad había de que un día en concreto perteneciera al rango en el que la epidemia está en su parte más influyente. Hay que tener en cuenta que en días con muchos casos de gripe (enero), éstos no disminuyen de forma repentina y al día siguiente vuelven a subir como se observa en el gráfico por lo que esta solución nos permite trabajar con valores más reales. Como resultado de esta modificación se llegó a la *Figura 3.1.i*

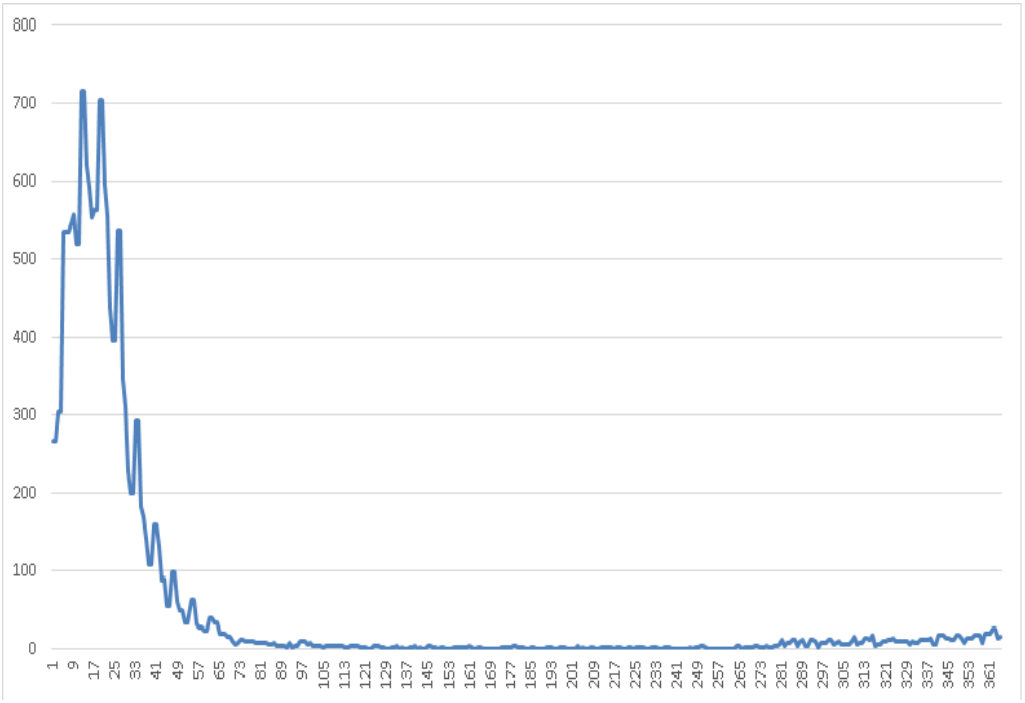


Figura 3.1.i: gráfico que recoge el número de casos de gripe en el año 2014 tras aplicarle el proceso a los días festivos

Ahora se puede observar como la caída es menor y por tanto más fiable para trabajar con estos datos. Este cambio se aplicó para los cuatro años del estudio (2013, 2014, 2015 y 2016).

Una vez llegados a este punto se decidió asignar a cada día, una probabilidad de estar en epidemia de gripe o no para cada año. Para ello, se realizó el siguiente proceso para cada año:

1. Se calculó el día medio donde más casos de gripe hay. Cabe destacar que el proceso está calculado de verano a verano y no en día naturales. Para ello se multiplica cada día por su número de casos y la suma total se divide por el número de casos totales del año.
2. Se calculó los días que se desvían de la media (desviación estándar). Para ello se resta cada día con el día medio, se eleva al cuadrado y se multiplica por el número de casos y la suma total se divide por el número de casos totales del año. Finalmente, se calcula la raíz cuadrada.
3. Se calcula la distribución gaussiana en base al día, el día medio y el valor de los días que se desvían de la media.

Con ellos obtenemos un valor para cada día del año que gráficamente deja el resultado de la *Figura 3.1.j* (año 2014) centrada en el día medio con más casos:

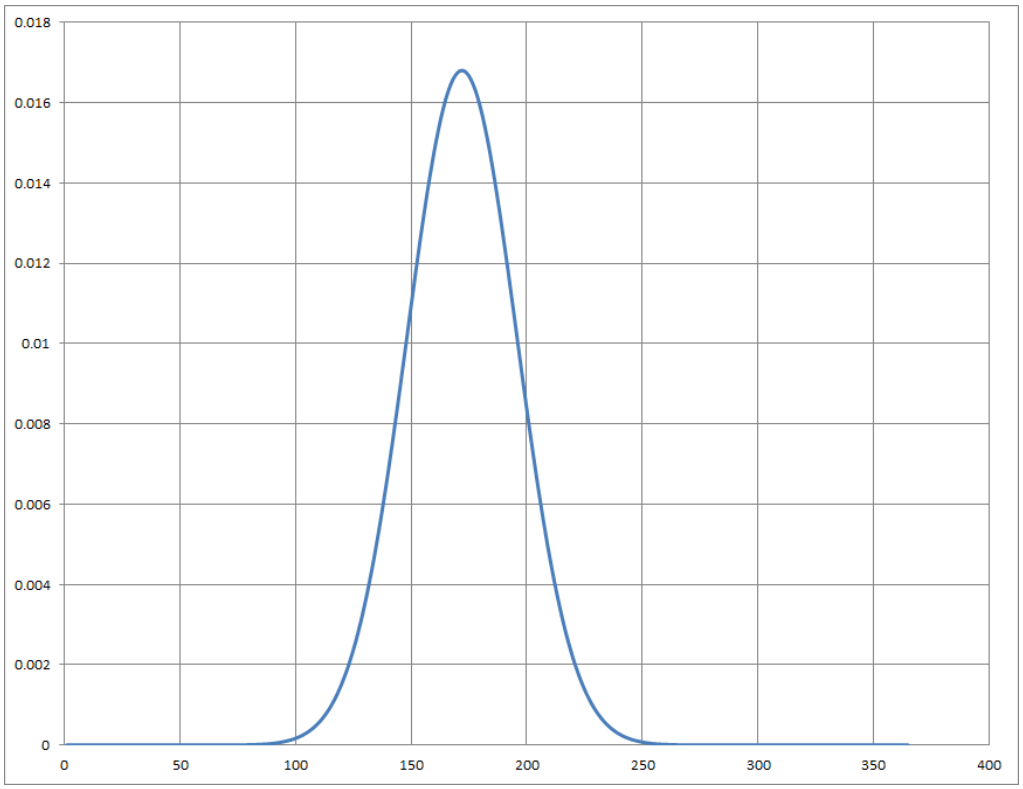


Figura 3.1.j: distribución gaussiana correspondiente a la época del año donde la gripe tiene más influencia (Año 2014). El día 0 corresponde al 1 de agosto y el 365 al 31 de julio

Llegados a este punto era necesario obtener un umbral que diferencie si un día se encuentra en epidemia de gripe o no. Para ello se observó las características de una distribución gaussiana (o normal) y se observó que cumple una muy interesante (Ver *Figura 3.1.k*):

- La proporción de mediciones situada entre la media y las desviaciones es una constante en la que:
- La media ± 1 * desviación estándar = cubre el 68,3% de los casos.
- La media ± 2 * desviación estándar = cubre el 95,5% de los casos.
- La media ± 3 * desviación estándar = cubre el 99,7% de los casos.

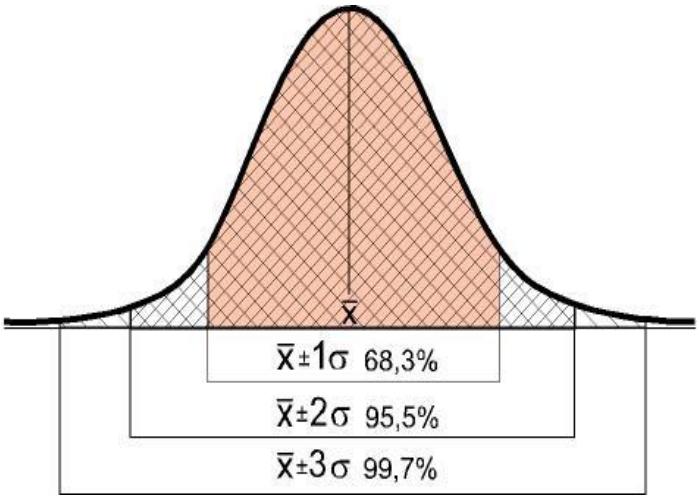


Figura 3.1.k: distribución gaussiana que muestra la cantidad de casos recogidos en base a la media y el número de desviaciones estándar. (Fuente: Curva de distribución normal. En jesusgarciaj.com [6])

Por tanto y realizando los cálculos para cada año se llegó a la conclusión de que con 2 desviaciones estándar se abarcaba gran cantidad de casos y además el umbral obtenido era bastante similar al propuesto inicialmente. Posteriormente se utilizó también el umbral con 1 desviación estándar para observar las diferencias.

Finalmente se calculó los días que quedaban por debajo y por encima del umbral calculado que separa los días de epidemia y los días de no epidemia. Siguiendo con el ejemplo del año 2014, el umbral calculado fue 0.002367796 y por tanto del día 125 del año (3 de diciembre) al 219 (7 de marzo) se considera epidemia de gripe. Como ya se ha comentado, se consideró el año de verano a verano y no como días naturales.

3.2.Datos clínicos y demográficos

3.2.1. Domicilios

Tanto para realizar el estudio de correlación como la predicción era indispensable que cada paciente del estudio tuviera asignado una estación meteorológica de donde obtener las temperaturas. Para ello se decidió obtener el domicilio de cada paciente del dataset de >=de 65 años para después asignarle la estación más cercana a ese domicilio.

En primer lugar se contactó con GDP (Gestor de Direcciones Postales) para poder obtener la latitud y longitud a partir de la dirección de los pacientes almacenados en LAKORA-TIS. De los 135953 pacientes iniciales del estudio se pudo conseguir la latitud y longitud de 110049. De hecho, son estas dos últimas variables las que nos interesan para después averiguar la estación más cercana. Por tanto, quedaban 25904 pacientes sin latitud ni longitud.

Tras esto, se llevó a cabo la construcción y ejecución de un programa en Python para obtener a través del domicilio y el código postal, la latitud y longitud. Para ello se utilizó los servicios de geolocalización de Google que arrojan un resultado bastante fiable aunque el límite de consultas por conexión cliente servidor está en 2500 consultas diarias. Esto supuso un pequeño hándicap a la hora de obtener los resultados. Además hay que sumar el problema de que algunas latitudes y longitudes devueltas por los servicios no se correspondían con direcciones correctas (fuera de Navarra, diferentes países, etc.). En este paso quedaban 7126 pacientes sin latitud ni longitud.

Por ello la única solución restante era mapear cada domicilio manualmente a través de Google Maps directamente. Fue un trabajo laborioso pero que dejó el número de pacientes sin coordenadas en 3575.

Tras esto, los pacientes restantes no poseían domicilio pero algunos mantenían el código postal. Por tanto, se les asignó un punto medio de la localidad a la que hacía referencia su código postal. A los 2946 restantes que no poseían ni domicilio ni código postal se les asignó un punto medio en base a su zona básica. Y a los 2832 pacientes que no poseían ni domicilio, ni código postal ni cupo médico se los descartó del estudio. El conjunto quedaba en 133121 pacientes.

3.2.2. Pacientes

El proyecto establecía el cupo de pacientes en enero de 2012 y que tuvieran 65 años o más el 1 de enero de 2013 con domicilio en Pamplona y comarca. También se consideró otro dataset con los pacientes de más de 16 años en enero de 2013 y de la zona de Pamplona y comarca. Sin embargo era más prioritario y más interesante, a priori, el primer dataset por lo que se trabajó más en este. Tras hacer esta criba en el primer dataset, el número de pacientes inicial para el estudio era de 135953. Tras esto era necesario que estos pacientes tuvieran domicilio en Navarra (tras aplicar el apartado anterior se quedaban en 133121) y que tuvieran centro de salud asignado (que tuvieran CIAS). De la misma forma se observó que 35843 no tenían GMA. El GMA es una variable asignada a cada paciente que se encuentra en un rango entre 0 y 5 en donde el valor más bajo significa en situación de poco riesgo y 5 de mucho riesgo. Esta variable se calcula cada 6 meses y solo para las personas que tienen CIAS. El CIAS es el cupo médico que tiene asignado cada paciente. Es decir, médico/a de cabecera y enfermero/a. Por tanto, se eliminaron aquellos pacientes que no tenían ni CIAS ni GMA ya que o no estaban en Navarra o no tenían un centro de salud asignado. Tras esta modificación el conjunto de pacientes quedó reducido a 105930.

3.2.3. Estructura de las bases de datos

La información almacenada en el Servicio Navarro de Salud sigue un esquema en estrella. Para comprender su estructura, se visualiza la *Figura 3.2.a*

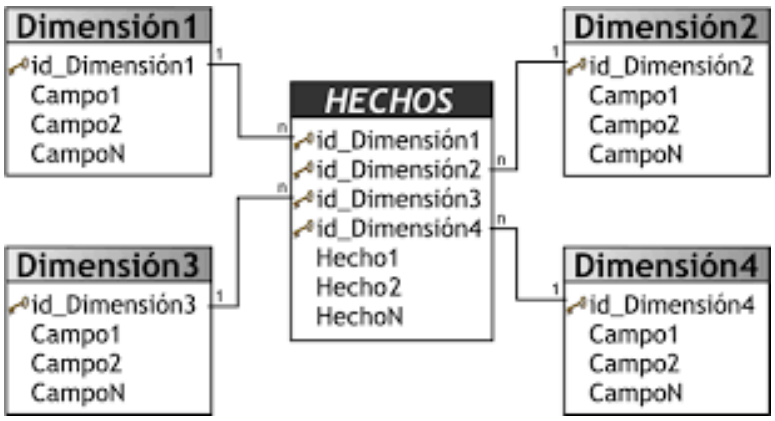


Figura 3.2.a: esquema de base de datos en estrella. (Fuente: Esquema en Estrella. En dataprix.com [7])

Se observa claramente como los datos están separados en dos secciones: Hechos y Dimensiones.

-La **tabla de hechos** contiene información de un evento específico como prescripciones de un paciente con su medicamento asignado, profesional que receta, fecha de prescripción, etc.

-La **tabla de dimensiones** contiene mucha más información (atributos) acerca del evento registrado aunque muchos menos registros. Siguiendo con el ejemplo anterior, una tabla podría ser la de medicamentos con su descripción, su grupo perteneciente, etc. o en la tabla de profesionales, sus datos personales, su centro asignado, cargo, etc.

Esta estructura tiene una serie de ventajas e inconvenientes. Por una parte, dado que la información importante para consultas se encuentra en la tabla de hechos, la velocidad para la extracción de datos es alta. Y si es necesario obtener algún campo más de las tablas de dimensiones las uniones se realizan mediante números enteros que no suponen gran carga al sistema.

Por otro lado, es posible que se guarde información redundante debido a la necesidad de almacenar varios atributos de diferentes tablas en una sola ya que es información que solo puede guardarse en otra tabla de dimensiones.

Por tanto, este tipo de esquema es útil para reestructurar los datos de una base de datos de una aplicación para explotar su información. Sin embargo, no sería muy aceptable para un almacenamiento normal de los datos debido a la perdida y redundancia de información.

Para mantener la base de poblacional actualizada se realiza una réplica semanal desde los sistemas operacionales y se aplican procesos de Extracción, Transformación y Carga (ETL). Con ello se consigue integrar y modelar la información para almacenarlo en la base de datos poblacional.

3.2.4. Fuentes de datos clínicos

Para la realización del proyecto se ha llevado a cabo una extracción de los datos considerados importantes para predecir las visitas a urgencias. Para ello, la información ha sido extraída en su mayoría de la base de datos poblacional. También se ha extraído información que no se encontraba en la poblacional para completar el dataset del estudio.

Fuentes de datos obtenidos de la base de datos poblacional:

a. dbo.DIM_PLAZAS_AP
b. farho.TH_DISPENSACIONES
c. farho.DIM_ARTICULOS
d. lamia.TH_PRESCRIPCIONES_DISPENSACIONES
e. lamia.DIM_MEDICAMENTOS
f. lamia.DIM_TIPO_PRESCRIPCION
g. leire.DIM_COMPLEJOS
h. leire.TH_URGENCIAS
i. leire.DIM_TIPO_ALTA_URGENCIAS
j. leire.DIM_TIPO_INGRESO
k. leire.DIM_TIPO_PROCEDENCIA_PACIENTE
l. an.DIM_PACIENTES

Datos obtenidos fuera de la base de datos poblacional:

a. CMBD: Posee información relativa a ingresos y diagnósticos codificados de forma manual y para su utilización se ha cogido la información de ficheros Excel.
b. Dependencia: Factor que marca si una persona es dependiente o no. Esta información fue obtenida de la base de datos de Bienestar Social.
c. Cotizacion_Farmacia: Indica el nivel de copago en farmacia en función del nivel de renta. Fue obtenido de LAKORA-TIS.
d. Fecha de nacimiento: En la base de datos poblacional anonimizada tan solo es visible la edad. Por tanto este valor fue obtenido de la versión sin anonimizar.
e. GMA: Valor que agrupa a un paciente dentro de uno de los cinco grupos de movilidad ajustados. Este valor está calculado en base a visitas a AP, probabilidad de ingreso urgente, edad, sexo, peso, enfermedades crónicas, etc. Este valor se

carga con la información de AP cada 6 meses en una tabla incremental de la base de datos poblacional llamada GMA_original.

3.2.5. Limpieza de datos

Tras la extracción de datos ha sido necesario realizar una limpieza para poder corregir e integrar de forma consistente los datos. A continuación se describen las tareas realizadas y en la documentación (ver *Anexo B*) se detalla la tareas aplicada a cada tabla.

-Fechas nulas: En multitud de ocasiones se han encontrado fechas del tipo 9999-12-31 o 1800-12-28 o 1899-12-30 que han sido reemplazados por el valor NULL.

-Cambio de tipo en fechas: Es bastante común que cada tabla de la base de datos poblacional tenga guardadas las fechas con un tipo diferente. Esto hace imposible el poder trabajar con ellas y se procedió al cambio de las mismas al tipo Date.

-Corrección de valores inconsistentes: En la tabla TH_URGENCIAS se pudo observar como en la fecha de alta había valores sin sentido como por ejemplo '0001010' debido a que es un campo rellenado a mano. Debido a que no había una gran cantidad de registros con estos valores, se consensuó establecer esa fecha de alta como el mismo día de la fecha de visita.

-Borrado de pacientes repetidos: Se pudo observar como algún paciente poseía dos (y hasta 3) claves primarias únicas. Se realizó un borrado de las más antiguas en base a su fecha de carga.

3.3.Base de datos generada

A continuación se muestra un esquema de la base de datos obtenida. La documentación extensa se encuentra en el *Anexo B*.

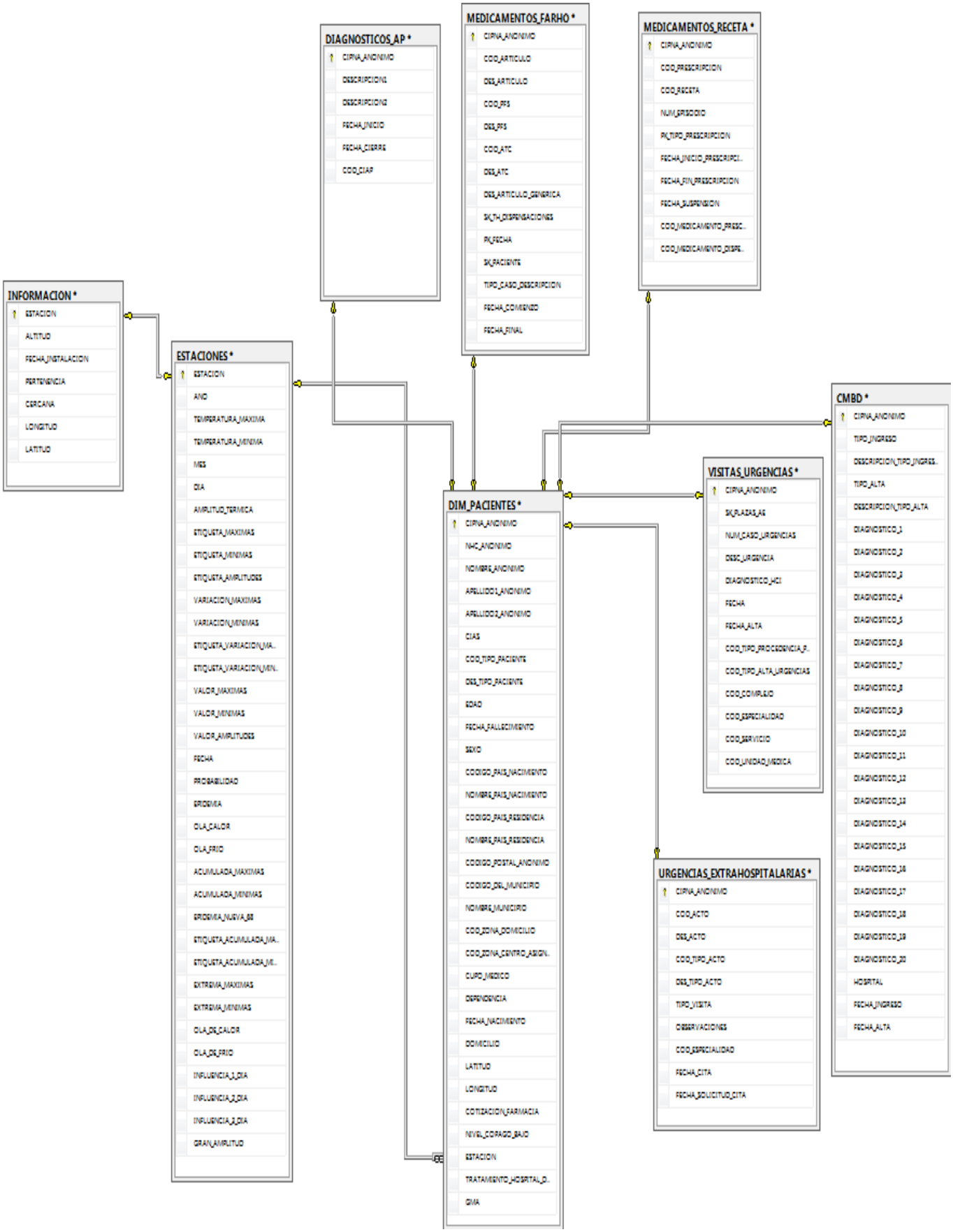


Figura 3.3.a: Esquema de la base de datos creada

4. Variables del estudio

En este apartado se explicará qué variables han sido consideradas para poder llevar a cabo el proyecto. En primer lugar, se presentarán las variables a nivel de paciente (*Sección 4.1*) y a continuación aquellas a nivel de día (*Sección 4.2*). Dentro de esta misma sección se hará mención a todas las variables creadas (*Sección 4.2.1*) y a aquellas finalmente utilizadas en el estudio y las predicciones (*Sección 4.2.2*).

Tras la extracción y limpieza de los datos se han definido las variables orientadas a una correcta predicción y aquellas que se creían de valor para realizar el estudio de correlación.

A continuación se mencionan junto con una descripción de cada una de ellas:

4.1.Variables a nivel de paciente

1) **SEXO**: Género del paciente. H:Hombre, M:Mujer
2) **EDAD**: Edad del paciente.
3) **FECHA_EVENTO**: Fecha utilizada para el cálculo del resto de variables. En el caso de CLASE positiva (Gente que ha ido alguna vez a urgencias) es el día que ha ido a urgencias. Para CLASE negativa (Gente que ha ido alguna vez a urgencias pero nunca en la fecha_evento) es cualquier día que no haya ido a urgencias.
4) **DEPENDENCIA**: Indica si el paciente es dependiente o no. 1: Es dependiente. 0: No es dependiente.
5) **NIVEL_COPAGO_BAJO**: Indica si el paciente pertenece a un nivel de copago bajo. Es decir si pertenece a los grupos TSI 001 (), TSI 002 (01) o TSI 003 (). 1: Pertenece a nivel de copago bajo. 0: No pertenece a nivel de copago bajo.
6) **ESTACION**: Es la estación más cercana asignada al paciente. Es utilizada para asignar cualquier valor atmosférico al paciente.
7) **TRATAMIENTO_HOSPITAL_DIA**: Indica si el paciente recibe tratamiento de hospital de día. 1: Recibe tratamiento en hospital de día. 0: No recibe tratamiento en hospital de día.
8) **NUMERO_VISITAS_URGENCIAS**: Indica el número de veces que el paciente ha ido a urgencias en el último año desde la FECHA_EVENTO
9) **NUMERO_VISITAS_URGENCIAS_EXTRA**: Indica el número de veces que el paciente ha ido a urgencias extrahospitalarias en el último año desde la FECHA_EVENTO.
10) **NUMERO_INGRESOS_URGENTES**: Indica el número de veces que el paciente ha sido ingresado de forma urgente en el último año desde la FECHA_EVENTO.
11) **NUMERO_MEDICAMENTOS_RECETA**: Indica el número de medicamentos prescritos en receta al paciente en el último año desde la FECHA_EVENTO.
12) **NUMERO_MEDICAMENTOS_FARHO**: Indica el número de medicamentos prescritos en farmacia hospitalaria al paciente en el último año desde la FECHA_EVENTO.

13) **GMA**: Indicador que clasifica al paciente dentro de uno de los cinco grupos creados en base a ciertos valores de diferentes variables. Estas variables son visitas a AP, gasto en farmacia según prescripción, probabilidad de ingreso urgente, edad, sexo, número de enfermedades crónicas, etc. 1: Grupo morbilidad muy baja. 2: Grupo morbilidad baja. 3: Grupo morbilidad media. 4: Grupo morbilidad alta. 5: Grupo morbilidad muy alta.
14) **CLASE**: Indica si la persona ha ido o no a urgencias. 1: Clase positiva (ha ido a urgencias). 0: Clase negativa (no ha ido a urgencias).

4.2.Variables a nivel de día

4.2.1. Variables totales

-**AMPLITUD_TERMICA**: Valor diario obtenido de la operación→TEMPERATURA_MAXIMA-TEMPERATURA_MINIMA.

-**ETIQUETA_MAXIMAS**: Variable cualitativa que tiene cinco posibles valores: Muy Baja, Baja, Media, Alta y Muy Alta. Hace referencia a cómo es la TEMPERATURA_MAXIMA ese día. Para ello se realizó un histograma con el número de días con cada TEMPERATURA_MAXIMA entre todas las estaciones del estudio. De esta forma podríamos dividir todo el rango de TEMPERATURA_MAXIMA en cinco partes diferenciadas. El histograma mencionado se visualiza en la *Figura 4.2.a*

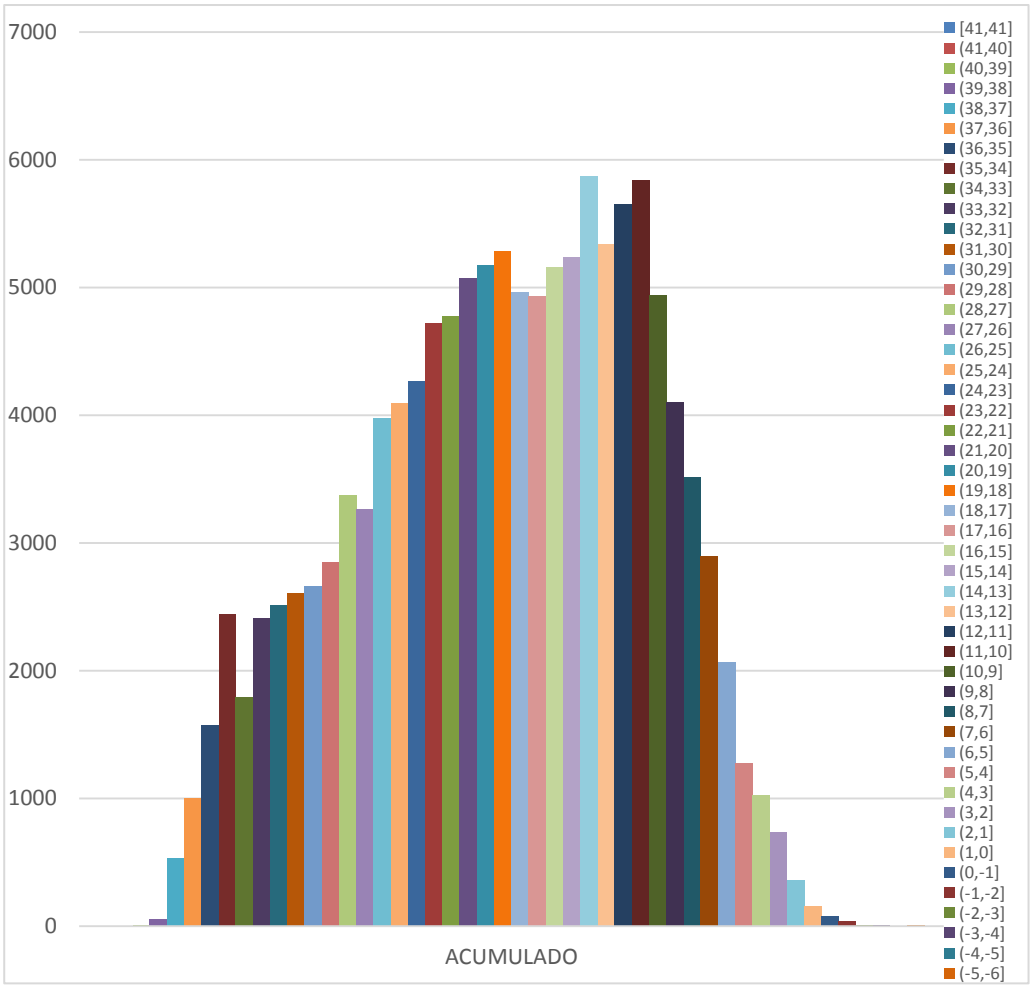


Figura 4.2.a: histograma con el número de días para cada temperatura máxima registrada en todas las estaciones en los cuatro años de estudio

Se puede observar como el rango de TEMPERATURA_MAXIMA va entre -6 y 41 grados centígrados. Vista esta distribución se llegó a la conclusión de etiquetar las temperaturas máximas de la siguiente forma:

-Muy Baja: [-6,6)

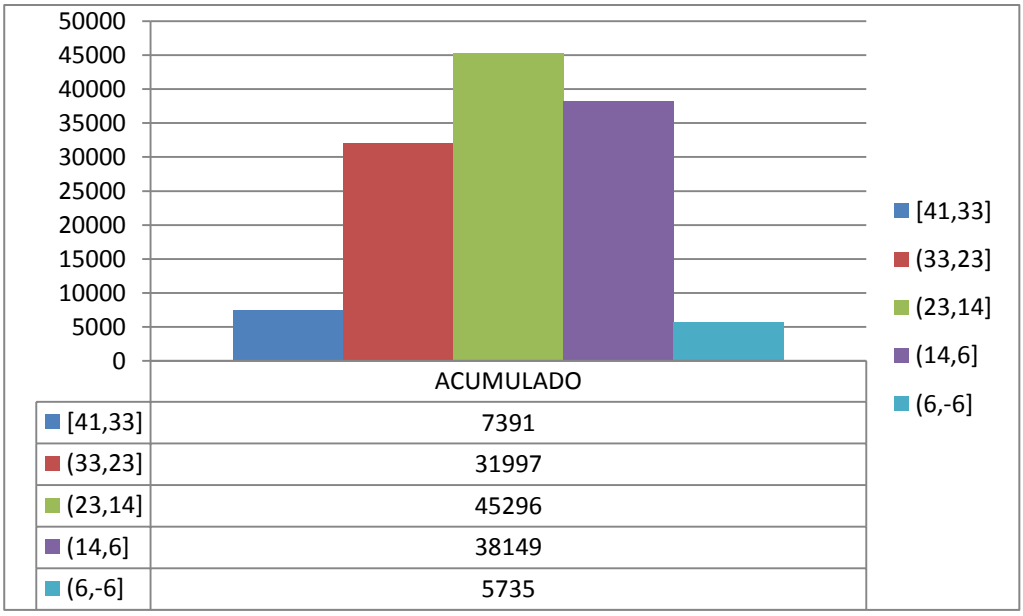
-Baja: [6,14)

-Media: [14,23)

-Alta: [23,33)

-Muy Alta: [33,41]

Para comprobar si esta distribución seguía una distribución normal se realizó una agrupación de los valores en los rangos descritos y el resultado se observa en la *Figura 4.2.b*



	ACUMULADO
[41,33]	7391
(33,23]	31997
(23,14]	45296
(14,6]	38149
(6,-6]	5735

Figura 4.2.b: histograma mediante acumulación de días en base a las cinco categorías mencionadas para las temperaturas máximas

-**ETIQUETA_MINIMAS**: Variable cualitativa que tiene cinco posibles valores: Muy Baja, Baja, Media, Alta y Muy Alta. Hace referencia a cómo es la TEMPERATURA_MINIMA ese día. Para ello se realizó un histograma con el número de días con cada TEMPERATURA_MINIMA entre todas las estaciones del estudio. De esta forma podríamos dividir todo el rango de TEMPERATURA_MINIMA en cinco partes diferenciadas. El histograma mencionado se visualiza en la *Figura 4.2.c*

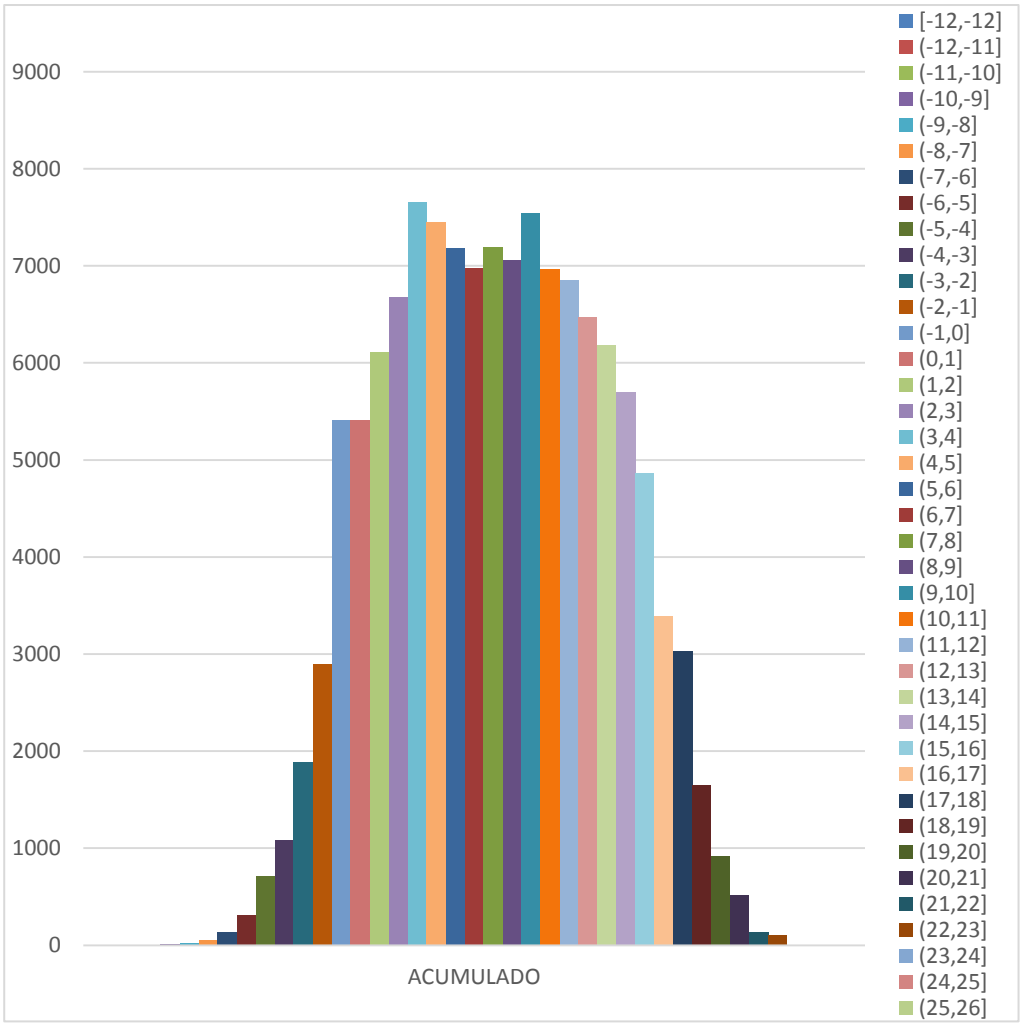


Figura 4.2.c: histograma con el número de días para cada temperatura mínima registrada en todas las estaciones en los cuatro años de estudio

Se puede observar como el rango de TEMPERATURA_MINIMA va entre -12 y 25.5 grados centígrados. Vista esta distribución se llegó a la conclusión de etiquetar las temperaturas mínimas de la siguiente forma:

-Muy Baja: [-12,-1]

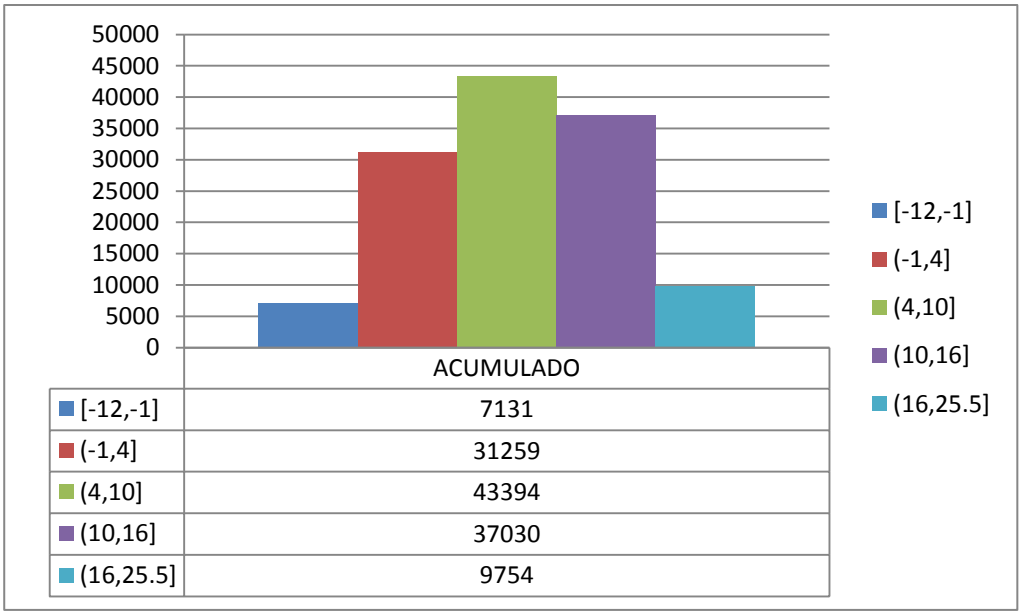
-Baja: (-1,4]

-Media: (4,10]

-Alta: (10,16]

-Muy Alta: (16,26]

Para comprobar si esta distribución seguía una distribución normal se realizó una agrupación de los valores en los rangos descritos y el resultado se observa en la *Figura 4.2.d*



	ACUMULADO
[-12,-1]	7131
(-1,4]	31259
(4,10]	43394
(10,16]	37030
(16,25.5]	9754

Figura 4.2.d: histograma mediante acumulación de días en base a las cinco categorías mencionadas para las temperaturas mínimas

-**ETIQUETA_AMPLITUDES**: Variable cualitativa que tiene cinco posibles valores: Muy Baja, Baja, Media, Alta y Muy Alta. Hace referencia a cómo es la AMPLITUD_TERMICA ese día. Para ello se realizó un histograma con el número de días con cada AMPLITUD_TERMICA entre todas las estaciones del estudio. De esta forma podríamos dividir todo el rango de AMPLITUD_TERMICA en cinco partes diferenciadas. El histograma mencionado se visualiza en la *Figura 4.2.e*

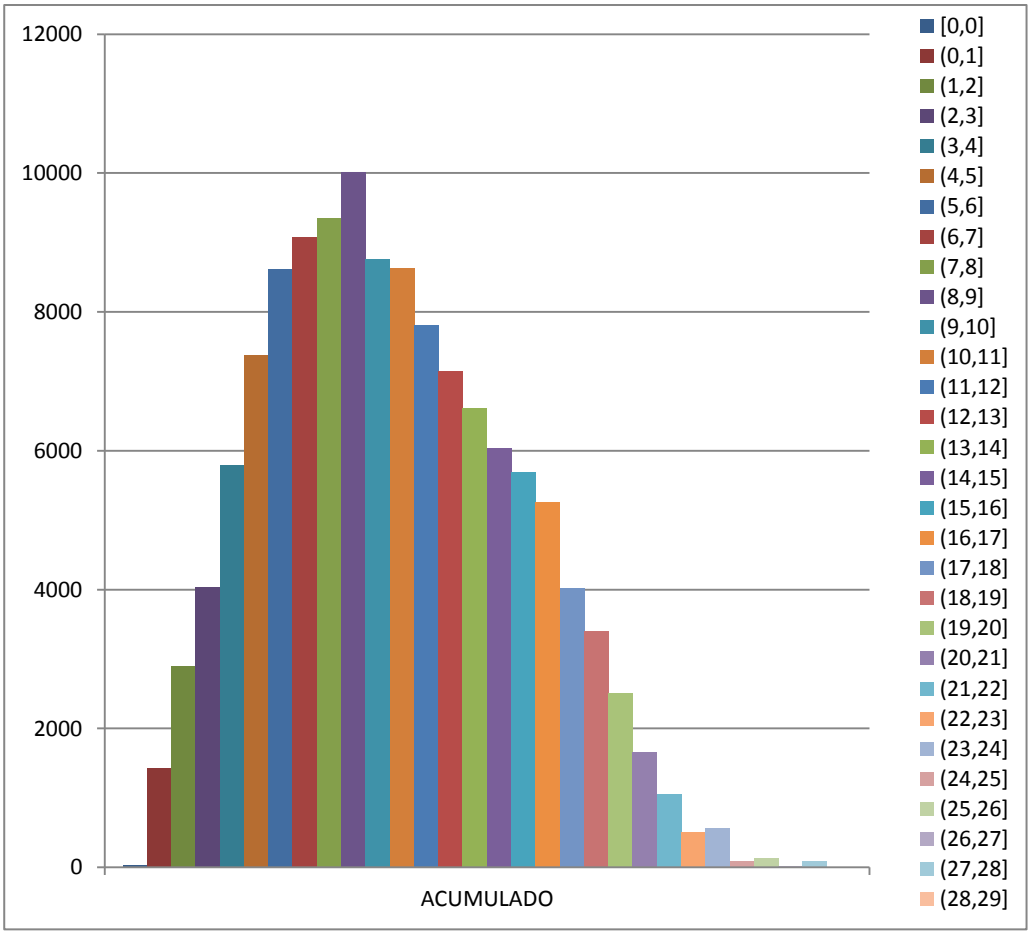


Figura 4.2.e: histograma con el número de días para cada amplitud térmica registrada en todas las estaciones en los cuatro años de estudio

Se puede observar como el rango de AMPLITUD_TERMICA va entre 0 y 29 grados centígrados. Vista esta distribución se llegó a la conclusión de etiquetar las amplitudes térmicas de la siguiente forma:

-Muy Baja: [0,3]

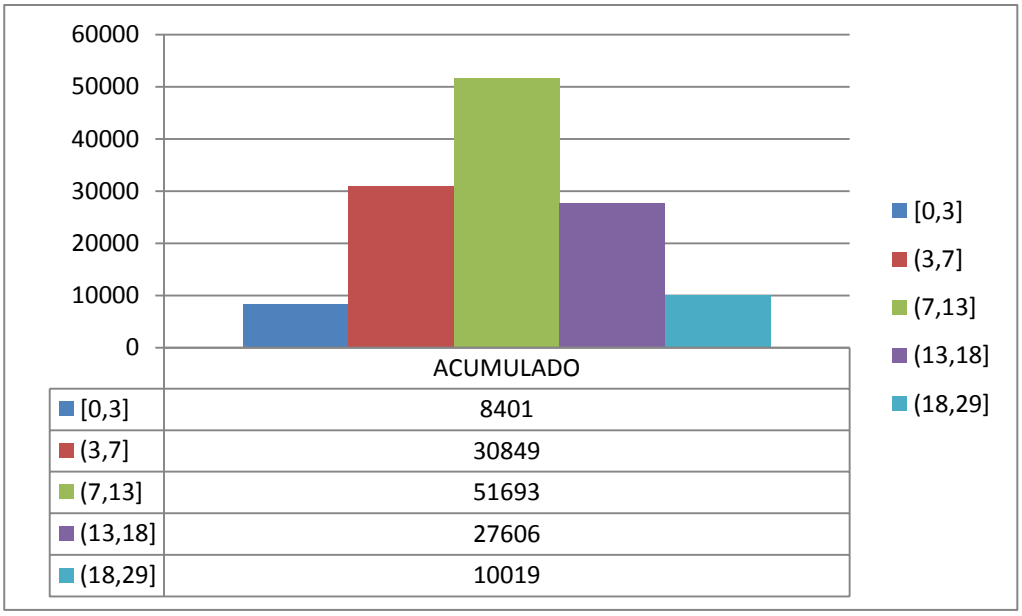
-Baja: (3,7]

-Media: (7,13]

-Alta: (13,18]

-Muy Alta: (18,29]

Para comprobar si esta distribución seguía una distribución normal se realizó una agrupación de los valores en los rangos descritos y el resultado se observa en la *Figura 4.2.f*



	ACUMULADO
[0,3]	8401
(3,7]	30849
(7,13]	51693
(13,18]	27606
(18,29]	10019

Figura 4.2.f: histograma mediante acumulación de días en base a las cinco categorías mencionadas para las amplitudes térmicas

-**VARIACION_MAXIMAS**: Valor numérico como resultado de la diferencia de TEMPERATURA_MAXIMA entre el día en cuestión y el día siguiente. Debido a ello el valor correspondiente al último día del estudio (31/12/2016) está sin valor.

-**VARIACION _MINIMAS**: Valor numérico como resultado de la diferencia de TEMPERATURA_MINIMA entre el día en cuestión y el día siguiente. Debido a ello el valor correspondiente al último día del estudio (31/12/2016) está sin valor.

-**ETIQUETA_VARIACION_MAXIMAS**: Variable cualitativa que tiene tres posibles valores: Baja, Media y Alta. Hace referencia a cómo es la VARIACION_MAXIMA ese día. Para ello se realizó un histograma con el número de días con cada VARIACION_MAXIMA entre todas las estaciones del estudio. De esta forma podríamos dividir todo el rango de VARIACION_MAXIMA en tres partes diferenciadas. El histograma mencionado se visualiza en la *4.2.g*

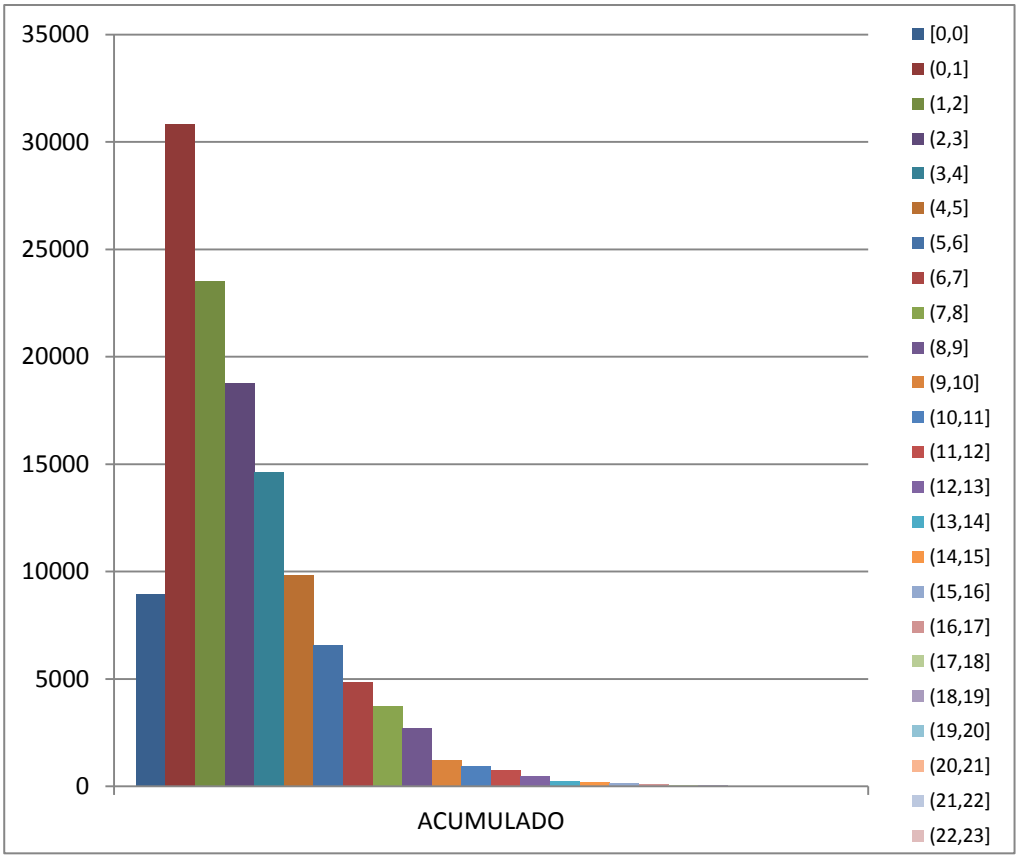


Figura 4.2.g: histograma con el número de días para cada variación de las temperaturas máximas registrada en todas las estaciones en los cuatro años de estudio

Se puede observar como el rango de VARIACION_MAXIMA va entre 0 y 23 grados centígrados. Vista esta distribución se llegó a la conclusión de etiquetar las variaciones de temperaturas máximas de la siguiente forma:

-Baja: [0,1]

-Media: (1,4]

-Alta: (4,23]

Para comprobar si esta distribución seguía una distribución normal se realizó una agrupación de los valores en los rangos descritos y el resultado se observa en la *Figura 4.2.h*

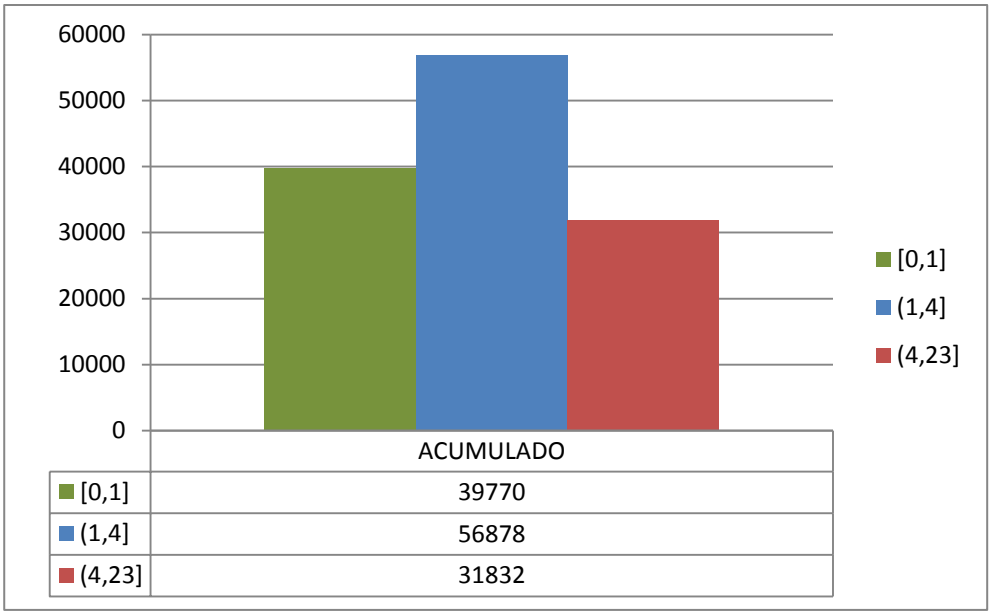


Figura 4.2.h: histograma mediante acumulación de días en base a las cinco categorías mencionadas para las variaciones de temperaturas máximas

-**ETIQUETA_VARIACION_MINIMAS**: Variable cualitativa que tiene tres posibles valores: Baja, Media y Alta. Hace referencia a cómo es la VARIACION_MINIMA ese día. Para ello se realizó un histograma con el número de días con cada VARIACION_MINIMA entre todas las estaciones del estudio. De esta forma podríamos dividir todo el rango de VARIACION_MINIMA en tres partes diferenciadas. El histograma mencionado se visualiza en la *Figura 4.2.i*

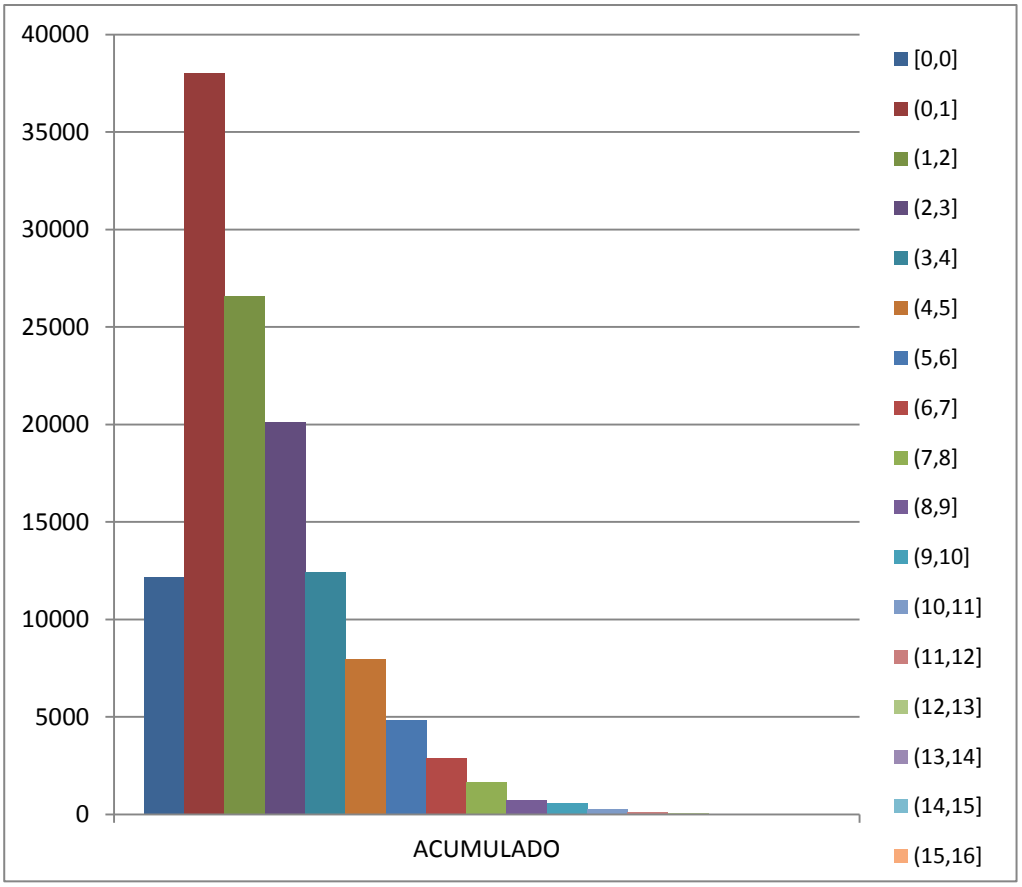


Figura 4.2.i: histograma con el número de días para cada variación de las temperaturas mínimas registrada en todas las estaciones en los cuatro años de estudio

Se puede observar como el rango de VARIACION_MINIMA va entre 0 y 16 grados centígrados. Vista esta distribución se llegó a la conclusión de etiquetar las variaciones de temperatura mínimas de la siguiente forma:

-Baja: [0,1]

-Media: (1,4]

-Alta: (4,16]

Para comprobar si esta distribución seguía una distribución normal se realizó una agrupación de los valores en los rangos descritos y el resultado se observa en la *Figura 4.2.j*

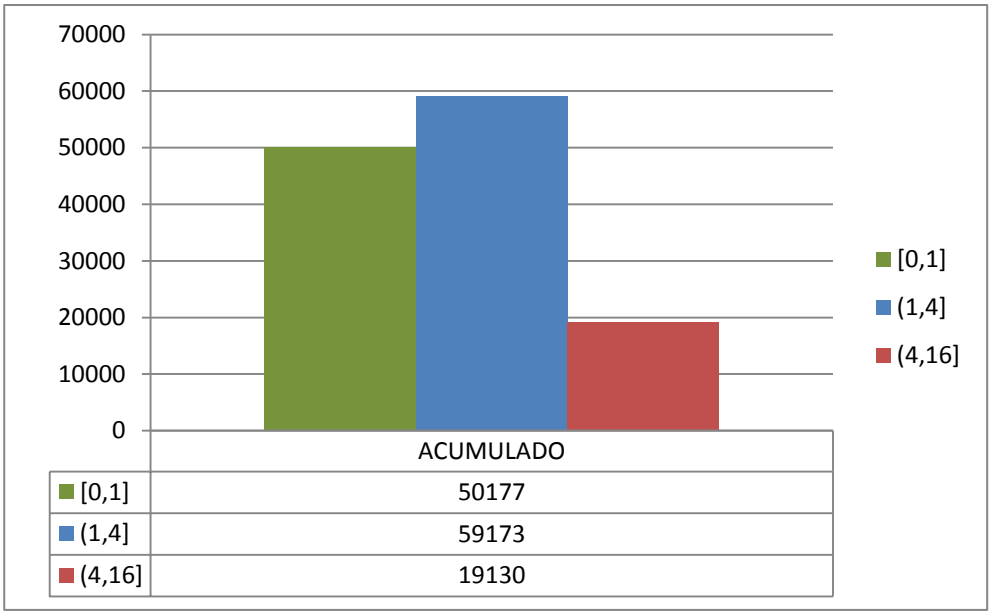


Figura 4.2.j: histograma mediante acumulación de días en base a las cinco categorías mencionadas para las variaciones de temperaturas mínimas

-**VALOR_MAXIMAS**: Valor que simboliza numéricamente los días con temperaturas altas que ha habido. Está calculada de la siguiente forma:

-Muy Alta:+1.5

-Alta:-0.5

-Media:-1

-Baja: 0

-Muy Baja: 0

Si el día en cuestión tiene uno de los siguientes valores en su ETIQUETA_MAXIMAS se le suma al valor de VALOR_MAXIMAS del día anterior sus valores respectivos.

-**VALOR_MINIMAS**: Valor que simboliza numéricamente los días con temperaturas bajas que ha habido. Está calculada de la siguiente forma:

-Muy Baja:+1.5

-Baja:-0.5

-Media:-1

-Alta: 0

-Muy Alta: 0

Si el día en cuestión tiene uno de los siguientes valores en su ETIQUETA_MINIMAS se le suma al valor de VALOR_MINIMAS del día anterior sus valores respectivos.

-**VALOR_AMPLITUDES**: Valor que simboliza numéricamente los días con grandes amplitudes que ha habido. Está calculada de la siguiente forma:

-Muy Alta:+1.5

-Alta:-0.5

-Media:-1

-Baja: 0

-Muy Baja: 0

Si el día en cuestión tiene uno de los siguientes valores en su ETIQUETA_AMPLITUDES se le suma al valor de VALOR_AMPLITUDES del día anterior sus valores respectivos.

-**FECHA**: El día, mes y año del registro en concreto.

-NUMERO_VISITAS: Visitas a urgencias en el día indicado en las zonas a estudiar.

-NUMERO_FALLECIMIENTOS: Fallecimientos en el día indicado en las zonas a estudiar.

-OLA_CALOR (S o N): Día de temperatura alta extrema (Si VALOR_MAXIMAS>=3).

-OLA_FRIO (S o N): Día de temperatura baja extrema (Si VALOR_MINIMAS>=3).

-GRAN_AMPLITUD (S o N): Si la amplitud térmica ha sido grande (Si VALOR_AMPLITUDES>=3).

-EPIDEMIA (S o N): Día de epidemia o no en base a una distribución gaussiana que recoge el 95,5% de los casos de gripe.

-EPIDEMIA_68 (S o N): Día de epidemia o no en base a una distribución gaussiana que recoge el 68.3% de los casos de gripe.

-DIA_SEMANA (L, M, X, J, V, S, D): El día de la semana.

-FESTIVO (S o N): Si el día en cuestión es festivo o no. El sábado y domingo siempre se cuentan como festivos.

-ACUMULADA_MAXIMAS (S o N): Está a “S” si para ese día la suma de la TEMPERATURA_MAXIMA de los 2 días anteriores + TEMPERATURA_MAXIMA del día es >=70.

-ACUMULADA_MINIMAS (S o N): Está a “S” si para ese día la suma de la TEMPERATURA_MINIMA de 2 días anteriores + TEMPERATURA_MINIMA del día es <=18.

-EXTREMA_MAXIMAS (S o N): Está a “S” si la TEMPERATURA_MAXIMA de ese día está por encima del 95 percentil, es decir, >= 33 grados.

-EXTREMA_MINIMAS (S o N): Está a “S” si la TEMPERATURA_MINIMA de ese día está por debajo del 5 percentil, es decir, <= -1 grado.

-OLA_DE_CALOR (S o N): Está a “S” si los dos días anteriores + día en cuestión han sido EXTREMA_MAXIMAS. Lo que realmente sería una ola de calor.

-OLA_DE_FRIO (S o N): Está a “S” si los dos días anteriores + día en cuestión han sido EXTREMA_MINIMAS. Lo que realmente sería una ola de frio.

-INFLUENCIA_1_DIA (S o N): Está a “S” si es el día siguiente a EXTREMA_MINIMAS.

-INFLUENCIA_2_DIA (S o N): Está a “S” si está dentro de los dos días siguientes a EXTREMA_MINIMAS.

-INFLUENCIA_3_DIA (S o N): Está a “S” si está dentro de los tres días siguientes a EXTREMA_MINIMAS.

-INFLUENCIA_SOLO_2_DIA (S o N): Está a “S” si es el segundo día tras EXTREMA_MINIMAS.

-INFLUENCIA_SOLO_3_DIA (S o N): Está a “S” si es el tercer día tras EXTREMA_MINIMAS.

-INFLUENCIA_1_DIA_MAXIMAS (S o N): Está a “S” si es el día siguiente a EXTREMA_MAXIMAS.

-INFLUENCIA_2_DIA_MAXIMAS (S o N): Está a “S” si está dentro de los dos días siguientes a EXTREMA_MAXIMAS.

-INFLUENCIA_3_DIA_MAXIMAS (S o N): Está a “S” si está dentro de los tres días siguientes a EXTREMA_MAXIMAS.

-INFLUENCIA_SOLO_2_DIA_MAXIMAS (S o N): Está a “S” si es el segundo día tras EXTREMA_MAXIMAS.

-INFLUENCIA_SOLO_3_DIA_MAXIMAS (S o N): Está a “S” si es el tercer día tras EXTREMA_MAXIMAS.

-GRIPE: Número de casos de gripe detectados en el día indicado.

-**PROBABILIDAD**: Es la probabilidad de un día en concreto de estar en epidemia de gripe o no.

-**TEMPERATURA_MAXIMA**: Temperatura máxima alcanzada en un día en concreto en una estación determinada.

-TEMPERATURA_MINIMA: Temperatura máxima alcanzada en un día en concreto en una estación determinada.

-TEMPERATURA_MEDIA: Temperatura media alcanzada en un día en concreto en una estación determinada.

4.2.2. Variables finalmente utilizadas

1) **NUMERO_VISITAS**: Visitas a urgencias en el día indicado en las zonas a estudiar.
2) **FECHA**: El día, mes y año del registro en concreto.
3) **OLA_CALOR** (S o N): Día de temperatura alta extrema (Si VALOR_MAXIMAS>=3).
4) **OLA_FRIO** (S o N): Día de temperatura baja extrema (Si VALOR_MINIMAS>=3).
5) **GRAN_AMPLITUD** (S o N): Si la amplitud térmica ha sido grande (Si VALOR_AMPLITUDES>=3).
6) **EPIDEMIA** (S o N): Día de epidemia o no en base a una distribución gaussiana que recoge el 95,5% de los casos de gripe.
7) **EPIDEMIA_68** (S o N): Día de epidemia o no en base a una distribución gaussiana que recoge el 68.3% de los casos de gripe.
8) **DIA_SEMANA** (L, M, X, J, V, S, D): El día de la semana.
9) **FESTIVO** (S o N): Si el día en cuestión es festivo o no. El sábado y domingo siempre se cuentan como festivos.
10) **ACUMULADA_MAXIMAS** (S o N): Está a "S" si para ese día la suma de la TEMPERATURA_MAXIMA de los 2 días anteriores + TEMPERATURA_MAXIMA del día es >=70.
11) **ACUMULADA_MINIMAS** (S o N): Está a "S" si para ese día la suma de la TEMPERATURA_MINIMA de 2 días anteriores + TEMPERATURA_MINIMA del día es <=18.
12) **EXTREMA_MAXIMAS** (S o N): Está a "S" si la TEMPERATURA_MAXIMA de ese día está por encima del 95 percentil, es decir, >= 33 grados.
13) **EXTREMA_MINIMAS** (S o N): Está a "S" si la TEMPERATURA_MINIMA de ese día está por debajo del 5 percentil, es decir, <= -1 grado.
14) **OLA_DE_CALOR** (S o N): Está a "S" si los dos días anteriores + día en cuestión han sido EXTREMA_MAXIMAS. Lo que realmente sería una ola de calor.
15) **OLA_DE_FRIO** (S o N): Está a "S" si los dos días anteriores + día en cuestión han sido EXTREMA_MINIMAS. Lo que realmente sería una ola de frio.
16) **INFLUENCIA_1_DIA** (S o N): Está a "S" si es el día siguiente a EXTREMA_MINIMAS.
17) **INFLUENCIA_2_DIA** (S o N): Está a "S" si está dentro de los dos días siguientes a EXTREMA_MINIMAS.
18) **INFLUENCIA_3_DIA** (S o N): Está a "S" si está dentro de los tres días siguientes a EXTREMA_MINIMAS.

19) **INFLUENCIA_SOLO_2_DIA** (S o N): Está a “S” si es el segundo día tras EXTREMA_MINIMAS.
20) **INFLUENCIA_SOLO_3_DIA** (S o N): Está a “S” si es el tercer día tras EXTREMA_MINIMAS.
21) **GRIPE**: Número de casos de gripe detectados en el día indicado.
22) **TEMPERATURA_MEDIA**: Temperatura media alcanzada en un día en concreto en una estación determinada.
23) **NUMERO_FALLECIMIENTOS**: Fallecimientos en el día indicado en las zonas a estudiar.

5. Estudio estadístico

En este apartado se explicará el estudio de correlación llevado a cabo en el proyecto. Se hará una pequeña introducción teórica a los estudios observacionales y la correlación (*Sección 5.1*), una explicación de la Regresión de Poisson (*Sección 5.2*) y de los Riesgos Relativos (*Sección 5.3*). A continuación se expondrán los resultados obtenidos (*Sección 5.4*).

Para poder observar si existía una correlación entre el número de visitas a urgencias y las temperaturas se empezó analizando la bibliografía referente a temas similares. Como se ha visto en la *Introducción* ([1]: Linwei Tian, et al*., 2016*, [2]: Anna Ponjoan, et al., 2017).y [3]: Tianqi Chen, et al., 2017) existía cierta correlación entre variables meteorológicas y clínicas por lo que se prosiguió a analizar la metodología utilizada. En la gran mayoría de estudios llevados a cabo la regresión de Poisson era la técnica más utilizada para la obtención de los riesgos relativos (RR).

5.1.Correlación y Estudio Observacional

En probabilidad y estadística, la correlación [8] indica la dirección y la fuerza de una relación lineal y la proporción entre dos variables estadísticas. Cuando los valores de una de las variables varían de forma sistemática con respecto a los mismos valores de la otra, se dice que existe una correlación entre variables cuantitativas. Por ejemplo, si tenemos dos variables (A y A2) existe correlación entre ellas si al disminuir los valores de A lo hacen también los de A2 y viceversa.

Hay que tener en cuenta que la correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad. Aunque no se debe concluir prematuramente que dos eventos correlacionados están ligados causalmente, una correlación puede ser un buen indicador de una relación causal. Aunque no siempre tiene porque ser así ya que un suceso puede ser multifactorial o puede ser causado por otro suceso no contemplado.

Un estudio observacional [9] es un estudio de carácter estadístico y demográfico, que pueden ser de tipo sociológico o biológico (estudios epidemiológicos) que están caracterizados en que no hay intervención por parte del investigador, y éste se limita a medir las variables que ha considerado en el estudio.

5.2.Regresión de Poisson (Poisson regression)

Una regresión de Poisson [10] es un tipo de modelo lineal generalizado de análisis regresivo que es usado en su mayoría de veces para modelar datos de conteo y tablas de contingencia. La regresión de Poisson asume que la variable independiente

"*Y*" sigue una distribución de Poisson y que el logaritmo de su valor esperado puede ser modelado por una combinación lineal de parámetros desconocidos.

La regresión de Poisson se utiliza para modelar fenómenos que pueden representarse mediante una variable aleatoria "*Y*" tal que para un valor $x \in \mathbb{R}^n$ de unas variables independientes,

$$Y|x \sim Poisson(\exp(a'x + b))$$

Es decir, el valor de "*Y*" condicionado a *x* sigue una distribución de Poisson de parámetro $\exp(a'x + b)$ para ciertos valores $a \in \mathbb{R}^n$ y $b \in \mathbb{R}$

Este modelo de Poisson es apropiado cuando la variable dependiente es un conteo, por ejemplo, número de llamadas a una central telefónica o como en nuestro caso, el número de visitas a urgencias que dependen de otras variables como, por ejemplo el día de la semana o en nuestro caso la temperatura de ese día. Los sucesos tienen que ser independientes.

En algunas ocasiones no se cumple la característica que debe cumplir una regresión de Poisson en dónde su media y su varianza deben ser iguales. Cuando la varianza es mayor que la media entonces puede indicar que el modelo no es apropiado y se produce una sobredispersión. Esto puede ser solucionado utilizando una quasi-Poisson en vez de una Poisson.

5.3.Riesgos Relativos (RR)

El riesgo relativo [11] es el ratio entre el riesgo en el grupo con el factor de exposición o factor de riesgo y el riesgo en el grupo de referencia (sin factor de exposición) como índice de asociación. Por tanto $RR = \frac{\text{incidencia acumulada en expuestos}}{\text{incidencia acumulada en no expuestos}}$

Es de gran utilidad utilizado en modelos de regresión, típicamente en el marco de regresiones de Poisson. Para su cálculo de forma manual se realiza una tabla como en la *Tabla 5.3.a:*

	Afectados	No Afectados	Total
Expuestos	a	b	a+b
No Expuestos	c	d	c+d
Total	a+c	b+d	N

Tabla 5.3.a: valores para el cálculo de Riesgos Relativos

Por tanto, $RR = \frac{a/(a+b)}{c/(c+d)}$

Características del riesgo relativo:

i. El riesgo relativo es una medida relativa del efecto porque indica cuanto más se tiende a desarrollar el evento en el grupo de sujetos expuestos al factor de exposición o factor de riesgo en relación con el grupo no expuesto.
ii. El riesgo relativo (RR) no tiene dimensiones.
iii. El rango de su valor oscila entre 0 e infinito.
iv. El RR=1 indica que no hay asociación entre la presencia del factor de riesgo y el evento.
v. El RR>1 indica que existe asociación positiva, es decir, que la presencia del factor de riesgo se asocia a una mayor frecuencia de suceder el evento. Cuanto mayor es el riesgo relativo, más fuerte es la prueba de una relación causal. Como se ha mencionado anteriormente, una alta correlación no implica causalidad por sí sola.
vi. El RR<1 indica que existe una asociación negativa, es decir, que no existe factor de riesgo, que lo que existe es un factor protector.
vii. El concepto de riesgo relativo es más difícil de interpretar que el de riesgo absoluto.
viii. Para interpretar el riesgo relativo se haría de la siguiente forma: un riesgo relativo de 20, quiere decir que los expuestos tienen 20 veces más probabilidad que los no expuestos de desarrollar la enfermedad.

5.4.Estudio y resultados

5.4.1. Python

Se han llevado a cabo dos implementaciones diferentes para observar el comportamiento de los datos. La primera se realizó en Python. En primer lugar se llevó a cabo la implementación de una regresión de Poisson para observar si existía correlación entre las variables de estudio y el número de visitas a urgencias mediante el lenguaje de programación Python. El dataset que se utilizó en un principio fue el mencionado en la *Sección 3.2.2: Pacientes*.

En principio se realizaron pruebas con todas las estaciones de Navarra y se vio que los resultados no eran muy fiables. Esto fue debido a que cada día debía tener un valor asignado que representara a todas las estaciones. En Navarra hay una gran diferencia en los valores ambientales en el mismo día dependiendo de la estación que lo mida. Debido a ello, para un día en concreto los valores se compensaban y un día en concreto podía ser considerado EXTREMA_MAXIMA (por ejemplo) para una estación y para otra no serlo. Debido a esto, se llegó al consenso de utilizar solamente pacientes y estaciones de la zona de Pamplona y comarca ya que los valores ambientales no cambian en exceso y la mayoría de pacientes se encuentran en esta zona.

Tras numerosas pruebas con diferentes variables se llegó a la decisión de volver a rehacer el dataset debido a que en epidemia de gripe las visitas o se mantenían sin cambios o incluso bajaban. Esto no cuadraba con lo que se sabe por parte del servicio de urgencias y del Complejo Hospitalario de Navarra ya que en esas fechas, se habilita una planta especial en el hospital y hay días en los que el servicio de urgencias se colapsa debido a que no pueden atender la demanda de pacientes.

Se procedió a extraer los pacientes de nuevo y ahora se comprobó que sí que en periodo de época de gripe el número de visitas a urgencias aumentaba. Debido a los cambios y eliminaciones llevados a cabo en la *Sección 3.2.2: Pacientes* y en la *Sección 3.2.1: Domicilios* era posible que se hubieran eliminado pacientes con valor aunque estos no tuvieran ni domicilio ni GMA ni CIAS. Por tanto, se concluyó que la manera más fiable de obtener resultados de valor era utilizar este dataset aunque no se tuviera cierta información de algunos pacientes. Para hacer algo más completo este dataset, se añadieron ciertos pacientes que habían fallecido entre los años de estudio y que en un principio no estaba contemplado. De esta forma los pacientes y las visitas a urgencias quedaron definidos de esta forma:

-Pacientes con >=65 años en el 2013, con CIAS y de la zona de Pamplona y comarca (Huarte, Villava, Burlada, Berriozar, San Jorge, Rochapea, Chantrea, Casco Viejo, II Ensanche, Milagrosa, Iturrama, San Juan, Ermitagaña, Zizur-Echavacoiz, Barañain, Azpilagaña, Mendillorri, Ansoain, Buztintxuri y Sarriguren) → 50127

-De esos que hayan visitado urgencias entre el 1 de enero del 2013 y el 31 de diciembre del 2016 en el Complejo Hospitalario de Navarra → 74553

-Pacientes con >=65 años en el 2013, sin CIAS, que hayan fallecido, que estuvieran vivos el 01-01-2013 y de la zona de Pamplona y comarca (Huarte, Villava, Burlada, Berriozar, San Jorge, Rochapea, Chantrea, Casco Viejo, II Ensanche, Milagrosa, Iturrama, San Juan, Ermitagaña, Zizur-Echavacoiz, Barañain, Azpilagaña, Mendillorri, Ansoain, Buztintxuri y Sarriguren) →10197

-De los pacientes con >=65 años en el 2013, sin CIAS, que hayan fallecido y de la zona de Pamplona y comarca (Huarte, Villava, Burlada, Berriozar, San Jorge, Rochapea, Chantrea, Casco Viejo, II Ensanche, Milagrosa, Iturrama, San Juan, Ermitagaña, Zizur-Echavacoiz, Barañain, Azpilagaña, Mendillorri, Ansoain, Buztintxuri y Sarriguren) que hayan visitado urgencias entre el 1 de enero del 2013 y el 31 de diciembre del 2016 en el Complejo Hospitalario de Navarra →30145

-Pacientes totales: 60324

-Urgencias totales: 104698

Además para poder asignarles unos valores de alguna estación se escogieron 5 estaciones que englobaban a todas esas zonas básicas y que a priori el valor para un día no cambiaba mucho debido a la proximidad de las mismas. Estas estaciones fueron: Pamplona GN, Pamplona MAN, Pamplona GN, Irotz y Arazuri INTIA.

Para decidir si un día en concreto era considerado positivo o negativo, al menos 3 de las 5 estaciones tenían que tener el valor a considerar para evitar errores.

Se procedió a realizar multitud de pruebas con diferentes variables para observar los resultados obtenidos y poder sacar soluciones. A continuación se muestran varias ejecuciones con los resultados:

-**Función_1**: = 'NUMERO_VISITAS ~ EXTREMA_MAXIMAS + EXTREMA_MINIMAS + OLA_DE_CALOR + OLA_DE_FRIO + FESTIVO + EPIDEMIA'

					NUMERO_VISITAS	
				EPIDEMIA	**N**	**S**
EXTREMA_MAXIMAS	**EXTREMA_MINIMAS**	**OLA_DE_CALOR**	**OLA_DE_FRIO**	**FESTIVO**		
N	**N**	**N**	**N**	**N**	67.357616	68.942652
				S	55.745819	59.801471
		S	**N**	**N**	69.000000	NaN
				S	64.000000	NaN
	S	**N**	**N**	**N**	69.666667	70.558824
				S	61.000000	61.705882
			S	**N**	67.000000	75.000000
				S	62.000000	63.000000
S	**N**	**N**	**N**	**N**	64.115385	NaN
				S	57.812500	NaN
		S	**N**	**N**	64.684211	NaN

Tabla 5.4.1.a: agrupación de las visitas a urgencias en base a las variables descritas en la función

	mean	hpd_2.5	hpd_97.5
Intercept	6.698443e+01	66.346269	6.761415e+01
EXTREMA_MAXIMAS[T.S]	9.756357e-01	0.937510	1.021227e+00
EXTREMA_MINIMAS[T.S]	1.030856e+00	1.005402	1.065062e+00
OLA_DE_CALOR[T.S]	1.005371e+00	0.958203	1.071320e+00
OLA_DE_FRIO[T.S]	1.040106e+00	0.954318	1.135137e+00
FESTIVO[T.S]	8.441270e-01	0.835158	8.582076e-01
EPIDEMIA[T.S]	1.037114e+00	1.023054	1.054154e+00
mu	4.138345e+15	1.015301	1.733041e+57

Tabla 5.4.1.b: resultados obtenidos tras aplicar la función de Poisson a las variables mencionadas. Se destacan en rojo las variables con diferencias estadísticas significativas.

-**Función_2**: 'NUMERO_VISITAS ~ EXTREMA_MINIMAS + OLA_DE_FRIO + FESTIVO + EPIDEMIA + INFLUENCIA_1_DIA + INFLUENCIA_2_DIA + INFLUENCIA_3_DIA + INFLUENCIA_SOLO_2_DIA + INFLUENCIA_SOLO_3_DIA'

	mean	hpd_2.5	hpd_97.5
Intercept	6.686087e+01	66.267653	6.744284e+01
EXTREMA_MINIMAS[T.S]	1.030122e+00	1.004090	1.067075e+00
OLA_DE_FRIO[T.S]	1.031581e+00	0.934865	1.121465e+00
FESTIVO[T.S]	8.436939e-01	0.832244	8.568469e-01
EPIDEMIA[T.S]	1.036183e+00	1.024221	1.053391e+00
INFLUENCIA_1_DIA[T.S]	9.977431e-01	0.939232	1.066855e+00
INFLUENCIA_2_DIA[T.S]	9.669932e-01	0.894080	1.048800e+00
INFLUENCIA_3_DIA[T.S]	1.029294e+00	0.969288	1.077442e+00
INFLUENCIA_SOLO_2_DIA[T.S]	1.018104e+00	0.957584	1.085739e+00
INFLUENCIA_SOLO_3_DIA[T.S]	1.004423e+00	0.967563	1.056609e+00

	mean	hpd_2.5	hpd_97.5
mu	3.589935e+09	1.219937	2.952088e+30

Tabla 5.4.1.c: resultados obtenidos tras aplicar la función de Poisson a las variables mencionadas. Se destacan en rojo las variables con diferencias estadísticas significativas.

-**Función_3**: 'NUMERO_VISITAS ~ EXTREMA_MAXIMAS + OLA_DE_CALOR + FESTIVO + EPIDEMIA + INFLUENCIA_1_DIA_MAX + INFLUENCIA_2_DIA_MAX + INFLUENCIA_3_DIA_MAX + INFLUENCIA_SOLO_2_DIA_MAX + INFLUENCIA_SOLO_3_DIA_MAX'

	mean	hpd_2.5	hpd_97.5
Intercept	6.701775e+01	66.473498	6.757685e+01
EXTREMA_MAXIMAS[T.S]	9.686417e-01	0.935480	1.010326e+00
OLA_DE_CALOR[T.S]	1.058951e+00	0.971041	1.146411e+00
FESTIVO[T.S]	8.454673e-01	0.832677	8.570569e-01
EPIDEMIA[T.S]	1.040466e+00	1.023699	1.055616e+00
INFLUENCIA_1_DIA_MAX[T.S]	9.309272e-01	0.861682	1.003598e+00
INFLUENCIA_2_DIA_MAX[T.S]	1.096589e+00	0.981666	1.226131e+00
INFLUENCIA_3_DIA_MAX[T.S]	9.909579e-01	0.919754	1.062571e+00
INFLUENCIA_SOLO_2_DIA_MAX[T.S]	9.350186e-01	0.859314	9.987115e-01
INFLUENCIA_SOLO_3_DIA_MAX[T.S]	1.006718e+00	0.963487	1.065212e+00
mu	1.601002e+90	1.520807	1.318131e+283

Tabla 5.4.1.d: resultados obtenidos tras aplicar la función de Poisson a las variables mencionadas. Se destacan en rojo las variables con diferencias estadísticas significativas.

-**Función_4**: 'NUMERO_VISITAS ~ EXTREMA_MAXIMAS + EXTREMA_MINIMAS + OLA_DE_CALOR + OLA_DE_FRIO + FESTIVO + EPIDEMIA+DIA_SEMANA'

	mean	**hpd_2.5**	**hpd_97.5**
Intercept	69.489247	67.287568	7.149475e+01
DIA_SEMANA_LETRA[T.J]	1.036438	0.998285	1.068179e+00
DIA_SEMANA_LETRA[T.L]	1.165878	1.135140	1.209144e+00
DIA_SEMANA_LETRA[T.M]	1.048665	1.014187	1.081913e+00
DIA_SEMANA_LETRA[T.S]	0.982154	0.958950	1.005875e+00
DIA_SEMANA_LETRA[T.V]	1.096840	1.062818	1.131694e+00
DIA_SEMANA_LETRA[T.X]	1.045820	1.016392	1.080000e+00
EXTREMA_MAXIMAS	0.976490	0.941614	1.011799e+00
EXTREMA_MINIMAS	1.024453	0.992886	1.051232e+00
OLA_DE_CALOR	1.000044	0.955595	1.064631e+00
OLA_DE_FRIO	1.001859	0.935965	1.073582e+00
EPIDEMIA	1.034725	1.020847	1.050161e+00
FESTIVO	0.898416	0.876434	9.219242e-01
mu	250642.513160	11.695738	2.850472e+14

Tabla 5.4.1.e: resultados obtenidos tras aplicar la función de Poisson a las variables mencionadas. Se destacan en rojo las variables con diferencias estadísticas significativas.

Como se ha explicado en el apartado teórico los riesgos relativos son la medida utilizada en los resultados de una regresión de Poisson. En las tablas de resultados para cada función se puede observar: la variable estudiada, el valor del riesgo relativo en media, el valor del riesgo relativo para el percentil 2.5 y el valor del riesgo relativo para el percentil 97.5. Como resultado, si los valores de los riesgos relativos de los dos percentiles son mayores que la unidad, entonces tenemos evidencias estadísticas significativas de que esa variable está correlacionada positivamente con el número de visitas. De la misma forma, si ambos percentiles se encuentran por debajo de la unidad entonces la variable analizada está correlacionada de forma negativa con el número de visitas.

Vistos estos resultados se observa que en FESTIVO las visitas siempre bajan respecto al número medio de estas. Además también se puede observar que cuando hay EPIDEMIA las visitas a urgencias siempre aumentan así como cuando hay un día con EXTREMA_MINIMAS. También se puede observar como los lunes y los viernes son los días de la semana que más visitas albergan. Por último, se observa un efecto de retardo de 2 días cuando ha habido una temperatura EXTREMA_MAXIMA en dónde las visitas disminuyen.

5.4.2. R

El dataset utilizado en esta segunda parte es el de las personas >=65 años utilizado en el apartado anterior. Se llevó a cabo una implementación en R en dónde se han utilizado funciones quasi-Poisson. Se ha utilizado el estudio de Antonio Gasparrini [12] como base para llevar a cabo este apartado y se ha contado con la colaboración de Ibai Tamayo y Julián Librero de Navarrabiomed. Hay que comentar que se han elegido las variables GRIPE, TEMPERATURA_MEDIA, FESTIVO Y EPIDEMIA en esta implementación ya que son las que más influyen en el número de visitas a urgencias. De la misma forma se creyó que era de interés el evaluar el número de fallecimientos de la misma forma que se va a evaluar el número de visitas. Por tanto cada procedimiento será aplicado a las visitas a urgencias y al número de fallecimientos.

A continuación se muestran los valores en el tiempo de las variables NUMERO_VISITAS, GRIPE, TEMPERATURA_MEDIA Y NUMERO_FALLECIMIENTOS.

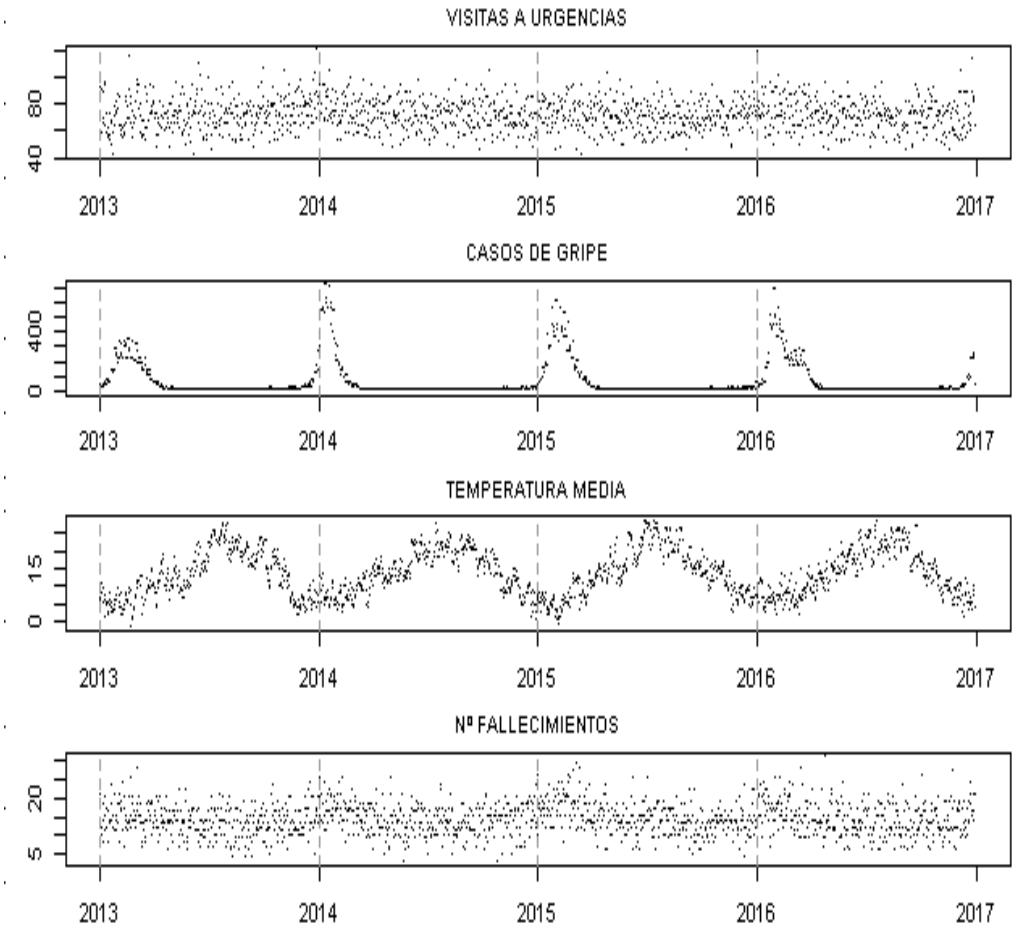


Figura 5.4.2.a: representación temporal de las variables NUMERO_VISITAS, GRIPE, TEMPERATURA_MEDIA Y NUMERO_FALLECIMIENTOS

Tras ello, se obtuvo una matriz de correlaciones entre las variables con el fin de observar qué variables pueden solapar los resultados de otras.

	NUMERO_VISITAS	GRIPE	FESTIVO	EPIDEMIA
NUMERO_VISITAS	1.00000000	0.12754669	-0.502601300	0.114847901
GRIPE	0.12754669	1.00000000	-0.010527995	0.648569076
FESTIVO	-0.50260130	-0.01052799	1.000000000	-0.004194726
EPIDEMIA	0.11484790	0.64856908	-0.004194726	1.000000000
TEMPERATURA_MEDIA	-0.04025638	-0.48523399	-0.004081647	-0.664760273
NUMERO_FALLECIMIENTOS	0.07613611	0.26969350	-0.049361945	0.290726213

	TEMPERATURA_MEDIA	NUMERO_FALLECIMIENTOS
NUMERO_VISITAS	-0.040256376	0.07613611
GRIPE	-0.485233991	0.26969350
FESTIVO	-0.004081647	-0.04936195
EPIDEMIA	-0.664760273	0.29072621
TEMPERATURA_MEDIA	1.000000000	-0.17780650
NUMERO_FALLECIMIENTOS	-0.177806503	1.00000000

Tabla 5.4.2.a: matriz de correlaciones entre las variables seleccionadas

Se puede observar como NUMERO_VISITAS tiene bastante correlación negativa con FESTIVO, GRIPE una correlación positiva con EPIDEMIA y EPIDEMIA una correlación negativa con TEMPERATURA_MEDIA.

Es de vital importancia recalcar que en esta parte se va a diferenciar entre modelo sin ajustar y modelo ajustado por estacionalidad. La estacionalidad es la variación periódica y por tanto predecible de una serie temporal. Un ejemplo claro sería la que se produce a lo largo de un año en relación a las temperaturas (verano e inverno), precipitaciones (estaciones húmedas y estaciones secas), duración de las horas de sol (más en verano, menos en inverno), etc. Aunque esto puede darse no solo en factores ambientales sino también en economía (las épocas navideñas siempre hacen que aumente el volumen de ventas), turismo (en verano y época de esquí aumentan de forma significativa las reservas en hoteles) y muchos campos más.

A la hora de analizar una serie temporal, como es este caso, es importante tener en cuenta que la estacionalidad puede llegar a jugar un papel clave. Para nuestro caso en concreto hay que analizar si las visitas a urgencias poseen un papel estacional (si siempre se producen en determinadas épocas del año) y cómo se podría llegar a analizar.

Para poder observar efectos en la serie temporal que no tengan que ver con la estacionalidad se lleva a cabo un método estadístico llamado desestacionalización o ajuste estacional. Este método consiste en eliminar el efecto estacional de una serie temporal con el fin de analizarla, compararla y observar la tendencia habiendo compensado los efectos estacionales.

Se van a llevar a cabo varias pruebas en donde se mostrarán resultados del modelo sin eliminar la estacionalidad y del modelo sin estacionalidad. Para este último se van a aplicar tres técnicas diferentes:

1. Mediante modelo de series estratificado: Una manera de controlar la estacionalidad es dividiendo el periodo de estudio en intervalos y estimar para cada intervalo un número base diferente de visitas a urgencias. Para ello, se introduce una variable indicadora para cada intervalo en el modelo. En este caso, aplicaremos un intervalo de un mes prediciendo, para cada día, una nueva variable con el número de visitas a urgencias al mes en cada año.

2. Mediante funciones de Fourier: Otra manera de controlar la estacionalidad es utilizando funciones de Fourier. Esto son pares de funciones seno-coseno representado un ciclo estacional completo (en nuestro caso un año natural) y son útiles para capturar patrones estacionales muy regulares, una característica que no es muy favorable para nuestros datos.

3. Mediante funciones spline: La última técnica implementada es mediante funciones spline. Son básicamente un número de diferentes curvas polinomiales (normalmente cúbicas) que son unidas punto a punto para cubrir todo el periodo. Para ello, se generarán un conjunto de variables base

las cuales son funciones de la variable temporal (en nuestro caso la variable FECHA) que se incluirán en el modelo.

Finalmente se visualizarán los residuos generados por los modelos sin estacionalidad con el fin de observar si se ha conseguido el propósito planteado.

A continuación se van a mostrar tres técnicas utilizadas para eliminar la estacionalidad junto con los residuos obtenidos tras aplicar estas técnicas.

-Para las visitas a urgencias:

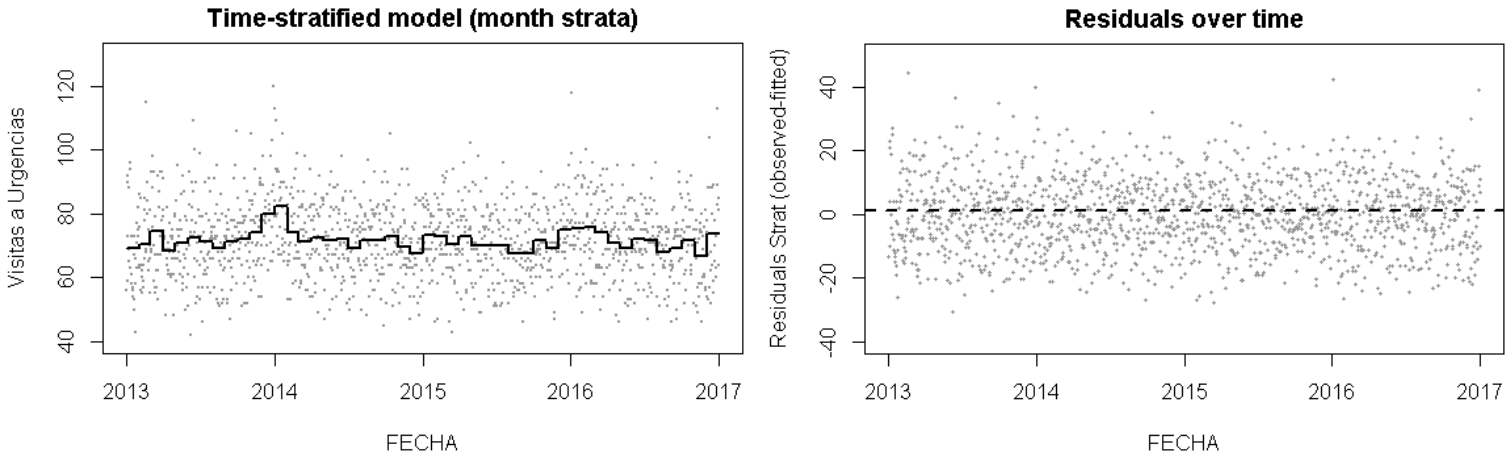
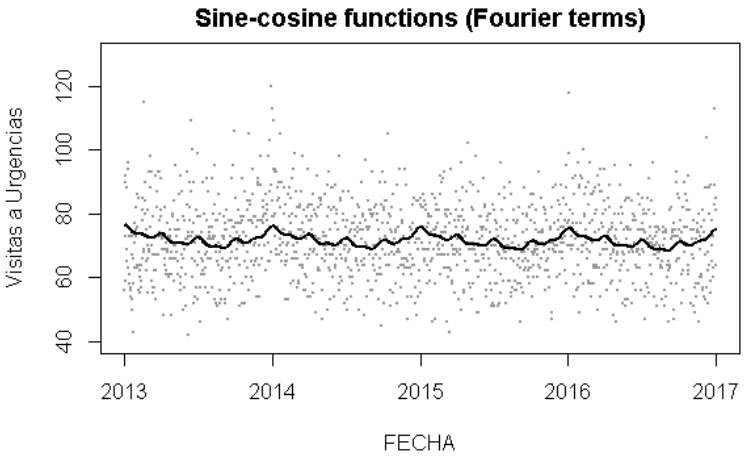


Figura 5.4.2.b: efecto de la estacionalidad mediante modelo de series estratificado y los residuos generados.



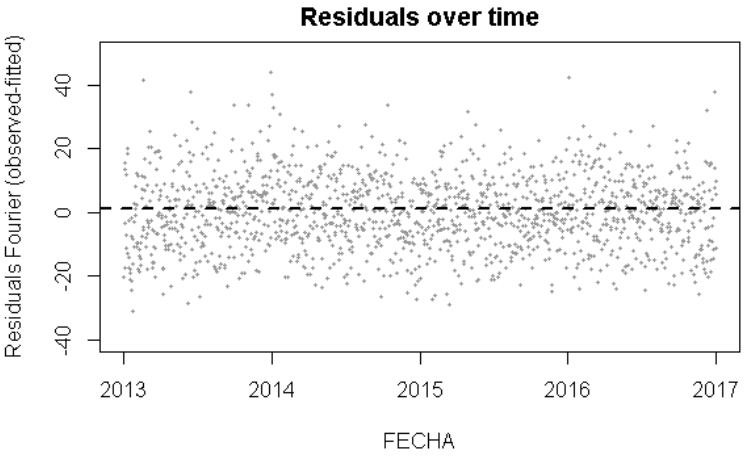
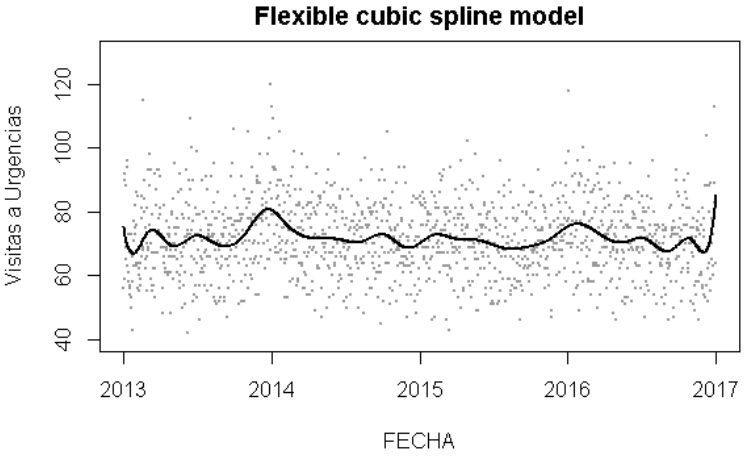


Figura 5.4.2.c: efecto de la estacionalidad mediante funciones de Fourier y los residuos generados



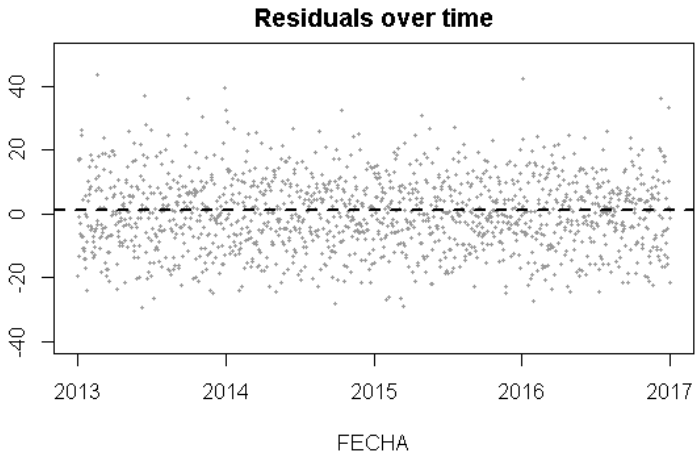


Figura 5.4.2.d: efecto de la estacionalidad en base a funciones spline y los residuos generados.

-Para los fallecimientos:

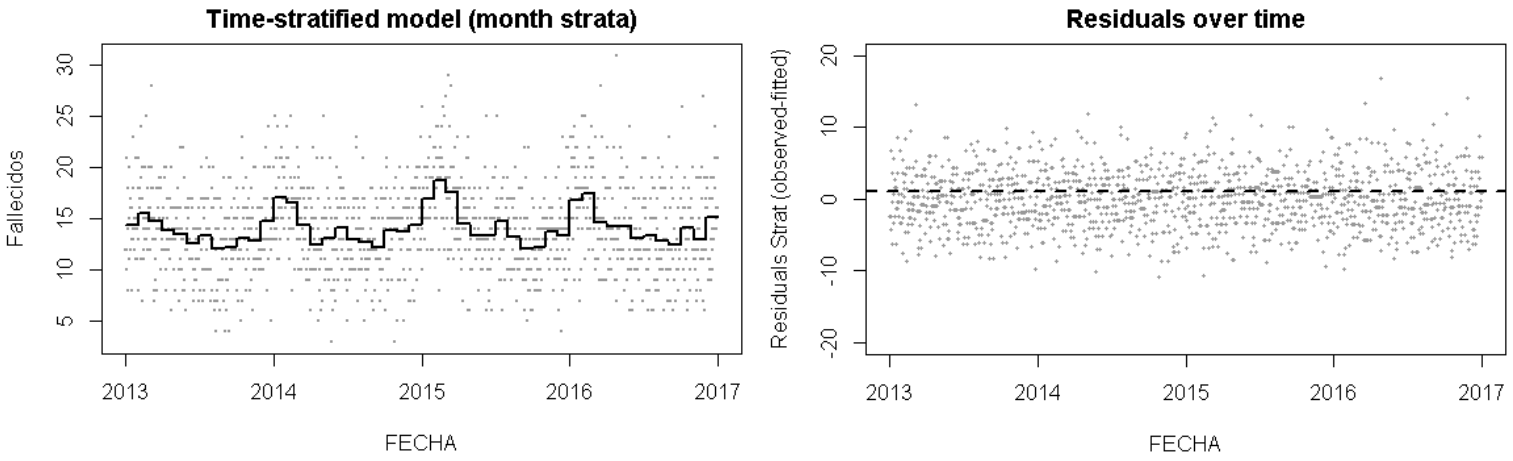


Figura 5.4.2.e: efecto de la estacionalidad mediante modelo de series estratificado y los residuos generados.

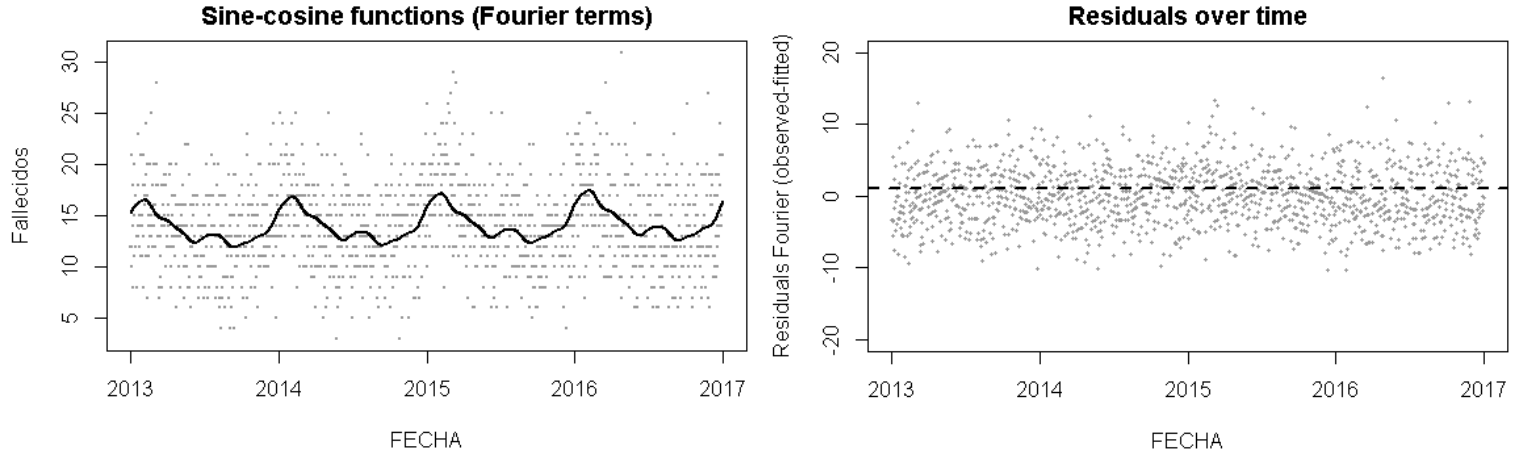


Figura 5.4.2.f: efecto de la estacionalidad mediante funciones de Fourier y los residuos generados

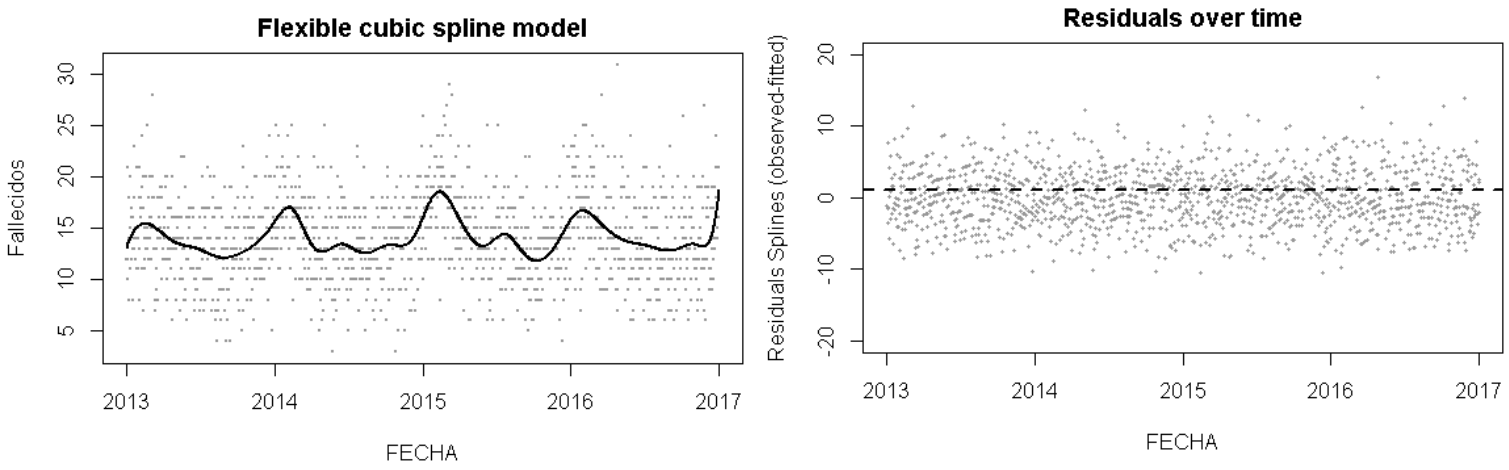


Figura 5.4.2.g: efecto de la estacionalidad en base a funciones spline y los residuos generados.

Se van a evaluar los resultados obtenidos por las funciones de quasi-Poisson para el modelo sin ajustar y para el modelo en bases a funciones spline ya que es la que ofrece un resultado con una estacionalidad menor de los modelos previamente analizados. Se va a crear una nueva variable "Gripe100" que indica el número de casos de gripe por cada 100 casos de gripe reales para poder obtener una perspectiva más real en cuanto a resultados. Es decir, por cada día se divide el número de casos de gripe entre 100 para obtener una escala más simplificada.

<u>-Para las visitas a urgencias:</u>

	Exp(Est)	2,5%	97,5%
Gripe100	1.0170	1.0105	1.0235
FESTIVO	0.8363	0.8233	0.8495
TEMPERATURA_MEDIA	1.0005	0.9993	1.0018

Tabla 5.4.2.b: resumen simplificado del efecto sin ajuste sobre el número de visitas a urgencias

	Exp(Est)	2,5%	97,5%
Gripe100	1.0060	0.9966	1.0155
FESTIVO	0.8357	0.8231	0.8485
TEMPERATURA_MEDIA	1.0063	1.0040	1.0087

Tabla 5.4.2.c: resumen simplificado del efecto ajustado sobre el número de visitas a urgencias

Las visitas a urgencias de la población de este dataset son mayores cuando hay más diagnósticos de gripe en AP.

Sin embargo, el efecto de la gripe sobre éstas desaparece al aplicar desestacionalidad. Esto, por otra parte, es lógico ya que la gripe siempre se da en la misma época del año y sin estacionalidad éste hecho desaparece.

Las visitas a urgencias de la población de este dataset son menores en festivos y al aplicar desestacionalidad sigue visitando menos gente en festivo, si bien con un efecto marginalmente inferior.

Cuando eliminamos el efecto de la estacionalidad, la población visita más urgencias cuando se eleva la temperatura media.

En conjunción con los dos modelos podemos concluir que: si aumenta la temperatura se incrementa el número de visitas a urgencias, en festivos la población va menos a urgencias y a más casos de gripe se incrementa el número de visitas a urgencias.

-Para los fallecimientos:

	Exp(Est)	2,5%	97,5%
Gripe100	1.0520	1.0391	1.0651
FESTIVO	0.9703	0.9403	1.0013
TEMPERATURA_MEDIA	0.9969	0.9943	0.9995

Tabla 5.4.2.d: resumen simplificado del efecto sin ajuste sobre el número de fallecimientos

	Exp(Est)	2,5%	97,5%
Gripe100	0.9999	0.9813	1.0189
FESTIVO	0.9734	0.9440	1.0037
TEMPERATURA_MEDIA	1.0113	1.0065	1.0162

Tabla 5.4.2.e: resumen simplificado del efecto ajustado sobre el número de fallecimientos

La mortalidad de la población de este dataset es mayor cuando hay más diagnósticos de gripe en AP.

Sin embargo, el efecto de la gripe sobre las muertes desaparece al aplicar desestacionalidad igual que para las visitas a urgencias.

No se observa un incremento de la mortalidad en festivos en ninguno de los dos modelos.

En el modelo sin ajustar parece que el incremento de la temperatura media hace que muera menos gente (más gente fallece por temperaturas bajas). Sin embargo, al ajustar por desestacionalidad este efecto se revierte: la gente muere más con mayores temperaturas medias. Esto puede deberse a que al eliminar la estacionalidad, la gripe no produce tantos muertos y entonces los días con temperaturas más altas, como en verano, hacen que fallezcan más personas.

En conjunción con los dos modelos podemos concluir: disminuir la temperatura media incrementa el número de muertes y a más casos de gripe se incrementa la mortalidad. El hecho de eliminar la estacionalidad hace que el efecto de la gripe desaparezca y saca a relucir el efecto de la temperatura media llegando al resultado de a más temperatura media más fallecimientos.

Por último, vamos a estudiar el efecto retardo de estas variables que pueden explicar ciertos comportamientos vistos hasta ahora. El "*Lag-effect"* o efecto retardo es una situación en dónde los valores de una variable en un día en concreto afectan a los valores de otras variables en otros días diferentes. En otras palabras, puede haber una correlación entre una variable hoy y otra variable otro día diferente. Para el caso tratado un ejemplo sería: El aumento de la temperatura media hoy podría explicar el aumento en el número de visitas de mañana. Esto se definiría como lag 1 ya que hay un día de diferencia entre el desencadenante (aumento temperatura) y el consecuente (aumento mortalidad).

Para modelar esto hay distintas técnicas que se diferencian en cómo tratan los valores de los diferentes días a la hora de introducirlos en el modelo.

Se van a mostrar tres técnicas para modelar estos retardos:

1. Sin ajustar entre días: Cada retardo no está ajustado entre días; los retardos han sido ajustados a la vez.
2. Ajustado entre días: Los días son introducidos en el modelo de forma simultánea. Esto se conoce como modelo de retardo distribuido. Esta técnica puede evitar que los efectos retardo estén confundidos entre ellos.
3. Ajustado entre días con una restricción: La gran desventaja de la técnica anterior es que las condiciones de los retardos puede que estén altamente correlacionadas y la colinealidad en el modelo puede resultar en estimaciones imprecisas (amplios rangos de confianza). Para solucionarlo es posible imponer una restricción en el efecto estimado por los diferentes retardos.

A continuación se muestran los resultados de estas técnicas tanto para las visitas a urgencias como para los fallecidos.

-Para las visitas a urgencias:

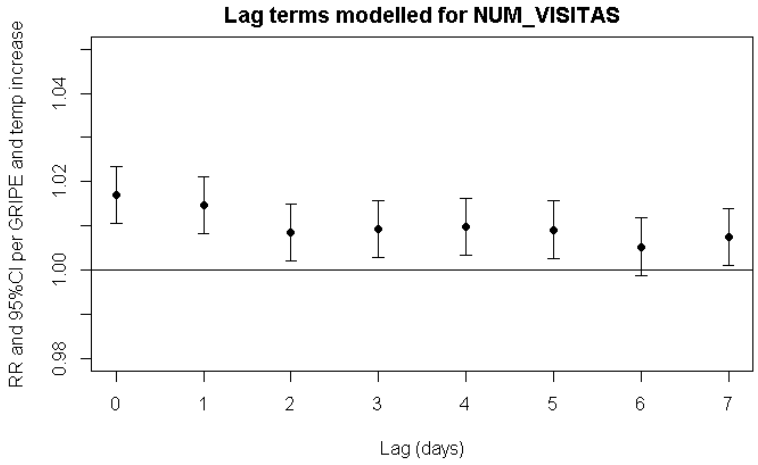


Figura 5.4.2.h: Efecto del retardo en el número de visitas para el modelo con estacionalidad sin ajustar entre días.

En este caso no se observa una correlación clara entre días ya que todos ellos presentan diferencias estadísticas (el RR es mayor que 1) y no hay diferencias entre días.

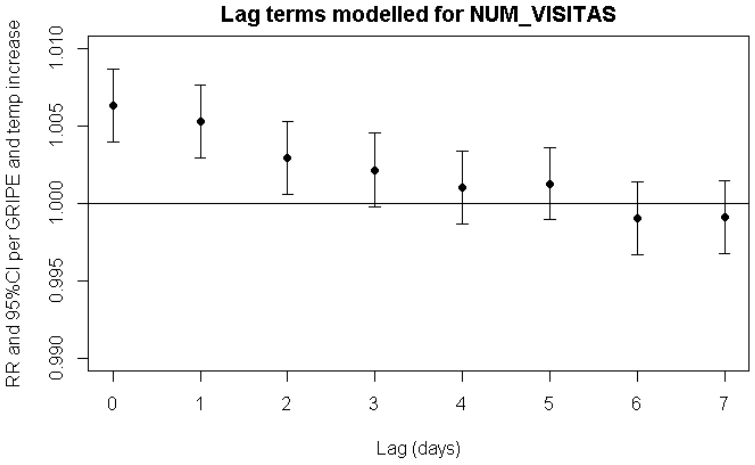


Figura 5.4.2.i: Efecto del retardo en el número de visitas para el modelo sin estacionalidad sin ajustar entre días.

Observando esta figura, no podemos concretar nada debido a que el día 0 posee más RR que el resto y esto nos indica que el mayor efecto en el número de visitas se produce el mismo día que aumentan los casos de gripe y la temperatura y no a los días siguientes.

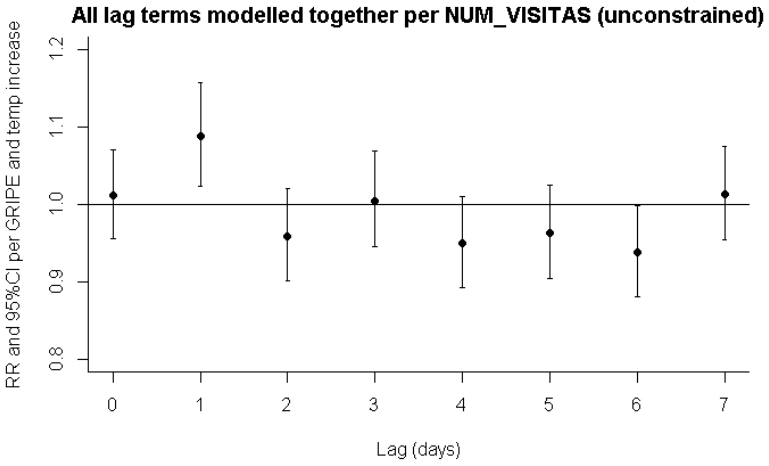


Figura 5.4.2.j: Efecto del retardo en el número de visitas para el modelo sin estacionalidad ajustado entre días.

En este caso, se puede observar cómo es posible que exista un lag de 1 día debido a que el RR del día 1 es mayor que el día 0. El resto de días no presentan diferencias significativas o están muy cerca del valor del día 0 por lo que no se ve una correlación clara.

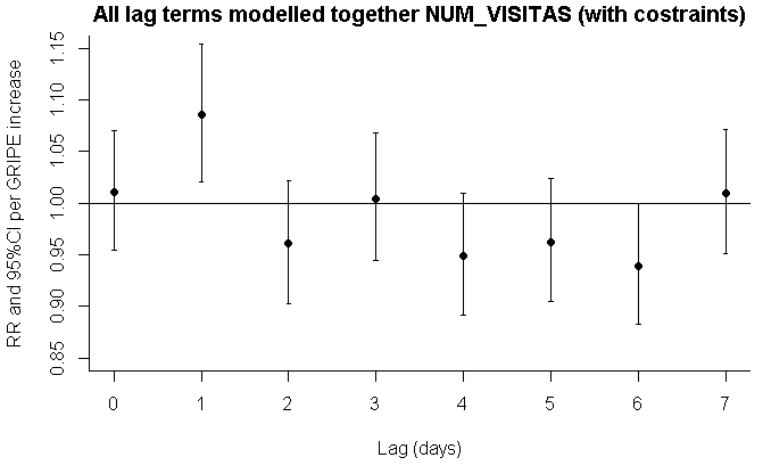


Figura 5.4.2.k: Efecto del retardo en el número de visitas para el modelo sin estacionalidad en base a los casos de gripe ajustado entre días en base a una restricción.

Se observa el mismo comportamiento que el gráfico anteriormente descrito.

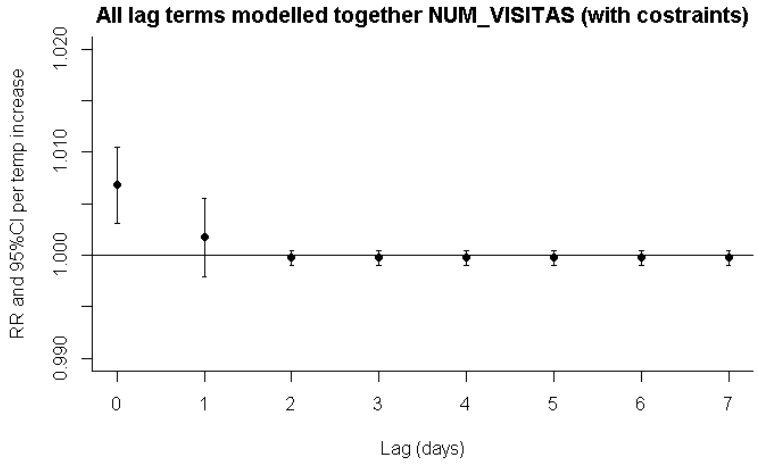


Figura 5.4.2.l: Efecto del retardo en el número de visitas para el modelo sin estacionalidad en base a la temperatura media ajustado entre días en base a una restricción.

En este caso no podemos concluir nada debido a que el riesgo relativo del día 0 es mayor que el del resto de días por lo que no se observaría un efecto retardo.

Por tanto, podemos concluir que en estas figuras puede que exista un efecto retardo de 1 día entre el número de visitas y los casos de gripe. Por tanto, si un día ha habido gran cantidad de casos de gripe esto se traduciría en un aumento del número de visitas al día siguiente.

<u>-Para los fallecimientos</u>

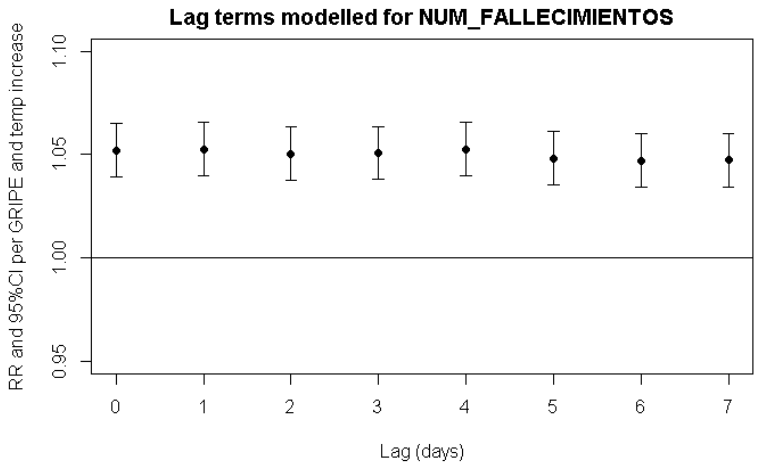


Figura 5.4.2.m: Efecto del retardo en el número de fallecidos para el modelo con estacionalidad sin ajustar entre días.

En este caso ocurre lo mismo que para las visitas a urgencias en este modelo. No se observa una correlación clara entre días ya que todos ellos presentan diferencias estadísticas (el RR es mayor que 1) y no hay diferencias entre días.

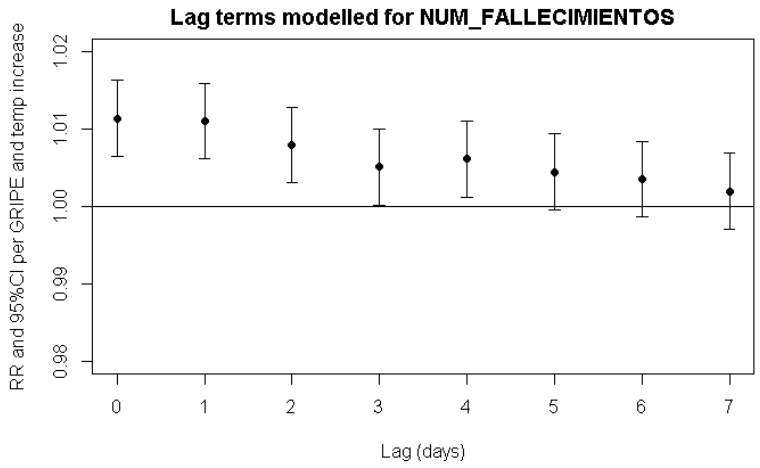


Figura 5.4.2.n: Efecto del retardo en el número de fallecidos para el modelo sin estacionalidad sin ajustar entre días.

Ocurre lo mismo que en el caso anterior.

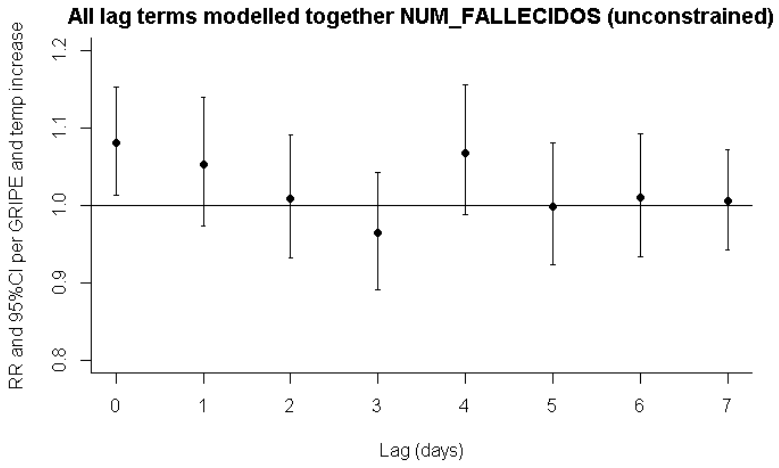


Figura 5.4.2.o: Efecto del retardo en el número de muertes para el modelo sin estacionalidad ajustado entre días.

En este modelo se observa que el día 4 y el día 0 poseen riesgos relativos similares y el resto de días las diferencias estadísticas son menores. Es por ello que no se puede hablar de efecto retardo claro.

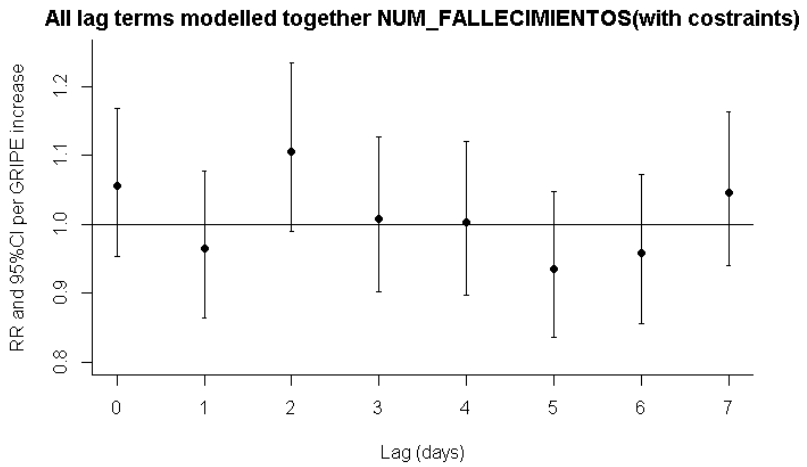


Figura 5.4.2.p: Efecto del retardo en el número de fallecimientos para el modelo sin estacionalidad en base a los casos de gripe ajustado entre días en base a una restricción.

En este caso sí que se observa un ligero aumento del riesgo relativo en el día 2 respecto al día 0 que podría explicar un retardo en los fallecimientos en base a los casos de gripe.

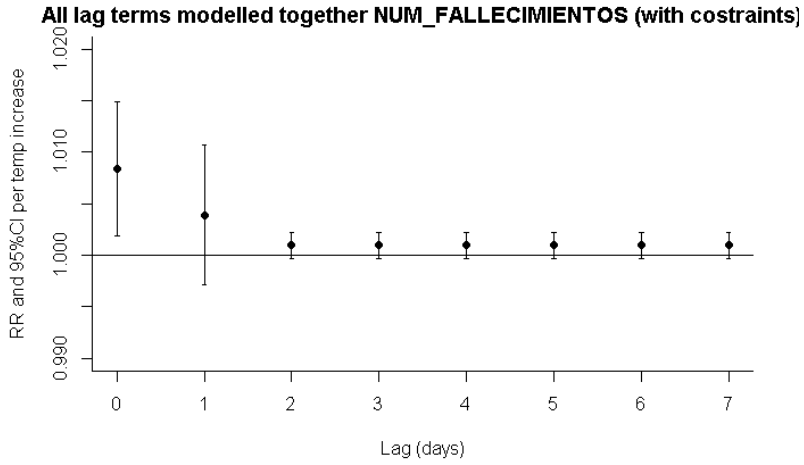


Figura 5.4.2.q: Efecto del retardo en el número de fallecimientos para el modelo sin estacionalidad en base a temperatura media ajustado entre días en base a una restricción.

Para este último gráfico no se observa nada reseñable.

Lo que podemos concluir de estas figuras es que puede que exista un efecto retardo de 2 días entre el número de muertes y los casos de gripe. Por tanto, si un día ha habido gran cantidad de casos de gripe esto se traduciría en un aumento del número de muertes a los 2 días.

-Conclusiones:

Una vez obtenidos los resultados de las dos implementaciones podemos concluir que: en FESTIVO se detectan menos visitas a urgencias; en EPIDEMIA tanto el número de visitas como el número de fallecimientos aumenta; cuando la TEMPERATURA_MEDIA aumenta (y por tanto lo hacen la TEMPERATURA_MAXIMA y la TEMPERATURA_MINIMA) hay más visitas a urgencias y más fallecimientos.

Como información adicional se puede afirmar que cuando hay un día con EXTREMA_MINIMAS el número de visitas aumenta aunque esto está relacionado con la variable EPIMEDIA ya que la correlación entre estas variables es clara. Además los lunes y viernes las visitas a urgencias son mayores respecto al resto de días de la semana y puede que exista un retardo de 1 día en los casos de gripe para las visitas a urgencias y 2 días en los casos de gripe para los fallecimientos.

6. Predicción a nivel de paciente

En este apartado se explicará el proceso realizado para intentar predecir si un paciente en concreto visitará urgencias en base a las variables a nivel de paciente creadas y las temperaturas. Primero se realizará una explicación de conceptos teóricos como Aprendizaje supervisado (*Sección 6.1*), Clasificación (*Sección 6.2*) y las medidas de rendimiento de un clasificador (*Sección 6.3*). Además se explicará el concepto de Regresión Logística (*Sección 6.4*) y Naive Bayes (*Sección 6.5*). A continuación se expondrán los resultados obtenidos (*Sección 6.6*).

Los antecedentes clínicos de un paciente nos muestran gran parte de la vida sanitaria de la persona y cómo ha influido en la misma. Debido a que a priori ofrecen bastante información sobre cada paciente se lanzó un modelo predictivo con el fin de poder predecir si un paciente iba a ir un día en concreto a urgencias en base a la temperatura de ese día. Para ello, una vez extraída la información relativa a las temperaturas y obtenidas las variables de cada paciente (*Ver Sección 4.1: Variables a nivel de paciente*) se procedió a montar el dataset. Este dataset se construyó con los pacientes mayores o iguales a 65 años descrito en la *Sección 3.2.2: Pacientes.*

Para la predicción se utilizaron modelos de clasificación ya que son los indicados para discernir en dos clases un conjunto de datos. En nuestro caso, las dos clases se dividían en gente que iba a urgencias y gente que no iba a urgencias.

Se probaron varios modelos de clasificación con el fin de obtener diferentes resultados. Los modelos utilizados fueron: Regresión logística, Gaussian Naive Bayes y Bernoulli Naive Bayes.

6.1.Aprendizaje supervisado

El aprendizaje supervisado [18] es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo es crear una función capaz de predecir a partir de una serie de ejemplos (entrenamiento), el valor correspondiente a cualquier objeto de entrada.

6.2.Clasificación

Otro término para el aprendizaje supervisado es la clasificación. Una amplia gama de clasificadores están disponibles, cada uno con sus fortalezas y debilidades. Los clasificadores más utilizados son las redes neuronales, (como el perceptrón

multicapa); las máquinas de vectores de soporte; el algoritmo de los K-vecinos más cercanos, los modelos de mixturas; el clasificador bayesiano ingenuo; los árboles de decisión y las funciones de base radial.

6.3.Medidas de rendimiento de un clasificador

Para valorar cómo se comporta un algoritmo de clasificación existen distintas medidas que nos proporcionan información para valorar el trabajo del clasificador [19]:

1) Accuracy:

Es el número de predicciones correctas dividido entre el número total de predicciones realizadas. También se puede expresar en porcentaje.

2) Matriz de confusión (o tabla de contingencia):

Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases o en qué está fallando más a la hora de clasificar.

Si en los datos de entrada el número de muestras de clases diferentes cambia mucho, la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. Si por ejemplo hay 990 muestras de la clase 1 y sólo 10 de la clase 2, el clasificador puede tener fácilmente un sesgo hacia la clase 1. Si el clasificador clasifica todas las muestras como clase 1 su precisión será del 99%. Esto no significa que sea un buen clasificador, pues tuvo un 100% de error en la clasificación de las muestras de la clase 2.

	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Tabla 6.3.a: Matriz de confusión para clasificación binaria

3) Precision:

Es el número de Verdaderos Positivos (True Positives) dividido entre la suma de Verdaderos Positivos (True Positives) y Falsos Positivos (False Positives). Es una medida de exactitud.

4) **Recall:**

Es el número de Verdaderos Positivos (True Positives) dividido entre la suma de Verdaderos Positivos (True Positives) y Falsos Negativos (False Negatives). Es una medida de integridad.

5) **F1-score:**

Se calcula como: 2*((precision*recall)/ (precision+recall)). Muestra un balance entre Precision y Recall.

6) **Área bajo la curva ROC:**

La curva ROC [21] representa los pares Precision-Recall, codificados como ratios de Verdaderos Positivos (True Positives) y Falsos Positivos (False Positives), según vamos variando la configuración. Una curva más alta indicará mejores valores de precisión a un mismo valor de cobertura.

6.4.Regresión logística

La regresión logística [22] es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (variable que puede adoptar un número limitado de categorías) en función de variables independientes o predictoras.

6.5.Naive Bayes

Es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadores adicionales [23]. Es por ello, que se suele resumir en la hipótesis de independencia entre las variables predictoras por lo que recibe el apelativo de ingenuo (Naive). Una ventaja de este clasificador es que solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios para la clasificación. Otras variantes de Naive Bayes consideradas serían:

1) **Gaussian Naive Bayes**

Tiene la peculiaridad de que asume que los valores continuos asociados a cada clase a tratar siguen una distribución gaussiana [24].

2) Bernoulli Naive Bayes

Clasifica de acuerdo a distribuciones multivariables de Bernoulli. Lo haría como una Multibinomial Naive Bayes pero para valores binarios/booleanos [25].

6.6.Estudio y resultados

Para este apartado se utilizaron los pacientes del dataset de personas mayores o iguales a 65 años descrito en la *Sección 3.2.2: Pacientes*. De este dataset compuesto por 105930 pacientes que hubieran ido a urgencias había 52974 pacientes y por tanto eran de la clase positiva (habían ido alguna vez a urgencias dentro del periodo de estudio) y 69292 pacientes eran de la clase negativa (pacientes que habían ido alguna vez a urgencias pero nunca en la fecha de visita a urgencias). En total, este dataset estaba formado por 122266 pacientes. Esta información junto con las variables utilizadas para esta parte están definidas en la *Sección 4.1: Variables a nivel de paciente*.

Se procedió a la implementación de diferentes algoritmos para intentar predecir si un día en concreto una persona iba a visitar urgencias o no. A priori, se perfilaba como un problema muy complejo debido a la gran cantidad de factores que afectan cuando una persona vista urgencias. Es por ello que aun sabiendo esto, y trabajando con las variables descritas en el apartado correspondiente se visualizaba un problema complicado de predecir.

Cabe destacar que la idea inicial era que, una vez realizada una primera clasificación, se le añadiera la temperatura de un día en concreto y a partir de ahí y con la predicción de la temperatura, poder decidir si un paciente iría a urgencias o no en un día concreto.

6.6.1. Regresión Logística

En primer lugar, se llevó a cabo la implementación de una regresión logística simple. Para ello era necesario escoger las variables que mejor resultado pueden ofrecer a la hora de predecir. Por tanto, de las variables mencionadas se creó una lista con todas las posibles combinaciones para evaluar cada resultado.

De todas las ejecuciones se van a mostrar los resultados de las variables que han arrojado el mejor accuracy.

-Accuracy: 0.6098691

-Variables: 'DEPENDENCIA', 'TRATAMIENTO_HOSPITAL_DIA', 'NUMERO_VISITAS_URGENCIAS', 'NUMERO_VISITAS_URGENCIAS_EXTRA', 'NUMERO_INGRESOS_URGENTES'

		Valor Predicho	
		No va a Urgencias	Va a Urgencias
Valor Real	No va a Urgencias	18803	2094
	Va a Urgencias	12312	3471

Tabla 6.6.a: Matriz de confusión para regresión logística

	Precisión	Recall	F1-score
0	0.60	0.90	0.72
1	0.62	0.22	0.33
Avg/total	0.61	0.61	0.55

Tabla 6.6.b: Valores de Precision, Recall y F1-score para regresión logística

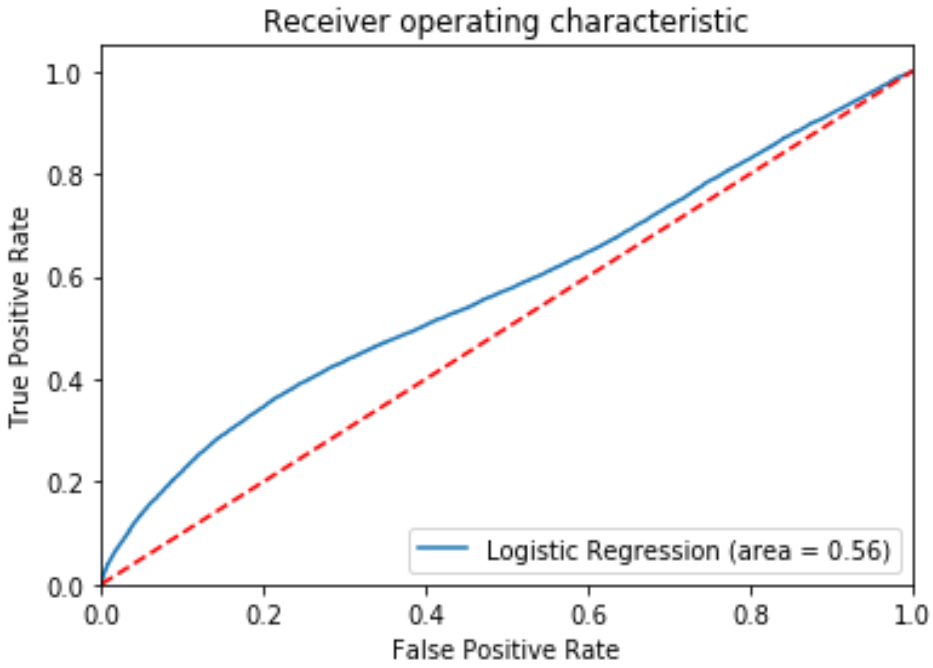


Figura 6.6.a: Área bajo la curva ROC para regresión logística

6.6.2. Gaussian Naive Bayes

En segundo lugar, se probó una gaussiana Naive Bayes.

De todas las ejecuciones se van a mostrar los resultados de las variables que han arrojado el mejor accuracy.

-Accuracy: 0.6138495

-Variables: 'SEXO', 'NIVEL_COPAGO_BAJO, 'NUMERO_VISITAS_URGENCIAS', 'NUMERO_MEDICAMENTOS_FARHO'

		Valor Predicho	
		No va a Urgencias	Va a Urgencias
Valor Real	No va a Urgencias	17852	3045
	Va a Urgencias	11691	4092

Tabla 6.6.c: Matriz de confusión para gaussian naive bayes

	Precisión	Recall	F1-score
0	0.60	0.85	0.71
1	0.57	0.26	0.36
Avg/total	0.59	0.60	0.56

Tabla 6.6.d: Valores de Precision, Recall y F1-score para gaussian naive bayes

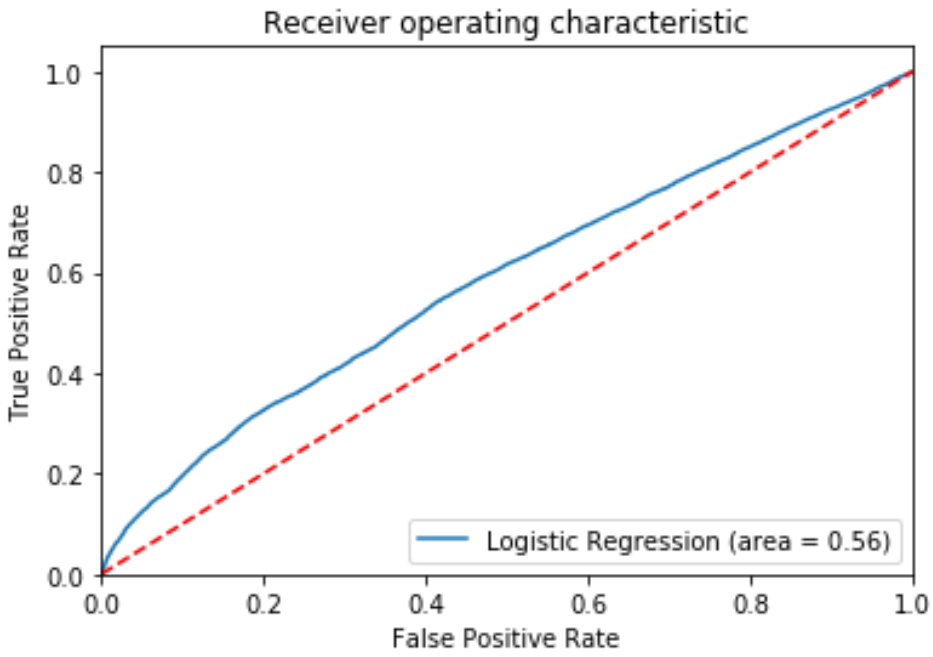


Figura 6.6.b: Área bajo la curva ROC para gaussian naive bayes

6.6.3. Bernoulli Naive Bayes

Por último, se implementó una bernoulli Naive Bayes.

De todas las ejecuciones se van a mostrar los resultados de las variables que han arrojado el mejor accuracy.

-Accuracy: 0.583451472192

-Variables: 'SEXO', 'DEPENDENCIA', 'NIVEL_COPAGO_BAJO, 'NUMERO_VISITAS_URGENCIAS', 'NUMERO_VISITAS_URGENCIAS_EXTRA', 'NUMERO_INGRESOS_URGENTES'

		Valor Predicho	
		No va a Urgencias	Va a Urgencias
Valor Real	No va a Urgencias	17087	3810
	Va a Urgencias	11643	4140

Tabla 6.6.e: Matriz de confusión para bernoulli naive bayes

	Precisión	Recall	F1-score
0	0.59	0.82	0.69
1	0.52	0.26	0.35
Avg/total	0.56	0.58	0.54

Tabla 6.6.f: Valores de Precision, Recall y F1-score para bernoulli naive bayes

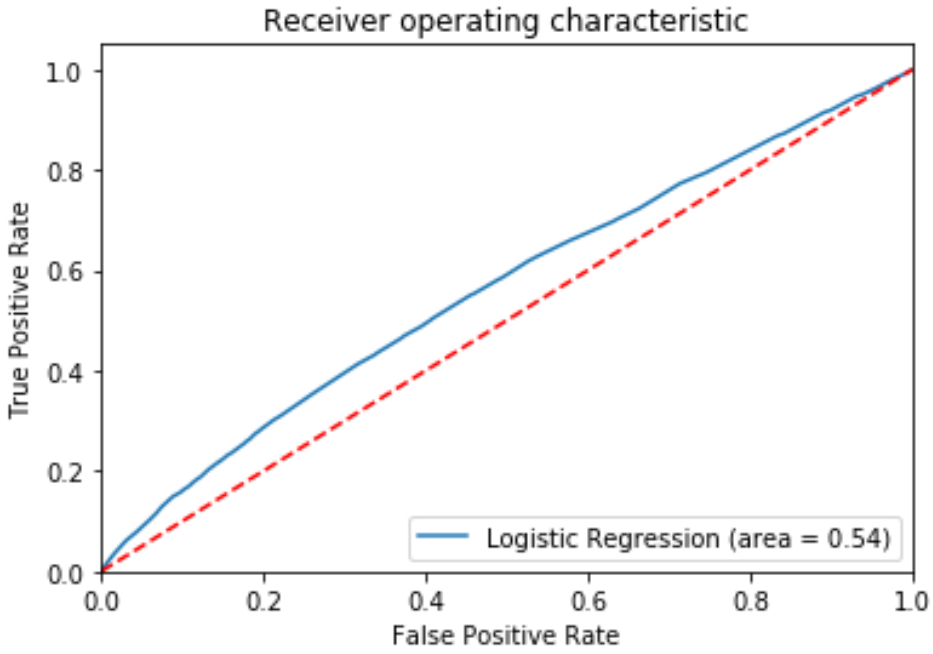


Figura 6.6.c: Área bajo la curva ROC para bernoulli naive bayes

-Conclusiones:

Se observa que la tasa de acierto ronda el 60%. No se ha podido encontrar ningún modelo que arroje una tasa de acierto mayor por lo que se concluye que realizar una predicción a nivel de paciente es un reto complicado. Debido a las causas que hacen que una persona vaya a urgencias o no, este proyecto no abarca todas las variables posibles que guardan relación. Además a nivel individual, las variables que se han obtenido de las BBDD de Salud ofrecen información pero no la suficiente como para obtener un modelo fiable y eficaz

7. Predicción a partir de series temporales

Vistos los resultados del apartado anterior y con el fin de obtener más conocimiento, en este apartado se presentará el trabajo realizado para predecir el número de visitas a urgencias a partir de las variables creadas a nivel de día. En primer lugar, se explicarán los conceptos de Serie temporal (*Sección 7.1*), ARIMA (*Sección 7.2*) y Red Neuronal (*Sección 7.3*), especialmente LSTM (*Sección 7.3.1*). Además se introducirá el término RMSE (*Sección 7.4)* que se ha utilizado en este apartado. Tras esto se presentarán los resultados obtenidos (*Sección 7.5*) tanto de SARIMAX (*Sección 7.5.1*) y LSTM (*Sección 7.5.2*) y por último una comparación entre resultados *(Sección 7.5.3*).

Las visitas a urgencias son un conjunto de valores que pueden ser tratados como una serie temporal y observar el comportamiento que siguen con el fin de poder obtener una predicción.

7.1.Serie temporal

Una serie temporal [13] es una secuencia de datos, observaciones o valores, medidos en momentos concretos y ordenados de forma cronológica. Estos datos pueden estar espaciados a intervalos iguales (por ejemplo cada hora) o desiguales (por ejemplo el peso de una persona cada vez que acude al médico). Uno de sus usos más habituales es su análisis para la predicción y pronóstico y por tanto multitud de ciencias consideran sus datos como series temporales. Un ejemplo sería la *Figura 7.1.a:*

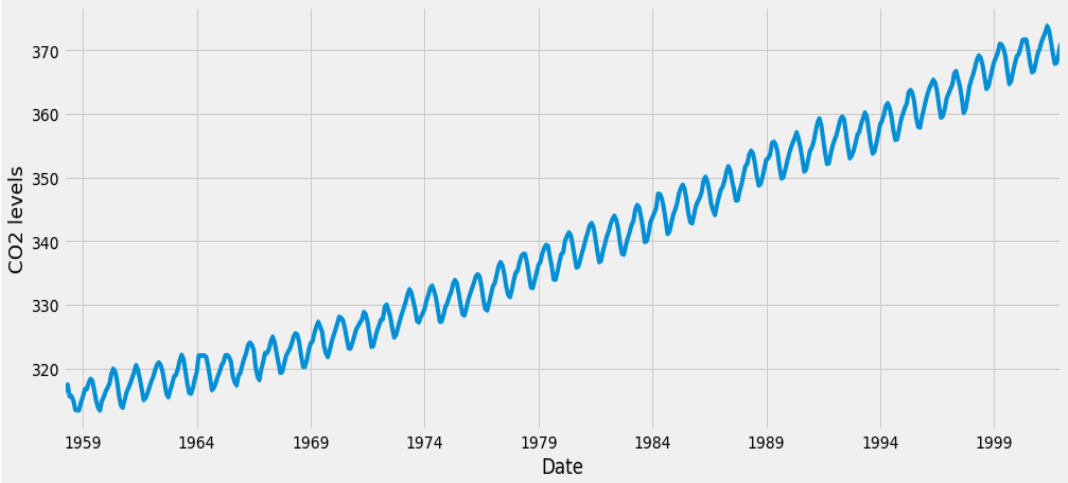


Figura 7.1.a: ejemplo de una serie temporal que muestra los niveles de CO2 en la isla de Hawaii desde 1958 hasta el año 2001. (Fuente: ARIMA Time Series Data Forecasting and Visualization in Python. En Digital Ocean [14])

7.2.ARIMA

Para tratar de predecir las urgencias se optó por utilizar el modelo autorregresivo integrado de promedio móvil (ARIMA por sus siglas en inglés) el cual [15] es un modelo estadístico utilizado en series temporales que utiliza variaciones y regresiones de datos estadísticos para encontrar patrones con los que poder realizar una predicción hacia el futuro. Es un modelo dinámico de series temporales por lo que las estimaciones futuras vienen explicadas por datos del pasado y no por variables independientes.

Para modelar ARIMA es necesario identificar los coeficientes y el número de regresiones a utilizar. Se suele expresar como $ARIMA\ (p, d, q)$ en dónde "*p*", "*d*" y "*q*" son los parámetros (números enteros no negativos) que indican:

-p: Componente autorregresiva: Nos permite incorporar el efecto de los valores del pasado en nuestro modelo.

-d: Componente integrada: Incluye términos al modelo que incorporan la diferencia a aplicar a la serie temporal.

-q: Componente de media móvil: Permite establecer el error de nuestro modelo como una combinación lineal de los errores observados en el pasado.

Para generalizar aún más y considerar el efecto de estacionalidad se utiliza el modelo SARIMAX (Seasonal ARIMA) que es el que vamos a utilizar debido a que necesitamos tener en cuenta la estacionalidad para nuestra predicción. En este caso, se denota como $ARIMA\ (p, d, q)\ (P, D, Q)\ s$ en donde (P, D, Q) hacen referencia a los mismos parámetros descritos arriba ((p, d, q)) aunque aplicados al componente estacional de la serie temporal. El término "*s*" hace referencia a la periodicidad (seasonal order).

7.3.Red neuronal

Una red neuronal [16] es un modelo computacional basado en un gran conjunto de unidades neuronales simples (neuronas artificiales). Funcionan de forma similar a como lo harían los axones de las neuronas en los cerebros biológicos. Cada unidad neuronal está conectada con muchas otras y los enlaces entre ellas pueden incrementar o inhibir el estado de activación de las neuronas adyacentes. Como se puede observar en la *Figura 7.3.a* cada nodo circular representa una neurona artificial y cada flecha representa una conexión desde la salida de una neurona a la entrada de otra.

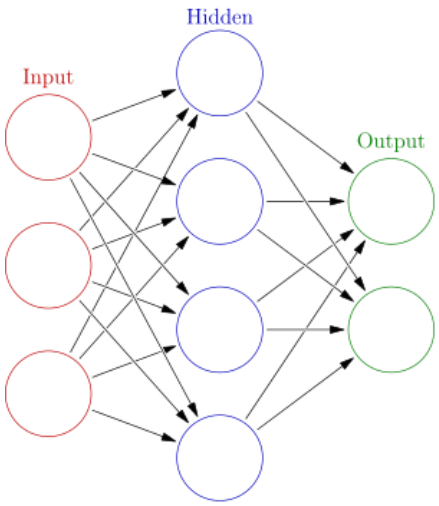


Figura 7.3.a: esquema de una red neuronal artificial. (Fuente: Red neuronal artificial. En Wikipedia [17])

7.3.1. LSTM

Debido a los resultados que arrojó ARIMA que se mostrarán más adelante se decidió implementar un modelo de redes neuronales. La memoria de largo a corto plazo (Long short-term memory, LSTM) [18] es una unidad de las redes neuronales recurrentes. Esta modalidad de red neuronal está compuesta de una celda, una puerta de entrada, una puerta de salida y una puerta de olvido. La celda es la responsable de ir recordando valores a partir de intervalos arbitrarios. Es por ello que recibe el nombre de "memoria". Cada una de las tres puertas funcionan como redes neuronales convencionales: con funciones de activación y pesos para cada neurona. La expresión de largo a corto plazo es utilizada debido a que LSTM es un modelo para la memoria a corto plazo que puede durar un gran periodo de tiempo. Son utilizadas para clasificar, procesar y predecir series temporales a partir del retardo entre sucesos (time lags).

Debido a que este modelo tiene la capacidad de almacenar ciertos valores y actualizarlos cada cierto tiempo, la cantidad de parámetros que se pueden establecer es más amplia que la que tendría una red neuronal sencilla. Estos serían el número de neuronas de la capa oculta, el número de capas ocultas, el número de ejemplos utilizados antes de actualizar los pesos de la red neuronal, número de veces que se entrena el modelo, número de lags, tipo de optimizador a utilizar al compilar la red neuronal, diferentes inicializaciones de los núcleos, etc.

Cabe recalcar que para predecir series temporales existen más métodos e implementaciones que podrían arrojar mejores resultados pero por cuestión de

tiempo no es viable probar muchos de ellos y este modelo ofrecía resultados interesantes.

7.4.RMSE

Para evaluar las predicciones una de las medidas más utilizadas es la raíz cuadrada del error cuadrático medio (Root-Mean-Square-Error en inglés [19]). Esta medida es usada para cuantificar las diferencias entre valores predichos y valores reales. Es una medida de precisión y se calcula como la raíz cuadrada de la media de los errores al cuadrado. En forma de ecuación sería:

RMSE=$\sqrt{\frac{\sum_{t=1}^{n}(\hat{a}_t - a_t)^2}{n}}$ en dónde $\hat{a}_t$ son los valores predichos, a_t los valores reales y *n* el número de predicciones realizadas.

7.5.Estudio y resultados

7.5.1. SARIMAX

Para este apartado se utilizó el dataset de >=65 años utilizados en la *Sección 5: Estudio estadístico*. Para la implementación de seasonal ARIMA (SARIMAX) es necesario obtener unos valores de los parámetros descritos anteriormente. Para ellos se fueron probando varios de ellos para observar cuál arrojaba el mejor AIC. El AIC (Akaike Information Criteria) es una medida de un modelo estadístico. Cuantifica la calidad y la simplicidad del modelo. Cuando se comparan dos modelos, el que menor AIC arroja, generalmente, funciona mejor. Por tanto, se probaron diferentes valores y algunos resultados se muestran en la *Tabla 7.4.a*

ARIMA(p,d,q)xseasonal_order	AIC
ARIMA(0, 0, 0)x(0, 0, 1, 12)	AIC:14625.7929249
ARIMA(0, 0, 0)x(0, 0, 2, 12)	AIC:13715.3650607
ARIMA(0, 0, 0)x(0, 0, 3, 12)	AIC:13143.482928
ARIMA(0, 0, 0)x(0, 0, 4, 12)	AIC:12659.8063332
ARIMA(0, 0, 0)x(0, 1, 2, 12)	AIC:10821.4480522
ARIMA(0, 0, 0)x(0, 1, 3, 12)	AIC:10716.8437523
ARIMA(0, 0, 0)x(0, 1, 4, 12)	AIC:10618.3538104
ARIMA(0, 0, 0)x(0, 1, 5, 12)	AIC:10531.8620835
ARIMA(0, 0, 0)x(0, 1, 6, 12)	AIC:10434.2611113
ARIMA(0, 0, 0)x(0, 2, 7, 12)	AIC:10622.674418
ARIMA(0, 0, 0)x(0, 2, 8, 12)	AIC:10633.1357769
ARIMA(0, 0, 0)x(0, 3, 1, 12)	AIC:12775.2414123

ARIMA(0, 0, 0)x(0, 3, 2, 12)	AIC:11564.1138416
ARIMA(0, 0, 0)x(0, 8, 2, 12)	AIC:14919.1168164
ARIMA(0, 0, 0)x(1, 0, 0, 12)	AIC:12027.9106662
ARIMA(0, 0, 0)x(1, 0, 1, 12)	AIC:10966.8595262
ARIMA(0, 0, 0)x(1, 0, 2, 12)	AIC:10857.8562739
ARIMA(0, 0, 0)x(1, 1, 4, 12)	AIC:10624.6465121
ARIMA(0, 0, 0)x(1, 1, 5, 12)	AIC:10503.7463204
ARIMA(0, 0, 0)x(1, 1, 6, 12)	AIC:10416.3298814
ARIMA(0, 0, 0)x(1, 1, 7, 12)	AIC:10305.613751
ARIMA(0, 0, 0)x(2, 0, 6, 12)	AIC:10474.3481808
ARIMA(0, 0, 0)x(2, 0, 7, 12)	AIC:10379.1719806
ARIMA(0, 0, 0)x(2, 0, 8, 12)	AIC:10237.5798269
ARIMA(0, 0, 0)x(2, 1, 0, 12)	AIC:11136.7637927
ARIMA(0, 0, 0)x(2, 3, 7, 12)	AIC:10878.687948
ARIMA(0, 0, 0)x(2, 4, 0, 12)	AIC:12755.1314537
ARIMA(0, 0, 0)x(2, 4, 1, 12)	AIC:12091.4399927
ARIMA(0, 0, 0)x(2, 4, 2, 12)	AIC:12069.2247531
ARIMA(0, 0, 0)x(2, 4, 3, 12)	AIC:11523.7337065
ARIMA(0, 0, 0)x(5, 2, 8, 12)	AIC:10043.7801556
ARIMA(0, 0, 0)x(5, 3, 0, 12)	AIC:11816.1889601
ARIMA(0, 0, 0)x(5, 3, 1, 12)	AIC:11171.1948539
ARIMA(0, 0, 0)x(5, 3, 2, 12)	AIC:10889.3395703
ARIMA(0, 0, 0)x(6, 3, 2, 12)	AIC:11787.438077
ARIMA(0, 0, 0)x(6, 3, 3, 12)	AIC:10517.2664924
ARIMA(0, 0, 0)x(6, 3, 4, 12)	AIC:10556.4477585
ARIMA(0, 0, 0)x(6, 3, 5, 12)	AIC:10570.1003357
ARIMA(0, 0, 0)x(7, 2, 6, 12)	AIC:10127.8915188
ARIMA(0, 0, 0)x(7, 2, 7, 12)	AIC:10101.5847208
ARIMA(0, 0, 0)x(7, 2, 8, 12)	AIC:9963.05008722
ARIMA(0, 0, 0)x(7, 3, 0, 12)	AIC:11108.5175654
ARIMA(0, 0, 0)x(7, 3, 1, 12)	AIC:10658.2091917
ARIMA(0, 0, 0)x(7, 3, 2, 12)	AIC:10437.3496645
ARIMA(0, 0, 0)x(8, 1, 3, 12)	AIC:10102.7510049
ARIMA(0, 0, 0)x(8, 1, 4, 12)	AIC:10092.3001835
ARIMA(0, 0, 0)x(8, 1, 5, 12)	AIC:10066.4326326
ARIMA(0, 0, 0)x(8, 1, 6, 12)	AIC:10064.1466068
ARIMA(0, 0, 0)x(8, 1, 7, 12)	AIC:10006.3042112

Tabla 7.4.a: Parámetros utilizados y AIC obtenido de algunas ejecuciones

Por tanto el menor AIC obtenido es para la combinación ARIMA(0,0,0)x(7,2,8,12). Con esta combinación obtenemos la *Tabla 7.4.b*

	coef	Std err	z	P>|z|	[0.025	0.975]
ar.S.L12	-1.3612	0.084	-16.274	0.000	-1.525	-1.197
ar.S.L24	-1.5667	0.163	-9.609	0.000	-1.886	-1.247
ar.S.L36	-1.4909	0.205	-7.271	0.000	-1.893	-1.089
ar.S.L48	-1.1807	0.191	-6.192	0.000	-1.554	-0.807
ar.S.L60	-0.8534	0.135	-6.320	0.000	-1.118	-0.589
ar.S.L72	-0.7740	0.070	-11.123	0.000	-0.910	-0.638
ar.S.L84	-0.0143	0.015	-0.985	0.325	-0.043	0.014
ma.S.L12	-0.5457	0.101	-5.382	0.000	-0.744	-0.347
ma.S.L24	-0.1466	0.076	-1.933	0.053	-0.295	0.002
ma.S.L36	-0.2858	0.074	-3.874	0.000	-0.430	-0.141
ma.S.L48	-0.3058	0.085	-3.605	0.000	-0.472	-0.140
ma.S.L60	0.0183	0.080	0.231	0.818	-0.138	0.174
ma.S.L72	0.3077	0.067	4.565	0.000	0.176	0.440
ma.S.L84	-0.5955	0.089	-6.713	0.000	-0.769	-0.422
ma.S.L96	0.5594	0.076	7.398	0.000	0.411	0.708
sigma2	98.5475	5.208	18.921	0.000	88.339	108.756

Tabla 7.4.b: Resumen obtenido de la ejecución de SARIMAX

Aquí podemos observar un resumen de los datos en dónde centraremos nuestra atención en la columna *coef.* Esta nos muestra la importancia de cada característica y cómo afecta a la serie temporal. Además la columna *P>|z|* muestra el nivel de significancia del peso de cada variable. Si la variable tiene un p-valor menor o igual que 0.05, es razonable mantenerla en el modelo.

Cuando implementamos modelos como SARIMAX es importante realizar comprobaciones para asegurarnos de que ninguna de las suposiciones iniciales son violadas. Estas serían garantizar que los residuos de nuestro modelo no están correlacionados y que siguen una distribución normal de media cero. Si el modelo no satisface estas propiedades, es un indicador de que puede ser mejorado.

En este caso, podemos observar en la *Figura 7.4.a* en la parte superior derecha como la distribución KDE sigue casi con total exactitud una distribución normal de media cero. Esto es un buen indicador de que los residuos están distribuidos de forma normal. En la parte inferior izquierda se puede observar como los residuos (puntos azules) siguen la línea de tendencia de una distribución normal N(0,1) aunque en la última parte se alejan un poco. Esto es otro buen indicador. Por último, en la parte superior izquierda son mostrados los residuos en el tiempo y no se observa ninguna estacionalidad de manera obvia. En este caso parece que se observa ruido blanco. Todo esto es, en parte, confirmado en la parte inferior derecha que muestra que los residuos de la serie temporal tienen baja correlación con versiones atrasadas de la misma serie temporal, aunque es cierto que para 4 y 7 días hay algo de la misma.

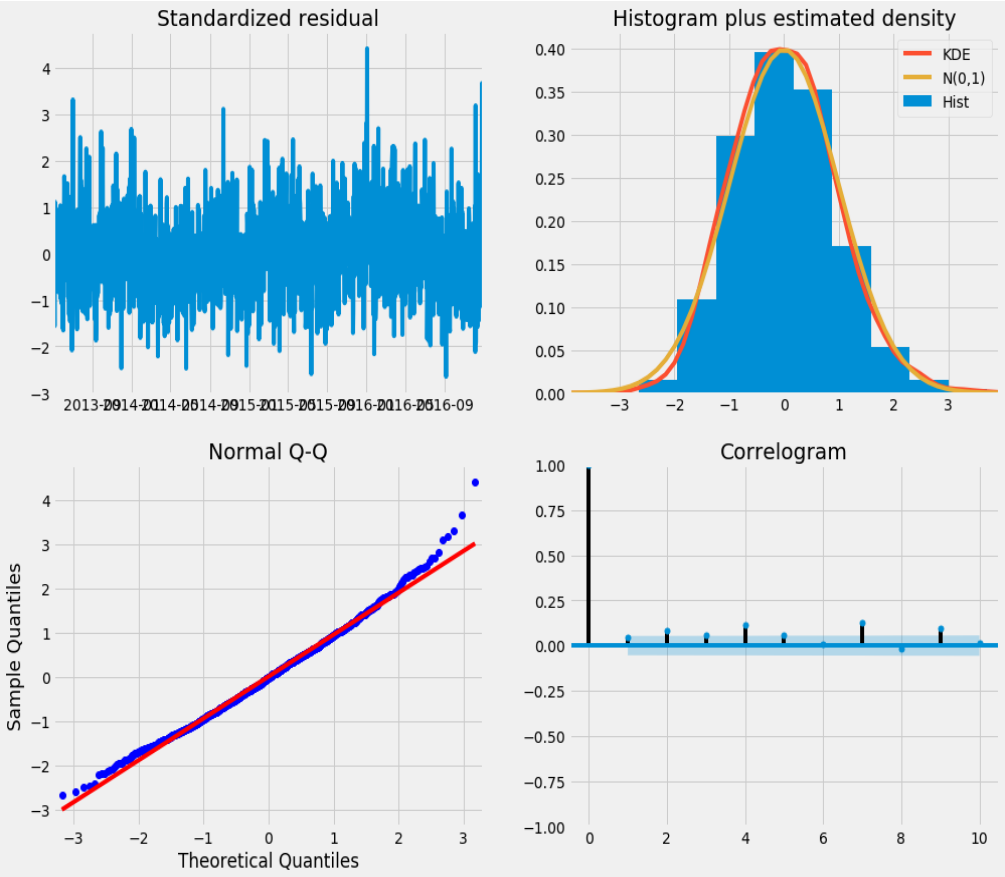


Figura 7.4.a: gráficos para observar el comportamiento de los datos para la configuración descrita anteriormente

Por tanto, esto nos hace concluir que nuestro modelo se ajusta bastante bien y podría ser usado para predecir valores futuros, a priori.

Para la predicción comprobaremos como funciona prediciendo el año 2016 y viendo el error cometido para ver cómo funciona el modelo. En la *Figura 7.4.b* vemos el resultado en marrón de la predicción para el año en cuestión.

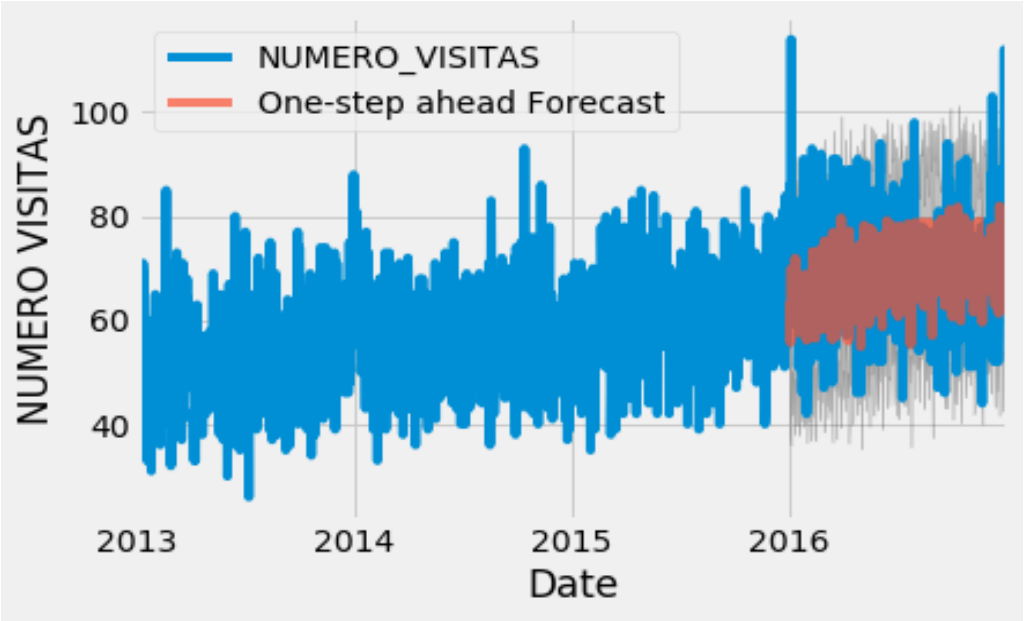


Figura 7.4.b: Representación temporal de las visitas a urgencias junto con la predicción obtenida por este modelo

A simple vista no se ve un mal resultado. Para comprobarlo de manera más objetiva usaremos la raíz cuadrada del error cuadrático medio (RMSE en inglés) para cuantificar el error obtenido. Para este caso el error ha sido de 10,48 lo cual es un número bastante bueno teniendo en cuenta que la media de visitas a urgencias para este dataset es de 71,76 visitas.

Los resultados son bastante optimistas y cabe recalcar que solo se tiene en cuenta la variable NUMERO_VISITAS para predecir y estos valores cambian bastante cada día y no siguen una tendencia clara.

Para visualizar mejor estos resultados se van a mostrar una serie de gráficos que ayudarán a entender la predicción obtenida.

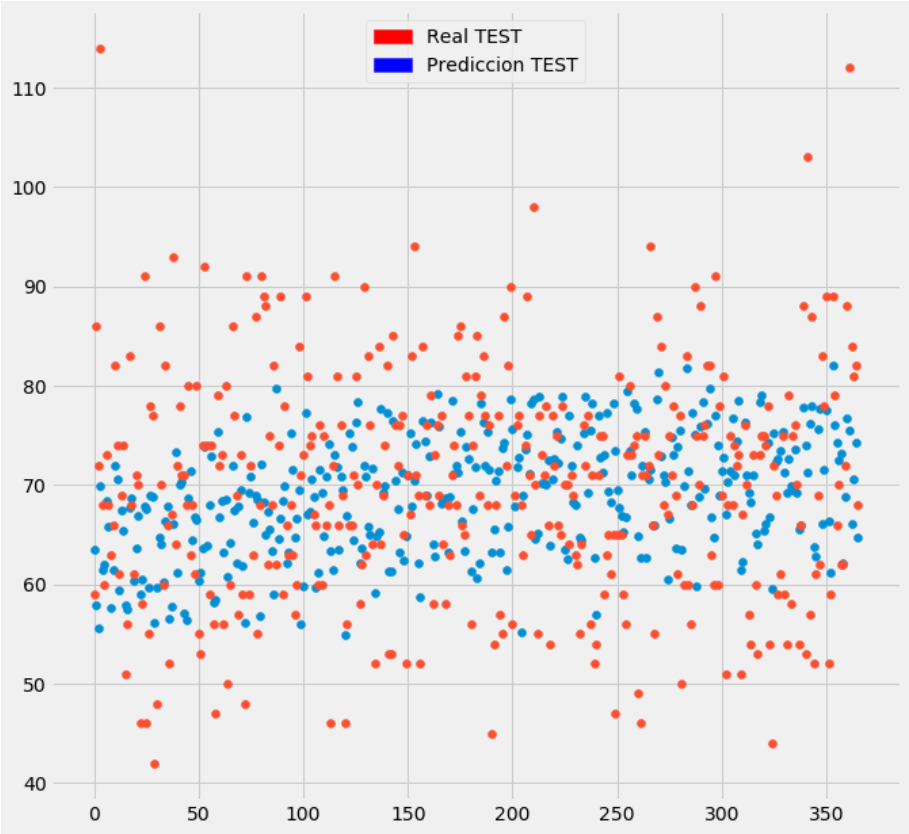


Figura 7.4.c: Representación gráfica en el tiempo de los valores reales y los valores predichos para la predicción obtenida (año 2016)

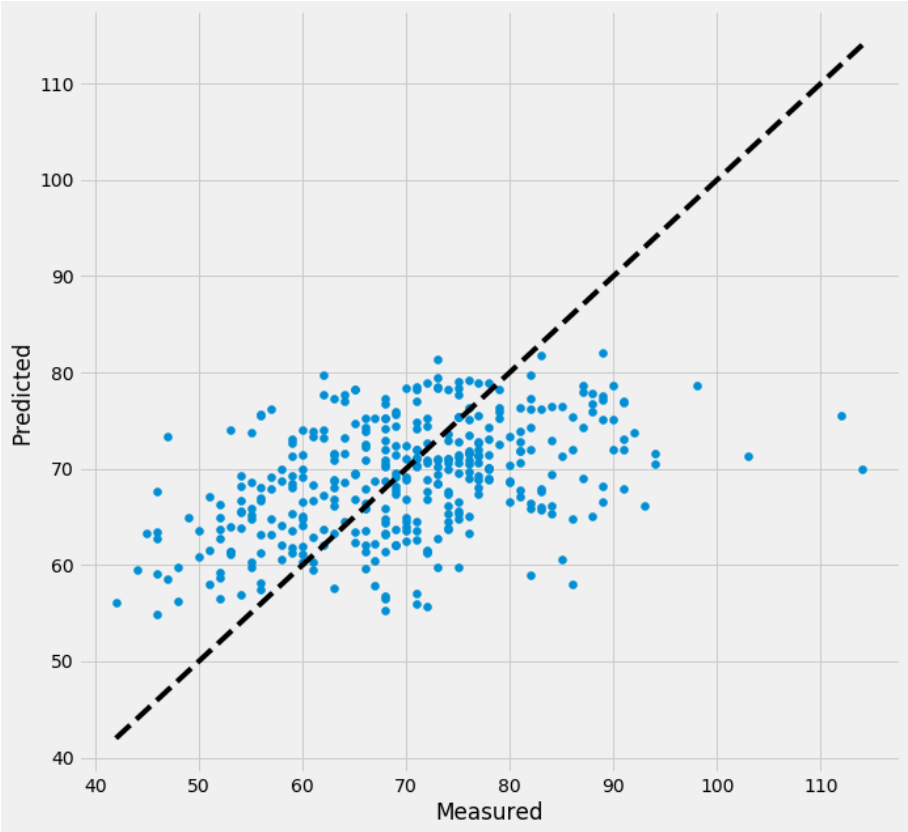


Figura 7.4.d: Representación gráfica de los valores predichos y los valores reales para la predicción obtenida del año 2016

La diagonal indica una predicción perfecta en dónde cada valor predicho coincide con cada valor real. Por tanto, podemos observar como la mayoría de los datos no se ajustan a la diagonal aunque están distribuidos de manera bastante homogénea.

count	366.00000
mean	-0.929995
std	10.453795
min	-44.070105
25%	-7.719751
50%	-0.763950
75%	6.936947
max	26.400518

Tabla 7.4.c: Resumen de la diferencia entre valores reales y valores predichos para la predicción obtenida

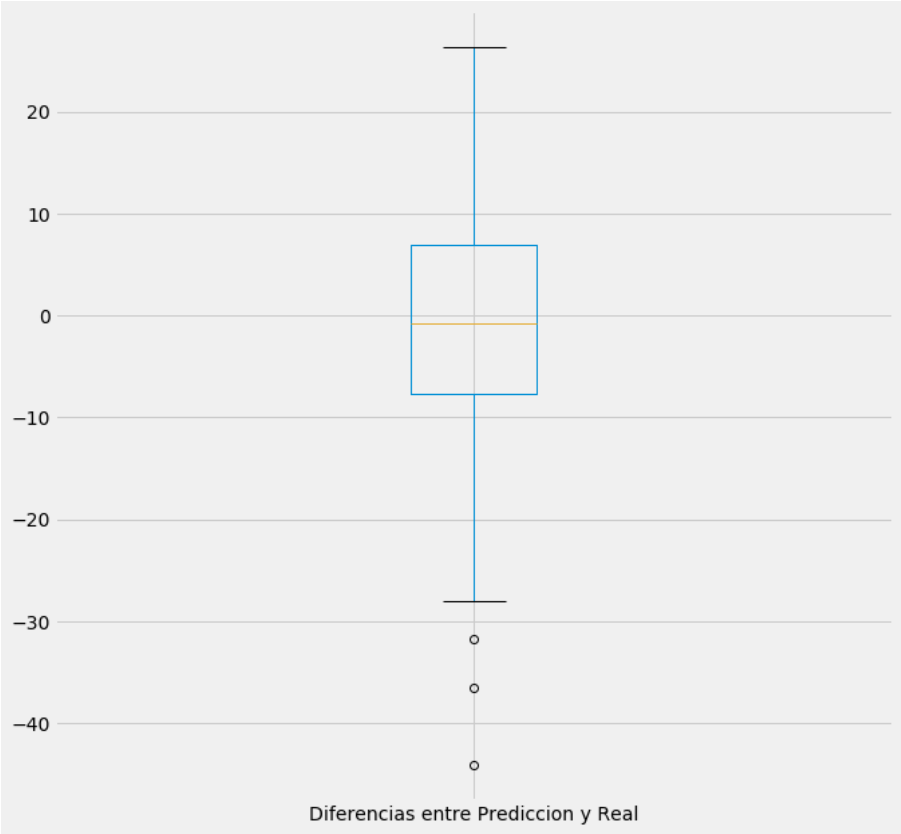


Figura 7.4.e: Representación de la diferencia entre la predicción y el valor real de la predicción obtenida

El borde superior de la caja azul representa el tercer cuartil, el borde inferior el primer cuartil, la línea naranja la mediana, los límites marcados con rayas negras engloban la mayoría de los datos y los puntos marcan los outliers.

En este caso se observa como la mayoría de los datos están entre -7,72 y 6,93 visitas fallidas a la hora de predecir. Es un buen resultado teniendo en cuenta la variabilidad de los datos y el uso de una única variable para predecir.

Tras ver estos resultados se comparará con el modelo LSTM que va a ser explicado a continuación y se extraerán conclusiones.

7.5.2. LSTM

Un segundo modelo fue implementado para intentar mejorar los resultados obtenidos hasta ahora. Este modelo fue implementado ya que ofrecía la posibilidad de usar más variables para predecir así como diferentes parámetros que ayudarían a construir un modelo más preciso.

El modelo LSTM permite modificar multitud de parámetros como se ha visto anteriormente. En este apartado se van a mostrar los resultados de la mejor combinación de ellos que han arrojado mejor resultado en la predicción. Cabe destacar que se crearon dos modelos con diferentes parámetros para los dos datasets mencionados. El dataset obtenido en la *Sección 5: Estudio estadístico* y el dataset con todas las visitas a urgencias del Complejo Hospitalario de Navarra en pacientes mayores o iguales a 16 años.

-Dataset >=65:

-Parámetros:

1-Número de días anteriores utilizados: 4

2-Numero de neuronas de la capa oculta: 1

3-Numero de épocas: 50

4-Estacionalidad: True

5-Numero de ejemplos de entrenamiento: 1095 (3 años)

6-Número de ejemplos de test: 365 (1 año)

7-Variables: NUMERO_VISITAS, FESTIVO, DIA_SEMANA, GRIPE

-Resultados:

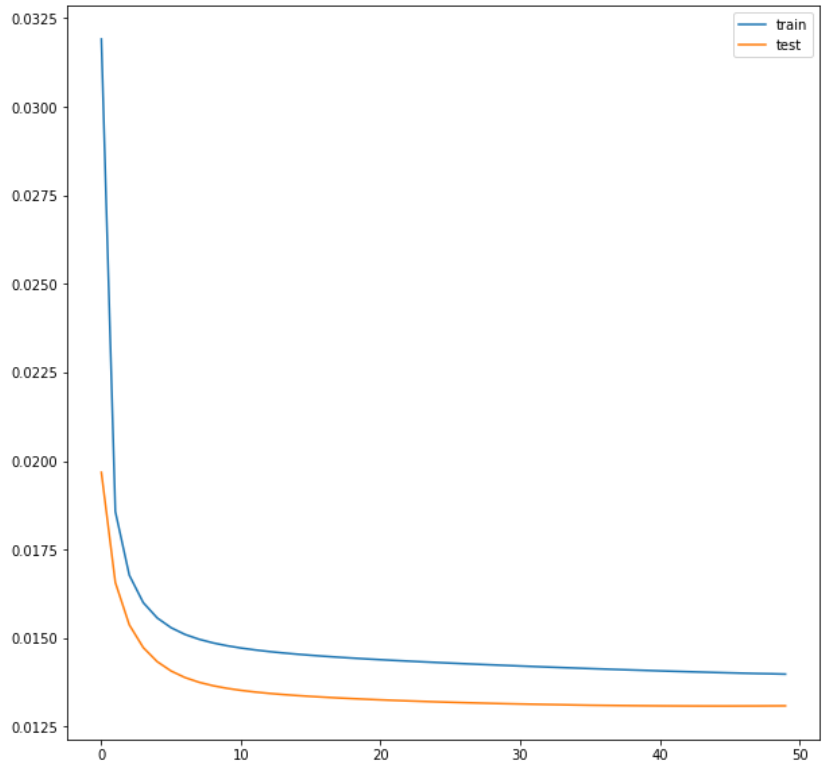


Figura 7.4.c: Convergencia para train y test con la mejor combinación de parámetros

Se observa como tanto train como test acaban arrojando un valor parecido. No hay sobreentrenamiento (bajos valores de error para train pero altos para test) ni se observa que el modelo pudiera mejorar con más entrenamiento del modelo (la tendencia es prácticamente lineal para ambos casos).

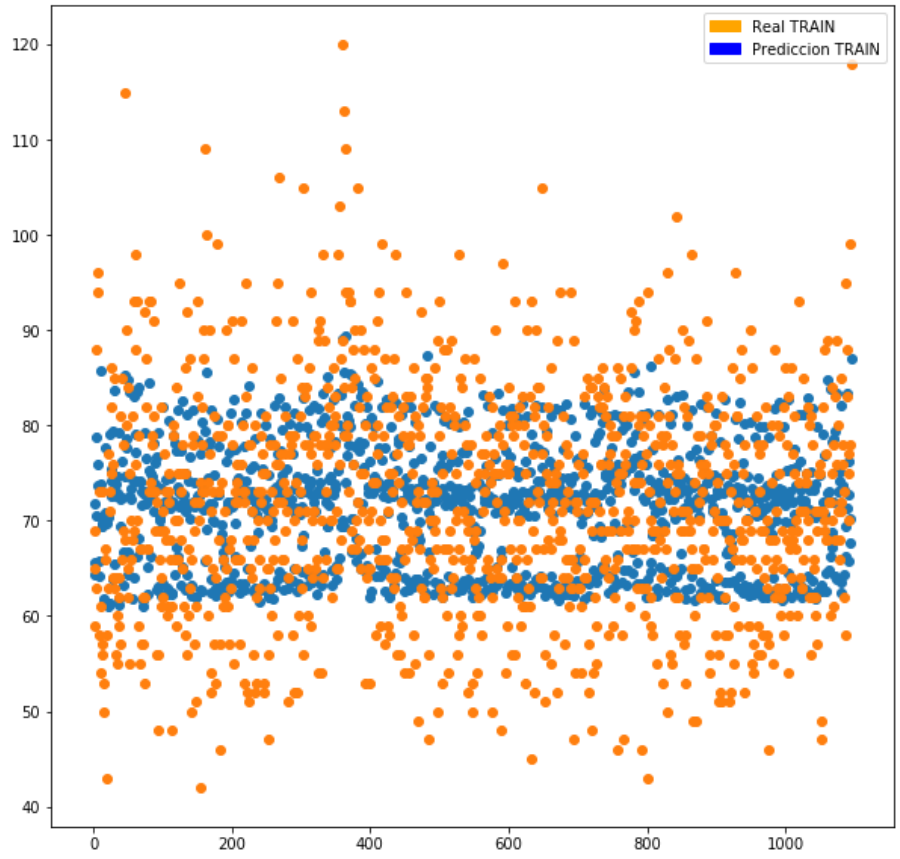


Figura 7.4.d: Representación gráfica en el tiempo de los valores reales y los valores predichos para el conjunto de entrenamiento (años 2013-2015)

Ya se puede observar a simple vista que los valores reales son muy dispersos y varían mucho de unos días a otros por lo que la predicción no puede ajustar muy bien los valores.

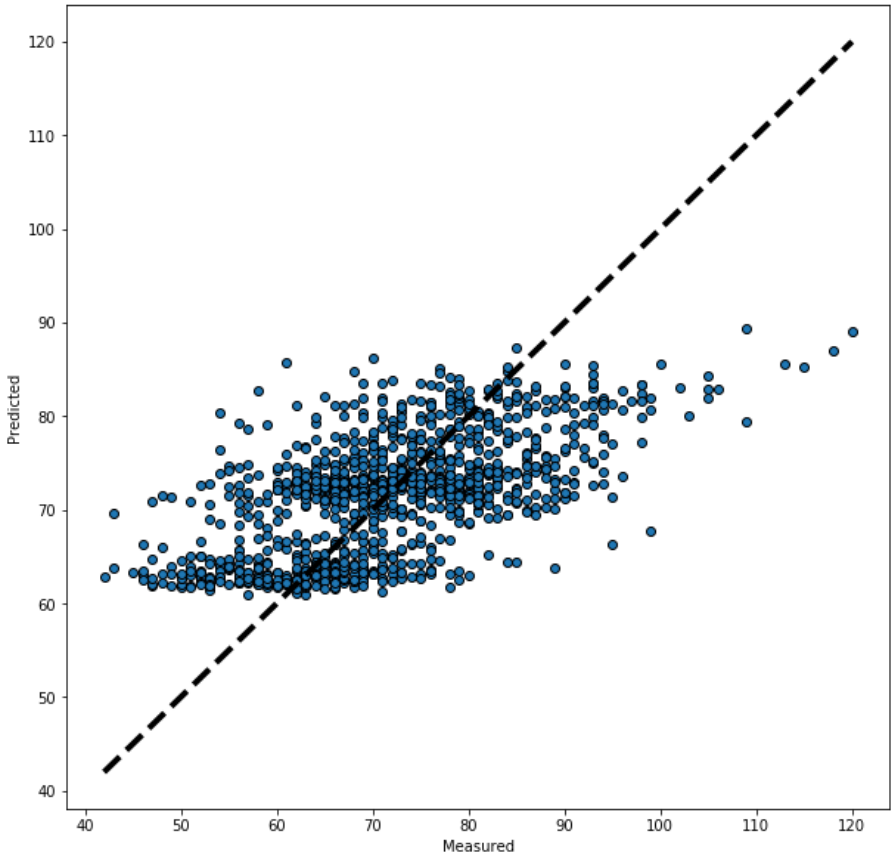


Figura 7.4.e: Representación gráfica de los valores predichos y los valores reales para el conjunto de entrenamiento

La diagonal indica una predicción perfecta en dónde cada valor predicho coincide con cada valor real. Por tanto, podemos observar como la mayoría de los datos no se ajustan a la diagonal pero se mantienen cerca de ella.

Para evaluar el error cometido por la predicción utilizaremos la raíz cuadrada del error cuadrático medio (RMSE en inglés). En este caso el RMSE ha sido de 9,346.

count	1095.000000
mean	0.455099
std	9.338836
min	-30.628983
25%	-5.638359
50%	0.900139
75%	6.839489
max	27.205101

Tabla 7.4.c: Resumen de la diferencia entre valores reales y valores predichos para el conjunto de entrenamiento

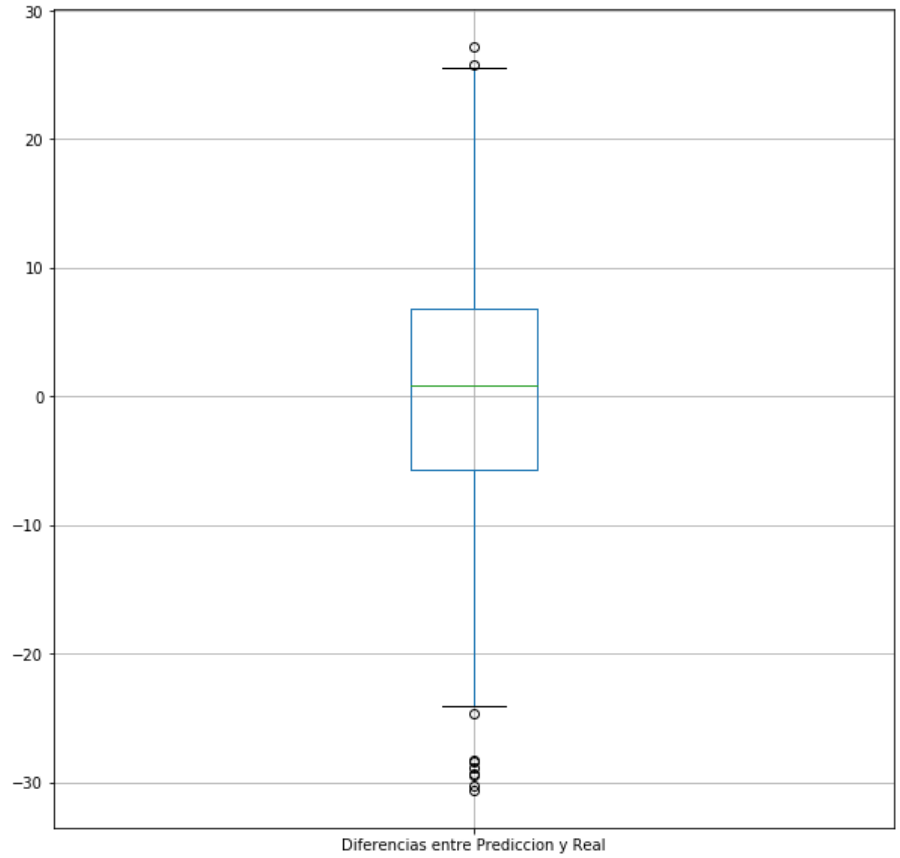


Figura 7.4.f: Representación de la diferencia entre la predicción y el valor real de los días de entrenamiento

El borde superior de la caja azul representa el tercer cuartil, el borde inferior el primer cuartil, la línea verde la mediana, los límites marcados con rayas negras engloban la mayoría de los datos y los puntos marcan los outliers.

En este caso se observa como la mayoría de los datos están entre -5,64 y 6,84 visitas fallidas a la hora de predecir. También hay bastantes valores que se predicen sin apenas error aunque es cierto que también hay algunos outliers con bastante error.

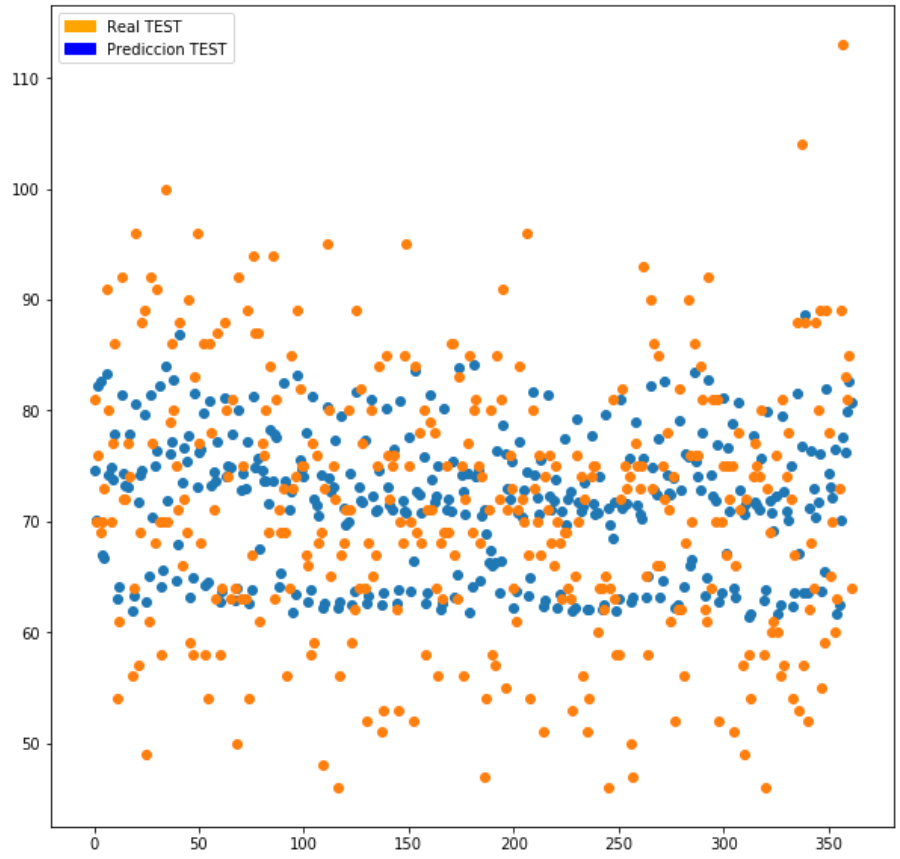


Figura 7.4.g: Representación gráfica en el tiempo de los valores reales y los valores predichos para el conjunto de test (años 2016)

Se observa un comportamiento parecido al conjunto de entrenamiento.

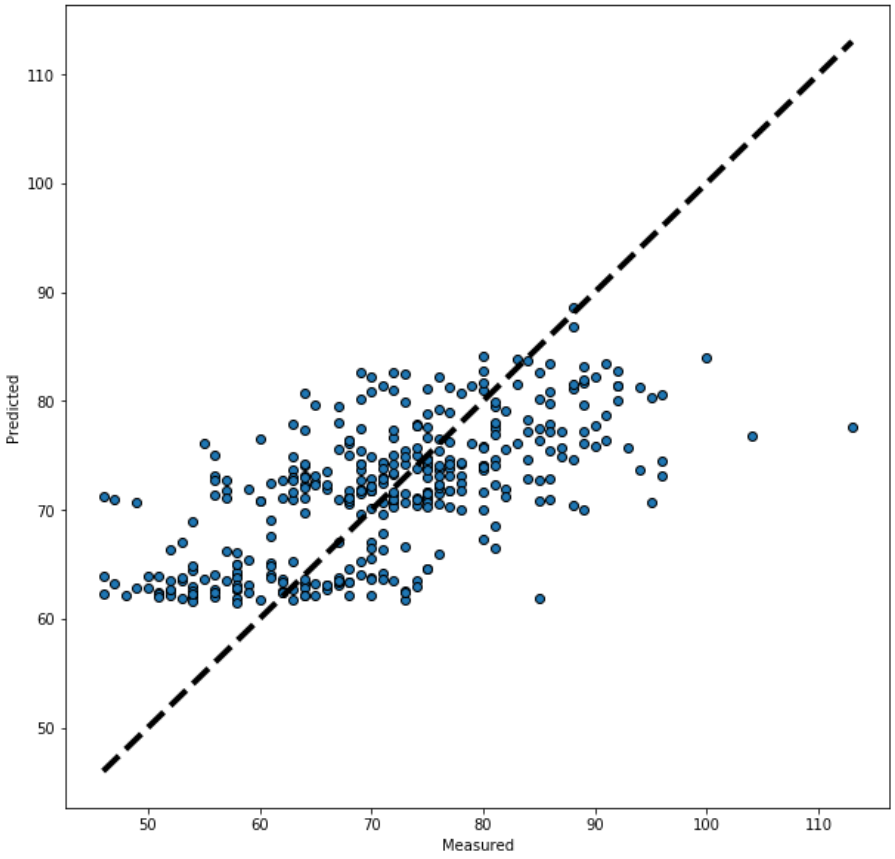


Figura 7.4.h: Representación gráfica de los valores predichos y los valores reales para el conjunto de test

La diagonal indica una predicción perfecta en dónde cada valor predicho coincide con cada valor real. Para el conjunto de test, los valores se distribuyen de una forma similar que para train.

En este caso la raíz cuadrada del error cuadrático medio (RMSE) ha sido de 8,881.

count	360.000000
mean	0.943174
std	8.842926
min	-32.676407
25%	-4.421303
50%	1.037189
75%	7.300253
max	26.189926

Tabla 7.4.d: Resumen de la diferencia entre valores reales y valores predichos para el conjunto de test

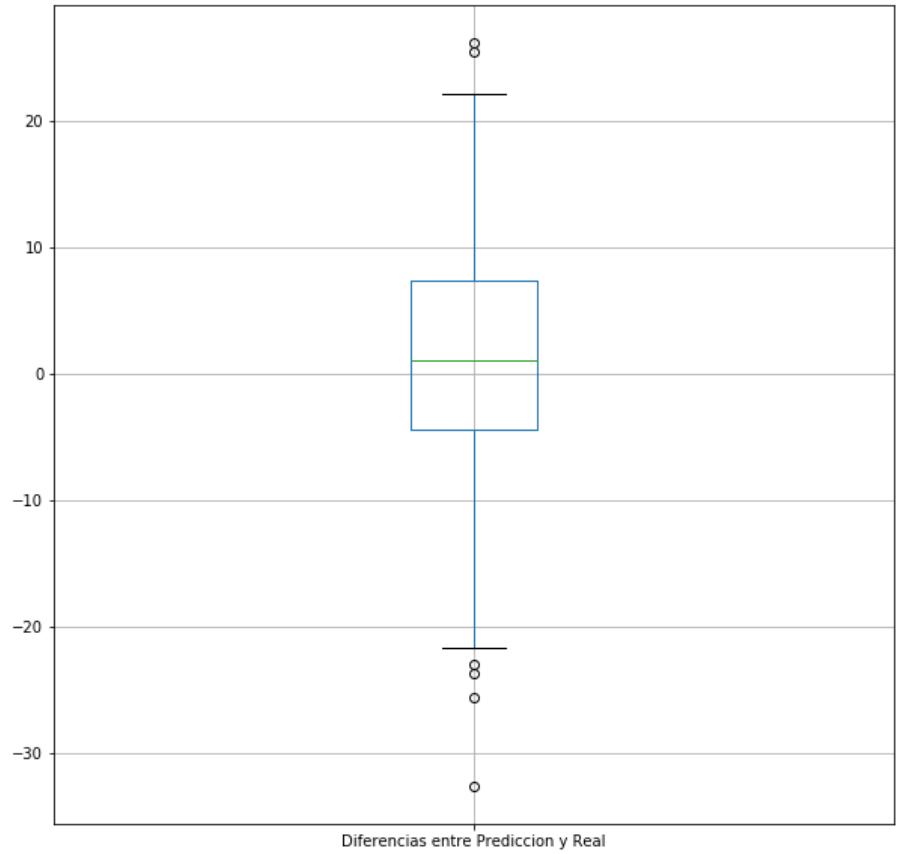


Figura 7.4.i: Representación de la diferencia entre la predicción y el valor real de los días de test

En este caso se observa como la mayoría de los datos están entre -4,42 y 7,30 visitas fallidas a la hora de predecir. La distribución para el conjunto de test es muy similar al conjunto de entrenamiento.

-Dataset todas las urgencias del Complejo Hospitalario de Navarra:

-Parámetros:

1-Número de días anteriores utilizados: 6

2-Numero de neuronas de la capa oculta: 2

3-Numero de épocas: 80

4-Estacionalidad: True

5-Numero de ejemplos de entrenamiento: 1095 (3 años)

6-Número de ejemplos de test: 365 (1 año)

7-Variables: NUMERO_VISITAS, FESTIVO, DIA_SEMANA, GRIPE

-Resultados:

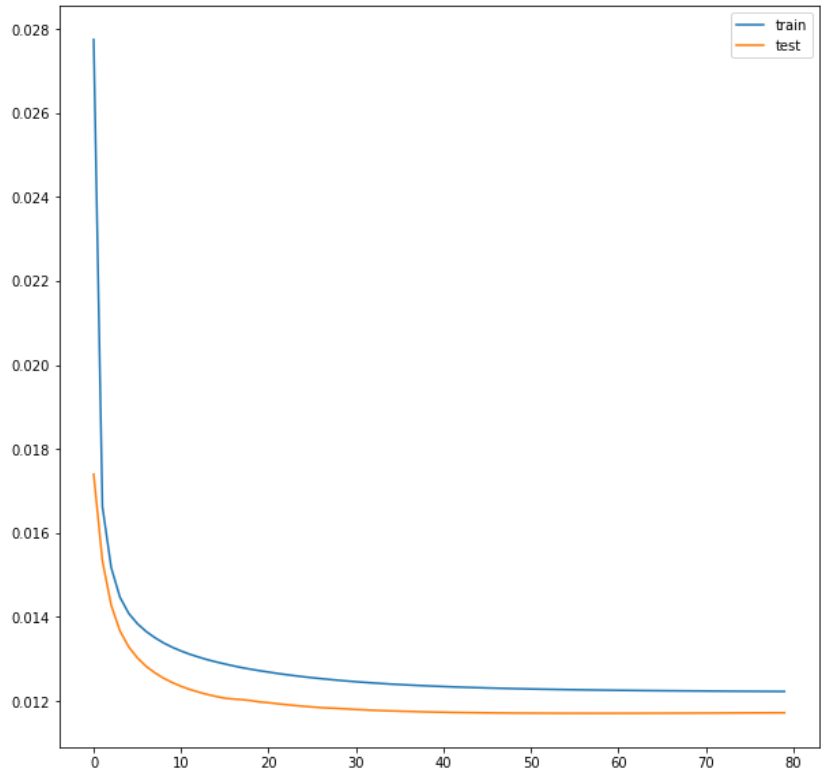


Figura 7.4.j: Convergencia para train y test con la mejor combinación de parámetros

Figura 7.4.k: Representación gráfica en el tiempo de los valores reales y los valores predichos para el conjunto de entrenamiento (años 2013-2015)

Se observa como tanto train como test acaban arrojando un valor parecido. No hay sobreentrenamiento ni se observa que el modelo pudiera mejorar con más entrenamiento del modelo (la tendencia es prácticamente lineal para ambos casos).

Se puede observar como para este dataset el comportamiento es algo similar al dataset anterior aunque en este caso parece que la predicción se comporta algo mejor.

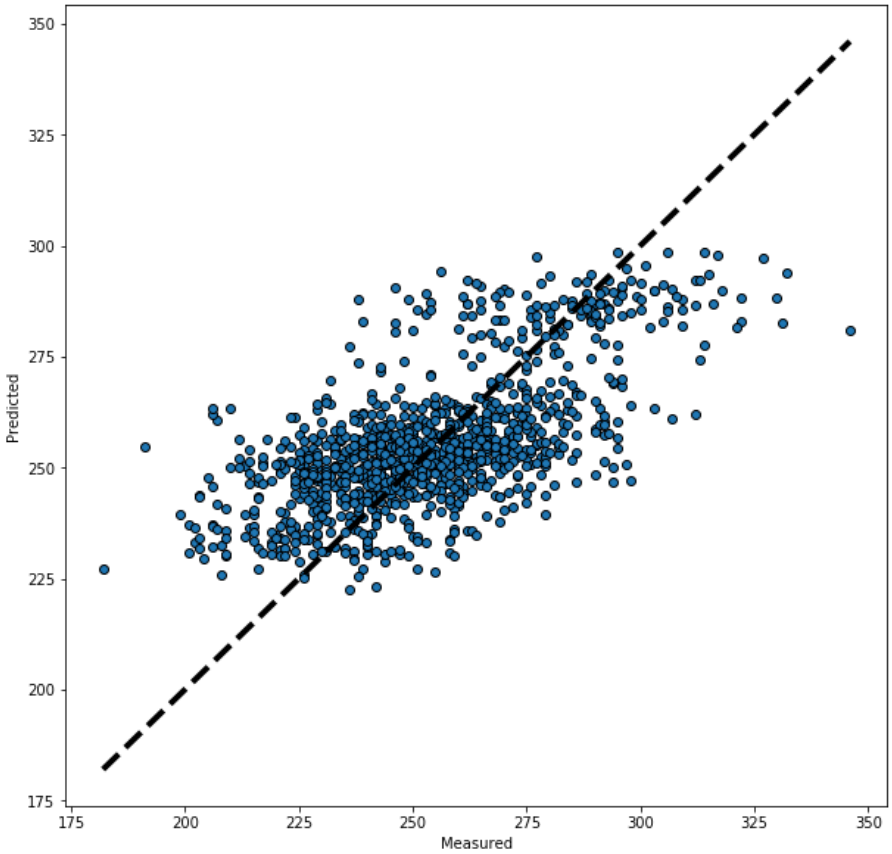


Figura 7.4.l: Representación gráfica de los valores predichos y los valores reales para el conjunto de entrenamiento

La diagonal indica una predicción perfecta en dónde cada valor predicho coincide con cada valor real. Observamos una distribución parecida al dataset anterior aunque da la sensación que los puntos están más próximos a la diagonal lo que daría un mejor resultado a priori.

Para evaluar el error cometido por la predicción utilizaremos la raíz cuadrada del error cuadrático medio (RMSE en inglés). En este caso el RMSE ha sido de 18,059

count	1095.000000
mean	1.633145
std	17.992762
min	-65.439697
25%	-10.383606
50%	1.488861
75%	13.551208
max	63.547104

Tabla 7.4.e: Resumen de la diferencia entre valores reales y valores predichos para el conjunto de entrenamiento

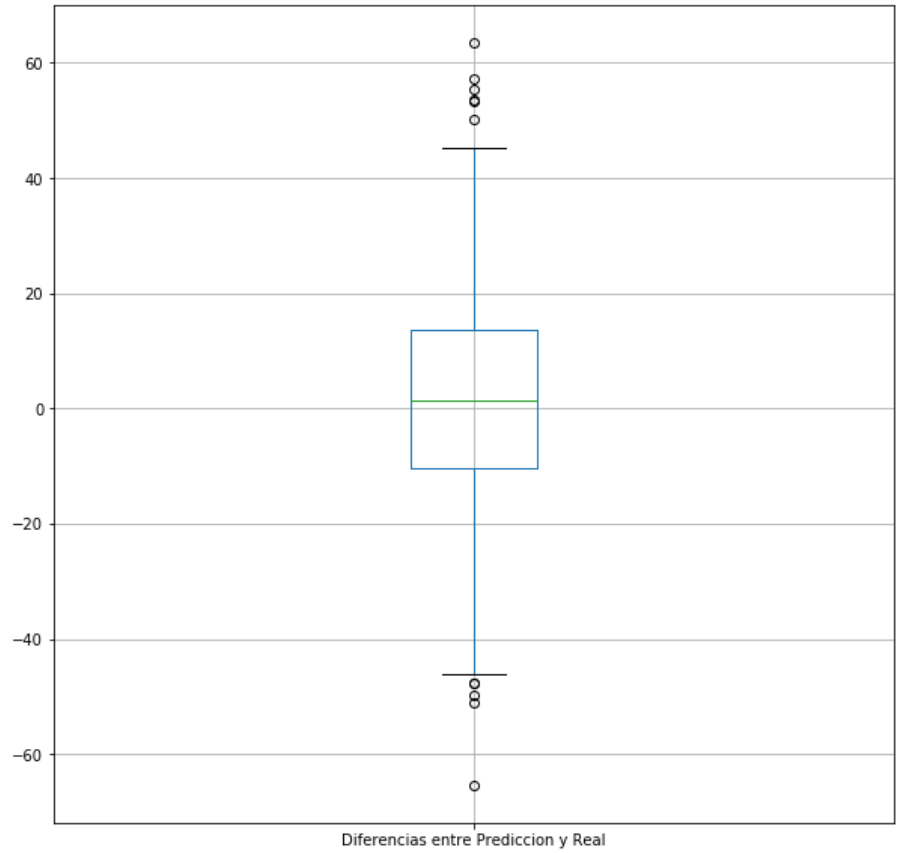


Figura 7.4.m: Representación de la diferencia entre la predicción y el valor real de los días de entrenamiento

El borde superior de la caja azul representa el tercer cuartil, el borde inferior el primer cuartil, la línea verde la mediana, los límites marcados con rayas negras engloban la mayoría de los datos y los puntos marcan los outliers.

En este caso se observa como la mayoría de los datos están entre -10,38 y 13,55 visitas fallidas a la hora de predecir.

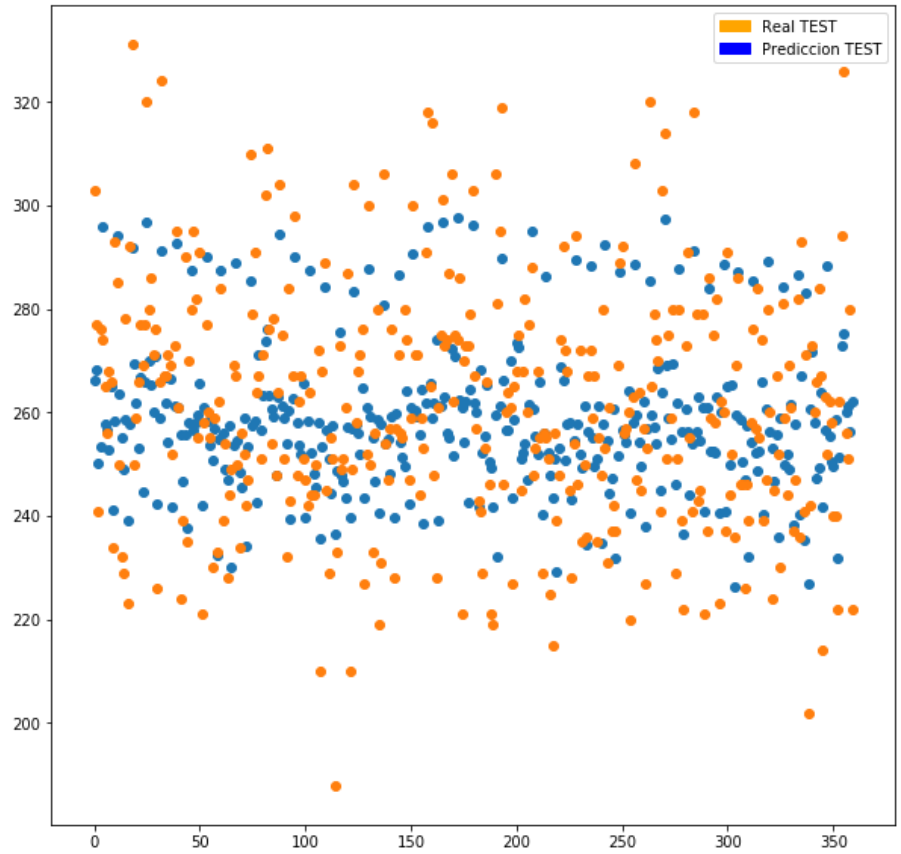


Figura 7.4.n: Representación gráfica en el tiempo de los valores reales y los valores predichos para el conjunto de test (años 2016)

El comportamiento es similar a los datos de entrenamiento.

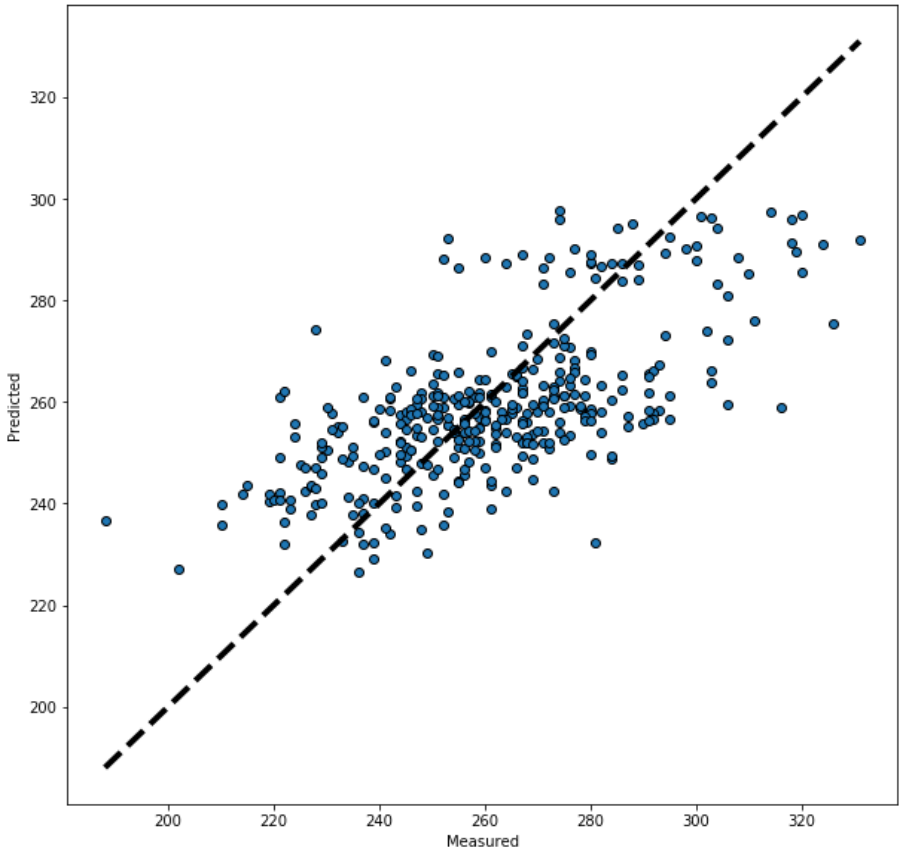


Figura 7.4.o: Representación gráfica de los valores predichos y los valores reales para el conjunto de test

La diagonal indica una predicción perfecta en dónde cada valor predicho coincide con cada valor real. En este caso los datos parecen estar más dispersos que para el conjunto de entrenamiento aunque parece que el resultado es mejor que para el dataset anterior.

En este caso la raíz cuadrada del error cuadrático medio (RMSE) ha sido de 17,822

count	360.000000
mean	-2.305096
std	17.697315
min	-56.974152
25%	-13.491772
50%	-2.072891
75%	10.038521
max	48.535751

Tabla 7.4.f: Resumen de la diferencia entre valores reales y valores predichos para el conjunto de test

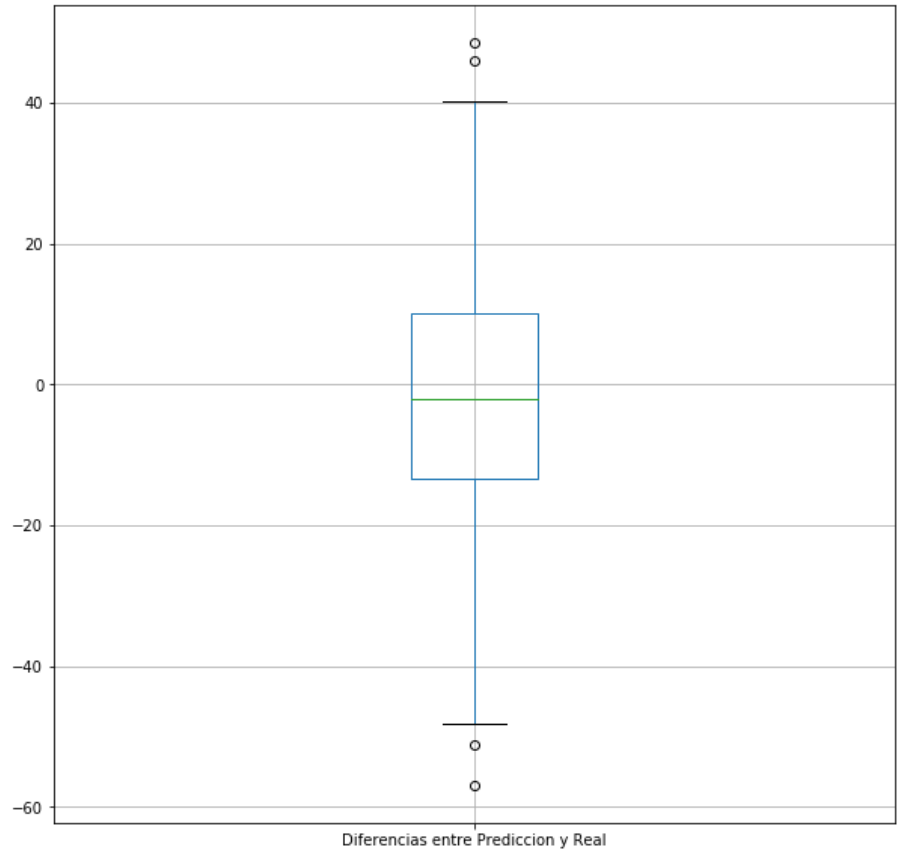


Figura 7.4.p: Representación de la diferencia entre la predicción y el valor real de los días de test

En este caso se observa como la mayoría de los datos están entre -13,49 y 10,03 visitas fallidas a la hora de predecir. También hay bastantes valores que se predicen sin apenas error aunque es cierto que también hay algunos outliers con bastante error.

Se realizaron dos gráficos más para ver el resultado de la predicción en base al día de la semana para poder visualizar mejor los datos y sacar alguna conclusión más.

Los puntos del mismo color indican el mismo día de la semana de la siguiente forma: Lunes: Verde, Martes: Azul, Miércoles: Rojo, Jueves: Cyan, Viernes: Magenta, Sábado: Amarillo y Domingo: Negro.

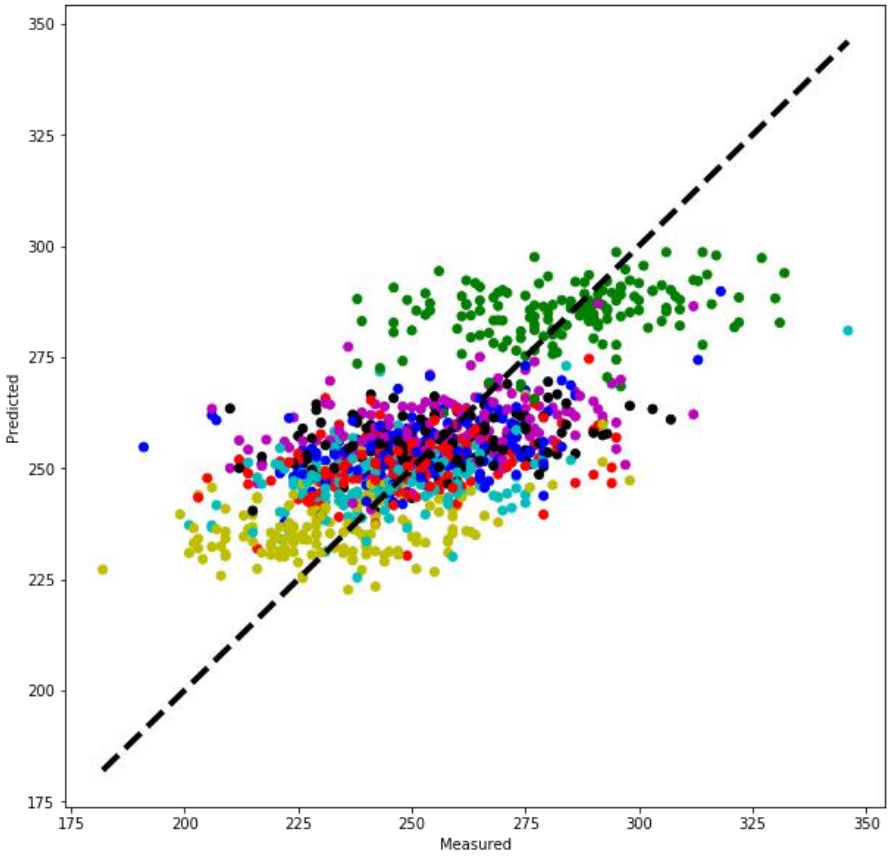


Figura 7.4.q: Representación gráfica de los valores predichos y los valores reales para el conjunto de entrenamiento con un color para cada día de la semana

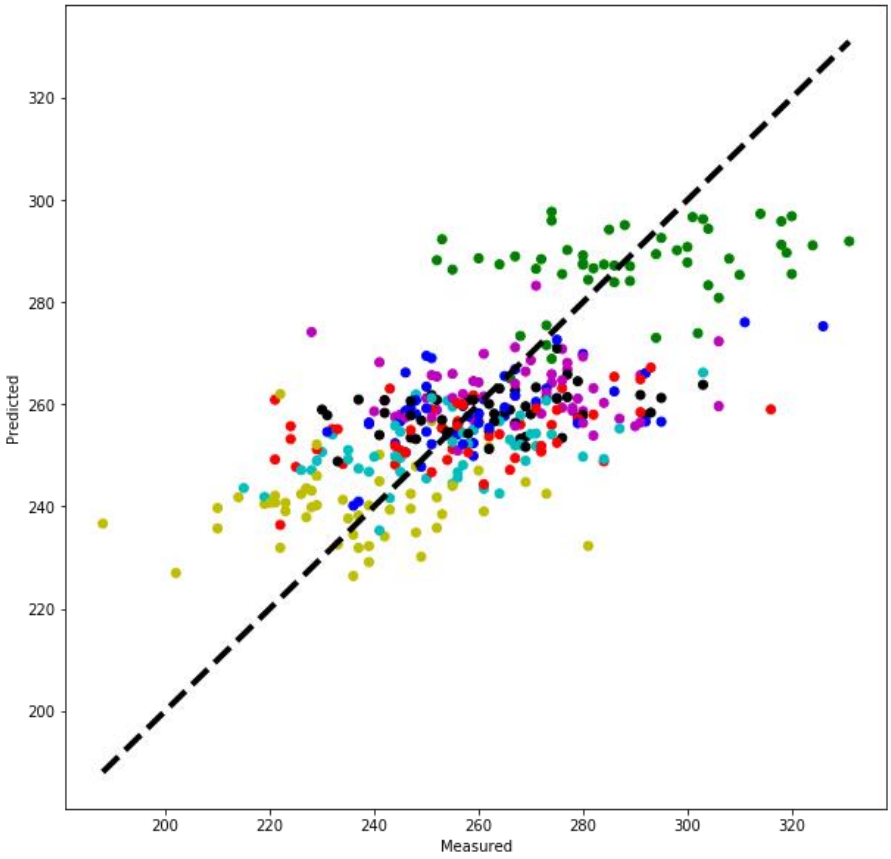


Figura 7.4.r: Representación gráfica de los valores predichos y los valores reales para el conjunto de test con un color para cada día de la semana

Para las dos gráficas se observa como los lunes son los días con más visitas y los sábados los días con menos aunque la distribución alrededor de la diagonal es similar al resto de días de la semana.

7.5.3. Comparación entre modelos

Con el fin de poder evaluar la calidad del modelo evaluado se realizó un conjunto de comparaciones. Se compararán los valores predichos por los modelos, tanto ARIMA como LSTM, los valores obtenidos prediciendo el número medio de visitas totales y prediciendo el número medio de visitas por cada día de la semana.

Es decir, se calculará el error obtenido comparando el conjunto de test predicho con el conjunto de test real; el error obtenido comparando el conjunto de test, como si predijera siempre el número medio de visitas, con el conjunto de test real; y el error obtenido comparando el conjunto de test, como si predijera el número medio de visitas para cada día de la semana, con el conjunto de test real.

De este modo nos podemos hacer una idea de la efectividad de la predicción y si mejoraría el resultado comparando con los modelos anteriores.

Primero se visualiza la comparación para el conjunto de test de todas las urgencias del Complejo Hospitalario de Navarra. En este caso es una comparación entre los modelos de LSTM. En la *Figura 7.4.s* se puede observar el error cometido por cada configuración. Una pequeña descripción de cada caja sería: la línea naranja marca la mediana la caja indica dónde se encuentran la mayoría de los errores, los límites de las líneas verticales marcan el primer y tercer cuartil y los puntos son los outliers.

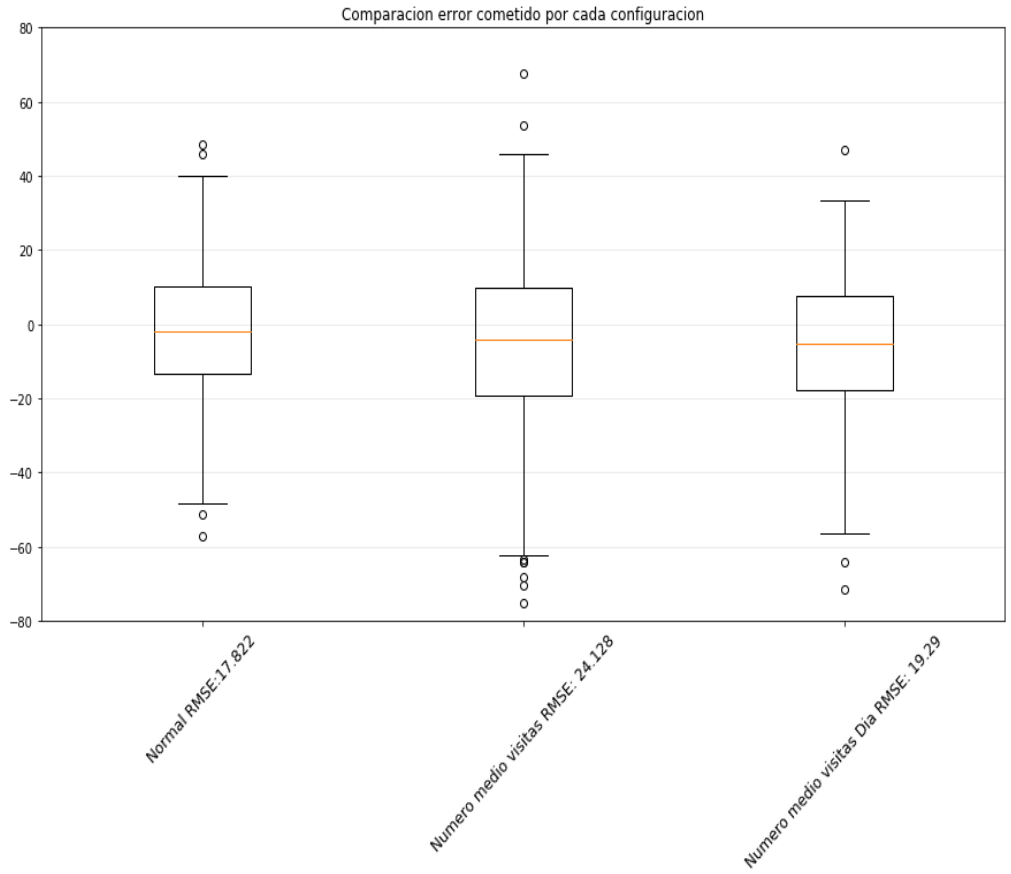


Figura 7.4.s: comparación del error de cada configuración en conjunto junto con el RMSE en el conjunto de test para todas las visitas a urgencias del Complejo Hospitalario de Navarra.

Ahora se puede observar como el modelo que realiza la predicción obtiene un error menor que las otras configuraciones así como un RMSE menor. Por tanto podemos concluir que el modelo implementado tiene bastante calidad para el problema planteado.

En cuanto al dataset generado por las personas >=65 años que han acudido al Complejo Hospitalario de Navarra se puede observar lo siguiente. (En este caso se encuentra incluido el modelo correspondiente al obtenido en el apartado de ARIMA).

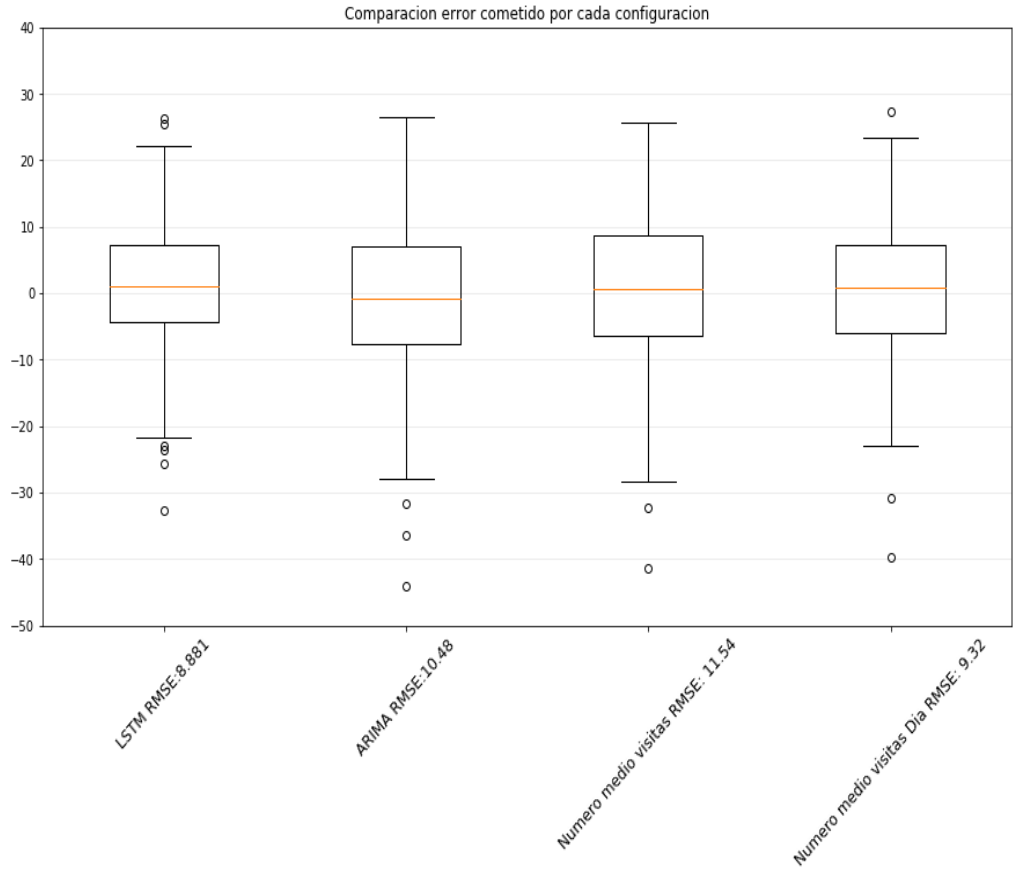


Figura 7.4.t: comparación del error de cada configuración en conjunto junto con el RMSE en el conjunto de test para las visitas a urgencias del dataset de personas >=65 años del Complejo Hospitalario de Navarra.

Se observa un comportamiento similar al anterior dataset por lo que podemos concluir que se obtiene un modelo de bastante calidad para este problema. El resultado de ARIMA es bastante aceptable teniendo en cuenta que está modelado tan solo con una variable y está bastante parejo al modelo LSTM. De todas formas, el modelo LSTM es el que mejor funciona al tener un comportamiento más ajustado al error 0. Sin embargo se puede observar como el comportamiento de los tres modelos es más parejo que para el dataset con todas las visitas a urgencias.

-Conclusiones:

Se observa como en general es bastante difícil predecir con exactitud el número de urgencias para un día en concreto. Las grandes diferencias que existen entre días bastante cercanos en el tiempo hacen que los datos no siguen una tendencia más o menos clara y por ello los errores cometidos son mayores de lo esperado. Con todo, el modelo LSTM es el que mejores resultados arroja en comparación con los otros modelos.

Analizando los datasets por separado en primer lugar en el dataset>=65 se ve como el RMSE es de 8,881 para LSTM. Si hacemos una media de las visitas a urgencias nos

sale 71,76. Por tanto el error cometido en media sería del 12,37% (8,881/71,76). Aunque si elegimos el día que más error se ha cometido en test (-32,67) el error en este caso sería del 28,93%. Esto se calcula dividiendo el valor predicho para ese día entre el valor real de ese día (80,335/113), multiplicándolo por 100 para obtener el porcentaje y restándole 100 para obtener el error. Como información adicional el día con menor error sería de 0,07 por lo que ese día la precisión sería total.

Sin embargo si nos fijamos en el otro dataset los resultados son algo más optimistas. El RMSE es de 17,822 y la media de visitas a urgencias es de 255,76. Por tanto el error cometido en media sería del 6,97% (17,822/255,76). Eligiendo el día con peor precisión en test (-56,97), el error cometido en este caso sería del 18,02%. Esto se calcula dividiendo el valor predicho para ese día entre el valor real de ese día (259,032/316), multiplicándolo por 100 para obtener el porcentaje y restándole 100 para obtener el error. Además el mejor día predicho arrojaría un error de 0,028 visitas ese día por lo que acertaría de pleno en ese día.

Es por ello que no se puede utilizar este modelo como una predicción muy fiable aunque en el caso de las visitas a urgencias totales el modelo funciona mejor y se podría utilizar como un indicador de la demanda asumiendo siempre el error cometido. De la misma forma, se ha comprobado que la tendencia es prácticamente siempre bien predicha. Es decir, se sabría de forma casi exacta si al día siguiente el número de visitas a urgencias va a aumentar o disminuir comparado con el día anterior. Con esta información conjunta los equipos sanitarios podrían obtener una ayuda a la hora de hacer una mejor distribución tanto del personal como del material sanitario en base al día en cuestión.

Además se ha visto como el modelo funciona mejor que una predicción basada en el número medio de visitas totales y el número medio de visitas por día lo que le añade cierto valor al modelo y puede llegar a ser útil para el conocimiento del servicio de urgencias.

8. Conclusiones y líneas futuras

Tras haber analizado en profundidad las visitas a urgencias y sus derivados se pueden extraer una serie de conclusiones que pueden facilitar el trabajo en el futuro.

Lo primero de todo se ha requerido un gran trabajo de extracción de datos y de familiarización con los sistemas informáticos en dónde la ayuda de Javier Gorricho y Garbiñe Basterra ha sido fundamental en el proceso.

Desde el principio del proyecto ya se suponía que el hecho de que una persona en particular acuda o no al servicio de urgencias era debido a multitud de factores que no podían ser analizados en este proyecto. Además, con los datos disponibles ya se ha podido observar como su comportamiento es bastante impredecible y varía mucho en días consecutivos sin un patrón claro.

Se ha visto como en época de gripe las visitas al servicio de urgencias aumentan de forma significativa en personas mayores o iguales de 65 años. Además se observa de forma clara como cuando es festivo este número baja notablemente. También se ha llegado a observar que un aumento de la temperatura media diaria está correlacionado con un aumento en las visitas a urgencias. Por otra parte, también hay que resaltar que temperaturas muy bajas también hacen que este número aumente.

Sabiendo lo anterior, se hacía bastante complicado el llegar a predecir con exactitud el número de personas que acudirán un día al servicio o si una persona en concreto haría uso del servicio de urgencias.

Como se ha podido comprobar, el realizar una predicción a nivel de paciente es muy complicado debido a la multitud de factores que entran en juego. Sin embargo, a nivel de día sí que se han podido obtener ciertos resultados de valor que pueden ser de cierto interés para el servicio de urgencias siempre y cuando se tenga en cuenta el error cometido.

Dicho esto, cabe recalcar que mediante la utilización de diferentes modelos, técnicas, variables, etc. se podría llegar a mejorar estos resultados, así como llegar a obtener diferentes conclusiones que ayudarían a obtener una visión tanto a presente como a futuro del comportamiento de las urgencias. Así mismo, con diferentes configuraciones para el algoritmo ARIMA o el algoritmo LSTM se podrían haber obtenido diferentes resultados que, por falta de tiempo, no ha sido posible el probar todas las configuraciones posibles.

Sería de gran valor, además, el poder poseer más información de cada visita a urgencias registrada. Por ejemplo, el tipo de urgencia registrada o la gravedad de la misma. Esto podría acotar mucho más los resultados y podríamos obtener unos resultados más ajustados. En nuestro proyecto hubiera sido de gran utilidad el saber qué urgencias habrían sido de tipo respiratorio o cardiovascular ya que son las que más impacto tienen en los factores ambientales. Por otro lado, la gravedad de la

visita arrojaría más calidad debido a que los casos urgentes podrían ser tratados de forma separada.

También sería de utilidad el uso de otro tipo de variables más difíciles de recoger que no sean atmosféricas, como el número de caídas, acontecimientos multitudinarios, etc. pero que tienen bastante impacto en el número final de visitas al servicio de urgencias.

9. Anexos

9.1. Anexo A: Estructura y funcionamiento general del Sistema Sanitario Navarro

Es de vital importancia conocer cómo está estructurado el actual Servicio Navarro de Salud y cuál es su funcionamiento para entender correctamente los servicios que se describen.

9.1.1. Funcionamiento del sistema sanitario

Actualmente el sistema sanitario español se caracteriza por tener dos tipos claramente diferenciados de atención al paciente: La atención primaria (AP) y la atención especializada (AE). La atención primaria es atendida en los centros de salud y cuando no puede ser atendida o es requerida una intervención, se deriva al especialista. Tras ello, el paciente puede volver a ser atendido por el médico de AP o seguir tratándolo el mismo. En cuanto a sus funciones en:

-Atención Primaria (AP): A cada paciente le es asignado una Unidad Básica de Atención (un médico/a y una enfermera/o) y la función básica es tener un seguimiento continuo del paciente durante toda su vida.

-Atención Especializada u Hospitalaria (AE): El paciente puede ser derivado desde Atención Primaria a un médico especialista en el área correspondiente para tratar un episodio concreto.

Cabe destacar que el médico especialista no tiene por qué conocer el historial clínico del paciente y por ello la atención primaria es básica. Por tanto, un paciente cuando lo requiera puede ir a su médico de Atención Primaria y éste si lo cree necesario lo delegará a un especialista. Además si es un caso de urgencia, el paciente deberá acudir a Urgencias en dónde es posible que sea ingresado u operado.

9.1.2. Estructuración geográfica

En Navarra, el sistema sanitario está dividido en áreas sanitarias y zonas básicas como se muestra en la *Figura 9.1.a*

Figura 9.1.a: mapa de las áreas sanitarias de Navarra junto con las zonas básicas correspondientes. (Fuente: Sistemas de recogida de información y evaluación de registros sobre Diabetes en Navarra. En es.slideshare.net [23])

-Áreas sanitarias: La comunidad se divide en tres áreas sanitarias: Pamplona (Naranja y verde), Estella (Amarillo) y Tudela (Azul). A la derecha tenemos la ciudad de Pamplona de cerca. Cada área sanitaria tiene un hospital asignado para los pacientes que residan en esas zonas. Así para Pamplona se tiene el Complejo Hospitalario de Navarra (CHN), para Estella el Hospital García Orcoyen y para Tudela el Hospital Reina Sofía.

-Zonas básicas: Cada área, por su parte, tiene múltiples zonas básicas que también pueden verse en el mapa. Las áreas tienen como referencia un hospital, y las zonas básicas un centro de salud.

Esta estructuración ofrece distintos tipos de atención. La atención básica se realiza en centros de salud o en su defecto, consultorios. En caso de ser necesaria una atención hospitalaria o especializada, el paciente es derivado a su hospital correspondiente. Si es necesaria una mayor atención o especialización entonces se le deriva

directamente al CHN (Compuesto por Hospital de Navarra, Hospital Virgen del Camino y Clínica Ubarmin).

9.1.3. Estructura organizativa del Departamento de Salud de Navarra

El Departamento de Salud de Navarra consta de:

- **Dirección General de Salud** que está subdividida en:
  - Servicio de Ciudadanía Sanitaria, Aseguramiento y Garantías: Encargado de Asociaciones (SIDA, cáncer, etc.), Inspección y Autorización de farmacia, etc.
  - Servicio de Planificación, Evaluación y Gestión del Conocimiento: Encargado de la Planificación Sanitaria, Desarrollo del Plan de Salud, Evaluación Sanitaria, Formación continua y de residentes, Acreditación de Formación, etc.
  - Instituto de Salud Pública y Laboral de Navarra (organismo autónomo): Sus funciones se basan en Vigilancia, Control e Intervención de brotes epidémicos y posibles situaciones de riesgo, Programas de Vacunación, Prevención de enfermedades, Promoción y Educación para la Salud, Prevención de riesgos laborales, Actividades de Formación, etc.
- **Servicio Navarro de Salud – Osasunbidea** (organismo autónomo): Sus funciones se basan en Asistencia Sanitaria (Atención Primaria, Atención Especializada, Hospitales, Salud Mental, etc.), Rehabilitación y Reinserción, Orientación sexual y familiar, Prestación de asistencia farmacéutica, Formación del personal sanitario, etc.
- **Fundación Miguel Servet o NavarraBiomed**: ocupada de fomentar, promover, liderar y gestionar un entorno científico y tecnológico dedicado a la investigación para mejorar la salud de los ciudadanos.
- **Secretaría General Técnica**: Atiende las necesidades del Departamento en materias comunes de índole técnica y jurídica, de recursos humanos y de gestión presupuestaria

Cabe destacar que los organismos autónomos pueden encargarse de forma propia de funciones como contratación y compra de material, sin necesidad de pasar por la Dirección General.

De forma esquemática se puede observar la *Figura 9.1.b*:

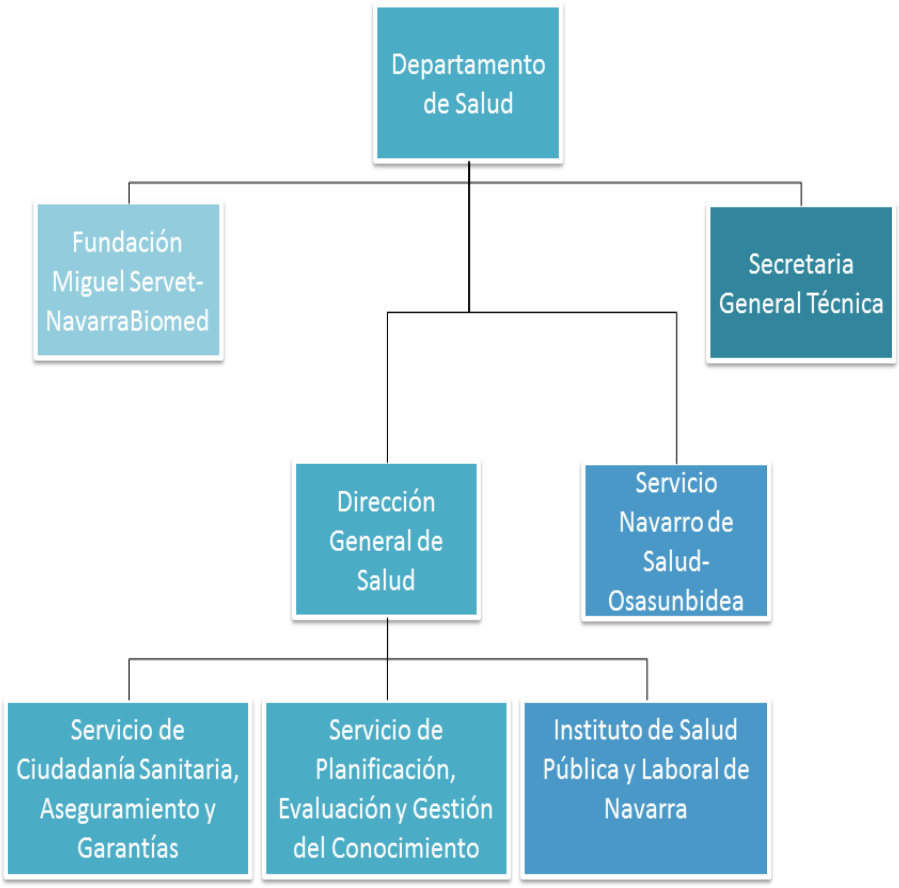


Figura 9.1.b: estructura organizativa del Departamento de Salud

La estructura organizativa del SNS-O es la observada en la *Figura 9.1.c*

Figura 9.1.c: estructura organizativa del SNS-O

Los órganos centrales trabajan para todo el sistema sanitario navarro y los periféricos para partes más concretas. Los órganos periféricos se componen de una gerencia por cada área sanitaria (Pamplona, Estella y Tudela), Atención Primaria y Salud Mental.

9.2. Anexo B: Documentación de la base de datos generada para el proyecto

Universo del estudio: Pacientes registrados en Servicio Navarro de Salud en enero de 2012 y mayores o iguales a 65 años en enero de 2013.

9.2.1. Pacientes

Extraído de an.DIM_PACIENTES

PACIENTES		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	
NHC_ANONIMO	varchar(64)	
NOMBRE_ANONIMO	varchar(64)	
APELLIDO1_ANONIMO	varchar(64)	
APELLIDO2_ANONIMO	varchar(64)	
CIAS	varchar(11)	
COD_TIPO_PACIENTE	varchar(1)	
DES_TIPO_PACIENTE	varchar(50)	
EDAD	int	
FECHA_FALLECIMIENTO	date	
SEXO	varchar(1)	H / M
CODIGO_PAIS_NACIMIENTO	varchar(3)	
NOMBRE_PAIS_NACIMIENTO	varchar(40)	
CODIGO_PAIS_RESIDENCIA	varchar(3)	
NOMBRE_PAIS_RESIDENCIA	varchar(40)	
CODIGO_POSTAL_ANONIMO	varchar(64)	

CODIGO_DEL_MUNICIPIO	varchar(3)	
NOMBRE_MUNICIPIO	varchar(50)	
COD_ZONA_DOMICILIO	varchar(4)	
COD_ZONA_CENTRO_ASIGNADO	varchar(4)	
CUPO_MEDICO	int	
DEPENDENCIA	int	Valores: • 0:NO DEPENDIENTE • 1:DEPENDIENTE
FECHA_NACIMIENTO	date	
DOMICILIO	nvarchar(150)	
LATITUD	float	
LONGITUD	float	
COTIZACION_FARMACIA	nvarchar(15)	Valores: • TSI 001 (): EXENTOS DE APORTACIÓN • TSI 002 (): APORTACIÓN DE UN 10% • TSI 002 (00): APORTACIÓN DE UN 10% • TSI 002 (01): APORTACIÓN DE UN 10% • TSI 002 (02): APORTACIÓN DE UN 10% • TSI 003 (): APORTACIÓN DE UN 40%

		• TSI 004 (): APORTACIÓN DE UN 50% • TSI 005 (): APORTACIÓN DE UN 60% • TSI 005 (03): APORTACIÓN DE UN 60% • TSI 006 (): EXCLUIDOS DE FARMACIA(MUFACE, MUGEJU, ISFAS) • TSI 006 (00): EXCLUIDOS DE FARMACIA(MUFACE, MUGEJU, ISFAS) • NULL: SIN INFORMACIÓN
NIVEL_COPAGO_BAJO	int	• 0: NIVEL COPAGO NO BAJO: TSI 002 (), TSI 002 (00), TSI 002 (02), TSI 004 (), TSI 005 (), TSI 005 (03), TSI 006 (), TSI 006 (00) • 1:NIVEL COPAGO BAJO: TSI 001 (), TSI 002 (01), TSI 003 ()
ESTACION	varchar(50)	
TRATAMIENTO_HOSPITAL_DIA	int	Valores: • 0:No recibe HOSPITAL_DIA • 1:Recibe

		HOSPITAL_DIA
GMA	float	Valores: • 1: Grupo morbilidad muy baja • 2: Grupo morbilidad baja • 3: Grupo morbilidad media • 4: Grupo morbilidad alta • 5: Grupo morbilidad muy alta

Cambios:

- FECHA_FALLECIMIENTO:9999-12-31→NULL

9.2.2. Diagnósticos AP

Extraído de atenea.TH_DIAGNOSTICOS, atenea.DIM_CIAPS

Fecha inicio >=01/01/2013 y <=31/12/2016

DIAGNOSTICOS_AP		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	
DESCRIPCION1	varchar(150)	
DESCRIPCION2	varchar(255)	
FECHA_INICIO	date	
FECHA_CIERRE	date	

COD_CIAP	varchar(15)	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA_FALLECIMIENTO:9999-12-31→NULL
- FECHA_INICIO: Varchar→date
- FECHA_CIERRE: Int→date
- FECHA_CIERRE:1800-12-28→NULL

9.2.3. Estaciones

ESTACIONES		
PK(ESTACION,FECHA)		
Nombre	Tipo	Valor
ESTACION	varchar(50)	
ANO	nvarchar(50)	
TEMPERATURA_MAXIMA	float	
TEMPERATURA_MINIMA	float	
MES	nvarchar(50)	
DIA	nvarchar(50)	
AMPLITUD_TERMICA	float	
ETIQUETA_MAXIMAS	nvarchar(50)	Valores: • Muy Baja: TEMPERATURA_MAXIMA <6 • Baja: TEMPERATURA_MAXIMA <14 & TEMPERATURA_MAXIMA

		>=6 • Media: TEMPERATURA_MAXIMA <23 & TEMPERATURA_MAXIMA >=14 • Alta: TEMPERATURA_MAXIMA <33 & TEMPERATURA_MAXIMA >=23 • Muy Alta: TEMPERATURA_MAXIMA >=33
ETIQUETA_MINIMAS	nvarchar(50)	Valores: • Muy Baja: TEMPERATURA_MINIMA <=-1 • Baja: TEMPERATURA_MINIMA >-1 & TEMPERATURA_MINIMA <=4 • Media: TEMPERATURA_MINIMA >4 & TEMPERATURA_MINIMA <=10 • Alta: TEMPERATURA_MINIMA >10 & TEMPERATURA_MINIMA <=16 • Muy Alta: TEMPERATURA_MINIMA >16
ETIQUETA_AMPLITUDES	nvarchar(50)	Valores: • Muy Baja: AMPLITUD_TERMICA<=3 • Baja: AMPLITUD_TERMICA>3 & AMPLITUD_TERMICA<=7

		• Media: AMPLITUD_TERMICA>7 & AMPLITUD_TERMICA<=13 • Alta: AMPLITUD_TERMICA>13 & AMPLITUD_TERMICA<=18 • Muy Alta: AMPLITUD_TERMICA>18
VARIACION_MAXIMAS	float	
VARIACION_MINIMAS	float	
ETIQUETA_VARIACION_MAXIMAS	nvarchar(50)	Valores: • Baja: VARIACION_MAXIMAS >=0 & VARIACION_MAXIMAS<=1 • Media: VARIACION_MAXIMAS>1 & VARIACION_MAXIMAS<=4 • Alta: VARIACION_MAXIMAS >4
ETIQUETA_VARIACION_MINIMAS	nvarchar(50)	Valores: • Baja: VARIACION_MINIMAS>=0 & VARIACION_MINIMAS<=1 • Media: VARIACION_MINIMAS>1 & VARIACION_MINIMAS<=4 • Alta: VARIACION_MINIMAS>4
VALOR_MAXIMAS	float	
VALOR_MINIMAS	float	
VALOR_AMPLITUDES	float	
FECHA	date	
PROBABILIDAD	float	

EPIDEMIA	nvarchar(1)	Valores: • S: Dentro del periodo de epidemia de gripe en base a distribución gaussiana para el 95.5% de los casos • N: Fuera del periodo de epidemia de gripe en base a distribución gaussiana para el 95.5% de los casos
OLA_CALOR	varchar(1)	Valores: • S: Si VALOR_MAXIMAS >= 3 • N: SI VALOR_MAXIMAS < 3
OLA_FRIO	varchar(1)	Valores: • S: Si VALOR_MAXIMAS >= 3 • N: SI VALOR_MAXIMAS < 3
ACUMULADA_MAXIMAS	float	
ACUMULADA_MINIMAS	float	
EPIDEMIA_68	varchar(1)	Valores: • S: Dentro del periodo de epidemia de gripe en base a distribución gaussiana para el 68.3% de los casos • N: Fuera del periodo de epidemia de gripe en base a distribución gaussiana para el

		68.3% de los casos
ETIQUETA_ACUMULADA_MAXIMAS	varchar(1)	Valores: • S: SI ACUMULADA_MAXIMAS >=70 • N: SI ACUMULADA_MAXIMAS <70
ETIQUETA_ACUMULADA_MINIMAS	varchar(1)	Valores: • S: SI ACUMULADA_MINIMAS <=18 • N: SI ACUMULADA_MINIMAS >18
EXTREMA_MAXIMAS	varchar(1)	Valores: • S: SI TEMPERATURA_MAXIMA>=95 PERCENTIL • N: SI TEMPERATURA_MAXIMA< 95 PERCENTIL
EXTREMA_MINIMAS	varchar(1)	Valores: • S: SI TEMPERATURA_MINIMA<= 5 PERCENTIL • N: SI TEMPERATURA_MINIMA> 5 PERCENTIL
OLA_DE_CALOR	varchar(1)	Valores: • S: SI LOS DOS DÍAS ANTERIORES + DÍA EN CUESTIÓN HAN SIDO EXTREMA_MAXIMAS • N: SI LOS DOS DÍAS ANTERIORES + DÍA EN CUESTIÓN NO HAN SIDO

		EXTREMA_MAXIMAS
OLA_DE_FRIO	varchar(1)	Valores: • S: SI LOS DOS DÍAS ANTERIORES + DÍA EN CUESTIÓN HAN SIDO EXTREMA_MINIMAS • N: SI LOS DOS DÍAS ANTERIORES + DÍA EN CUESTIÓN NO HAN SIDO EXTREMA_MINIMAS
INFLUENCIA_1_DIA	varchar(1)	Valores: • S: SI ES EL DÍA SIGUIENTE A UNA TEMPERATURA MÍNIMA EXTREMA • N: SI NO ES EL DÍA SIGUIENTE A UNA TEMPERATURA MÍNIMA EXTREMA
INFLUENCIA_2_DIA	varchar(1)	Valores: • S: SI ESTÁ DENTRO DE LOS DOS DÍAS SIGUIENTES A UNA TEMPERATURA MÍNIMA EXTREMA • N: SI NO ESTÁ DENTRO DE LOS DOS DÍAS SIGUIENTES A UNA TEMPERATURA MÍNIMA EXTREMA
INFLUENCIA_3_DIA	varchar(1)	Valores: • S: SI ESTÁ DENTRO DE LOS TRES DÍAS SIGUIENTES A UNA TEMPERATURA MÍNIMA EXTREMA

		• N: SI NO ESTÁ DENTRO DE LOS TRES DÍAS SIGUIENTES A UNA TEMPERATURA MÍNIMA EXTREMA
INFLUENCIA_SOLO_2_DIA	varchar(1)	Valores: • S: SI ES EL SEGUNDO DÍA TRAS UNA TEMPERATURA MÍNIMA EXTREMA • N: SI NO ES EL SEGUNDO DÍA TRAS UNA TEMPERATURA MÍNIMA EXTREMA
INFLUENCIA_SOLO_3_DIA	varchar(1)	Valores: • S: SI ES EL TERCER DÍA TRAS UNA TEMPERATURA MÍNIMA EXTREMA • N: SI NO ES EL TERCER DÍA TRAS UNA TEMPERATURA MÍNIMA EXTREMA
GRAN_AMPLITUD	varchar(1)	Valores: • S: Si VALOR_AMPLITUDES >= 3 • N: SI VALOR_AMPLITUDES < 3

9.2.4. Información

INFORMACION		
PK(ESTACION)		
Nombre	Tipo	Valor
ESTACION	nvarchar(50)	
ALTITUD	nvarchar(50)	
FECHA_INSTALACION	date	
PERTENENCIA	nvarchar(50)	Valores: • GN Y AEMET • GN • INTIA • MAPAMA
LONGITUD	nvarchar(50)	
LATITUD	nvarchar(50)	
CERCANA	nvarchar(50)	
Foreign Key	Referencia a	
ESTACION	ESTACIONES	

9.2.5. Medicamentos receta

Extraído de lamia.TH_PRESCRIPCIONES_DISPENSACIONES, lamia.DIM_MEDICAMENTOS, lamia.DIM_TIPO_PRESCRIPCION

MEDICAMENTOS_RECETA		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	

COD_PRESCRIPCION	int	
COD_RECETA	int	
NUM_EPISODIO	varchar(50)	
PK_TIPO_PRESCRIPCION	int	'Hay que dejar el PK porque no tiene código asignado'
FECHA_INICIO_PRESCRIPCION	date	
FECHA_FIN_PRESCRIPCION	date	
FECHA_SUSPENSION	date	
COD_MEDICAMENTO_PRESCRITO	int	
COD_MEDICAMENTO_DISPENSADO	int	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA_INICIO_PRESCRIPCION: Varchar→date
- FECHA_FIN_PRESCRIPCION: Varchar→date
- FECHA_SUSPENSION: Varchar→date
- FECHA_FIN_PRESCRIPCION: 1899-12-30→NULL

9.2.6. Medicamentos farho

Extraído de farho.TH_DISPENSACIONES, farho.TH_ARTICULOS

MEDICAMENTOS_FARHO		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	

COD_ARTICULO	varchar(7)	
DES_ARTICULO	varchar(50)	
COD_PF5	int	
DES_PF5	varchar(50)	
COD_ATC	varchar(7)	
DES_ATC	varchar(110)	
DES_ARTICULO_GENERICA	varchar(50)	
SK_TH_DISPENSACIONES	int	
PK_FECHA	varchar(8)	
SK_PACIENTE	int	
TIPO_CASO_DESCRIPCION	varchar(20)	
FECHA_COMIENZO	date	
FECHA_FINAL	date	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA_COMIENZO: Varchar→date
- FECHA_FINAL: Varchar→date

9.2.7. Visitas urgencias

Extraído de leire.TH_URGENCIAS,leire.DIM_TIPO_PROCEDENCIA_PACIENTE, leire.DIM_TIPO_ALTA_URGENCIAS,leire.DIM_COMPLEJOS, leire.DIM_ESPECIALIDADES, leire.DIM_SERVICIOS, leire.DIM_UNIDADES_MEDICAS, dbo.DIM_PLAZAS_AE

VISITAS_URGENCIAS		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	
SK_PLAZAS_AE	int	
NUM_CASO_URGENCIAS	varchar(8)	
DESC_URGENCIA	varchar(40)	
DIAGNOSTICO_HCI	varchar(8000)	
FECHA	date	
FECHA_ALTA	date	
COD_TIPO_PROCEDENCIA_PACIENTE	varchar(2)	
COD_TIPO_ALTA_URGENCIAS	varchar(2)	
COD_COMPLEJO	varchar(2)	
COD_ESPECIALIDAD	varchar(2)	
COD_SERVICIO	varchar(3)	
COD_UNIDAD_MEDICA	varchar(5)	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA: Varchar→date
- FECHA_ALTA. Varchar→date

9.2.8. CMBD

Extraído de CMBD, leire.DIM_TIPO_INGRESO, dbo.DIM_HOSPITALES

CMBD		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	nvarchar(255)	
TIPO_INGRESO	float	
DESCRIPCION_TIPO_INGRESO	nvarchar(255)	
TIPO_ALTA	float	
DESCRIPCION_TIPO_ALTA	nvarchar(255)	
DIAGNOSTICO_1	nvarchar(255)	
DIAGNOSTICO_2	nvarchar(255)	
DIAGNOSTICO_3	nvarchar(255)	
DIAGNOSTICO_4	nvarchar(255)	
DIAGNOSTICO_5	nvarchar(255)	
DIAGNOSTICO_6	nvarchar(255)	
DIAGNOSTICO_7	nvarchar(255)	
DIAGNOSTICO_8	nvarchar(255)	
DIAGNOSTICO_9	nvarchar(255)	
DIAGNOSTICO_10	nvarchar(255)	
DIAGNOSTICO_11	nvarchar(255)	
DIAGNOSTICO_12	nvarchar(255)	
DIAGNOSTICO_13	nvarchar(255)	
DIAGNOSTICO_14	nvarchar(255)	
DIAGNOSTICO_15	nvarchar(255)	
DIAGNOSTICO_16	nvarchar(255)	

DIAGNOSTICO_17	nvarchar(255)	
DIAGNOSTICO_18	nvarchar(255)	
DIAGNOSTICO_19	nvarchar(255)	
DIAGNOSTICO_20	nvarchar(255)	
HOSPITAL	float	
FECHA_INGRESO	date	
FECHA_ALTA	date	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA_INGRESO: Nvarchar→date
- FECHA_ALTA: Nvarchar→date

9.2.9. Urgencias extrahospitalarias

Extraído de atenea.TH_AGENDAS

URGENCIAS_EXTRAHOSPITALARIAS		
PK(CIPNA_ANONIMO)		
Nombre	Tipo	Valor
CIPNA_ANONIMO	varchar(64)	
COD_ACTO	Varchar(10)	
DES_ACTO	varchar(200)	
COD_TIPO_ACTO	varchar(2)	
DES_TIPO_ACTO	varchar(100)	
TIPO_VISITA	varchar(30)	
OBSERVACIONES	varchar(255)	

COD_ESPECIALIDAD	varchar(3)	
FECHA_CITA	date	
FECHA_SOLICITUD_CITA	date	
Foreign Key	Referencia a	
CIPNA_ANONIMO	PACIENTES	

Cambios:

- FECHA_CITA: Int→date
- FECHA_SOLICITUD_CITA: Int→date

10. Bibliografía

[1] Linwei Tian, et al. *Emergency Cardiovascular Hospitalization Risk Attributable to Cold Temperatures in Hong Kong*. *Circ Cardiovasc Qual Outcomes, 9: 135-142*, Marzo 2016

[2] Anna Ponjoan, et al. *Effects of extreme temperatures on cardiovascular emergency hospitalizations in a Mediterranean region: a self-controlled case series study. Environmental Health, vol 16, pp 1-32,* Abril 2017

[3] Tianqi Chen, et al. *Time-series Analysis of Heat Waves and Emergency Department Visits in Atlanta, 1993 to 2012*. *EHP: Environmental Health Perspectives*, Mayo 2017

[4]Gobierno de Navarra. http://www.navarra.es/home_es/Temas/Medio+Ambiente/Calidad+del+aire/, Diciembre 2017.

[5] Gobierno de Navarra. http://meteo.navarra.es/estaciones/mapadeestaciones.cfm, Diciembre 2017

[6] Jesús García Jiménez-Blog. https://jesusgarciaj.com/2010/01/22/la-curva-de-distribucion-normal/, Enero 2018

[7] Dataprix.TI http://www.dataprix.com/datawarehouse-manager, Enero 2018

[8] Wikipedia. https://es.wikipedia.org/wiki/Correlaci%C3%B3n, Febrero 2018

[9] Wikipedia. https://es.wikipedia.org/wiki/Estudio_observacional, Febrero 2018

[10] Wikipedia. https://es.wikipedia.org/wiki/Regresi%C3%B3n_de_Poisson, Febrero 2018

[11] Wikipedia. https://es.wikipedia.org/wiki/Riesgo_relativo. Febrero 2018

[12] Krishnan Bhaskaran et al. *Time series regression studies in environmental*

Epidemiology. International Journal of Epidemiology 2013; 42:1187–1195, Abril 2013

[13] Wikipedia. https://es.wikipedia.org/wiki/Serie_temporal, Febrero 2018

[14] DigitalOcean. https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3, Marzo 2018

[15] Wikipedia. https://es.wikipedia.org/wiki/Modelo_autorregresivo_integrado_de_media_m%C3%B3vil, Febrero 2018

[16] Wikipedia. https://es.wikipedia.org/wiki/Red_neuronal_artificial, Febrero 2018

[17] Wikipedia. https://en.wikipedia.org/wiki/Long_short-term_memory, Febrero 2018

[18] Wikipedia. https://es.wikipedia.org/wiki/Aprendizaje_supervisado, Febrero 2018

[19] Wikipedia. https://en.wikipedia.org/wiki/Root-mean-square_deviation, Marzo 2018

[20] Machine Learning Mastery. https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/, Enero 2018

[21] S3lab Security Blog. http://s3lab.deusto.es/medidas-calidad-algoritmos-clasificacion/, Enero 2018

[22] Wikipedia. https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica, Febrero 2018

[23] Wikipedia. https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo, Febrero 2018

[24] Scikit-learn. http://scikit-learn.org/stable/modules/naive_bayes.html, Febrero 2018

[25] SlideShare. https://es.slideshare.net/sanidadyconsumo/sistemas-de-recogida-de-informacin-y-evaluacin-de-registros-sobre-diabetes-en-navarra, Diciembre 2017