UNIVERSIDAD DE JAÉN

ESCUELA POLITÉCNICA SUPERIOR DE JAÉN
DEPARTAMENTO DE INFORMÁTICA

TESIS DOCTORAL

AVANCES DE SISTEMAS DE INFORMACIÓN GEOGRÁFICA Y MINERÍA DE DATOS PARA APLICACIONES DE E-SALUD

PRESENTADA POR:
JUAN JOSÉ CUBILLAS MERCADO

DIRIGIDA POR:
DR. D. FRANCISCO R. FEITO HIGUERUELA
DRA. DÑA. MARÍA ISABEL RAMOS GALÁN

JAÉN, 19 DE FEBRERO DE 2015

ISBN 978-84-8439-947-6

A mi Mujer y mis hijos: Carmen y Jorge.

RESUMEN

Hoy en día los sistemas de información son fundamentales en el ámbito sanitario, tanto desde el punto de vista de gestión de recursos, como desde el punto de vista puramente sanitario. En este trabajo avanzamos en la optimización de los recursos sanitarios basándonos en la minería de datos y en la obtención de variables espaciales relacionadas con los centros de salud, con el objetivo principal de crear modelos para predecir la afluencia de pacientes a los centros de salud y determinar las variables que afectan a dicha afluencia. El trabajo se divide en cuatro partes:

- **Parte I. diseñar un modelo que sea capaz de predecir en un determinado día el número de pacientes que demandan atención médica.**

  Todos hemos padecido en algún momento colas en los centros de salud, esto implican grandes dificultades para que los profesionales médicos puedan desarrollar su trabajo y asistir a los pacientes de una manera correcta. Este estudio se ha particularizado en los centros de Jaén. Aquí el uso de los servicios de salud es muy variable a lo largo del año. El objetivo es crear un modelo que sea capaz de predecir con antelación el número de pacientes que van a visitar el centro de salud. Esto será útil para llevar a cabo una gestión óptima de los recursos médicos haciendo así que el sistema de salud sea más sostenible y eficiente.

  Este estudio se basa en el número de pacientes que han acudido a los centros de atención primaria de Jaén entre los años 2007 y 2011. Los modelos predictivos se generan usando los datos desde año 2007 al 2010, y los modelos son validados con los datos de año 2011. Es un estudio supervisado. En este estudio se han usado técnicas de minería de datos. El algoritmo MDL se utiliza para analizar la relación de cada atributo con el atributo objetivo y en la generación de los modelos se utilizan los algoritmos de regresión: SVM (con kernel (Gaussiano y lineal) y GLM.

  Como resultado hemos obtenido un modelo predictivo con un error absoluto de 2,29% en la predicción de pacientes del año 2011. En los días Laborables el algoritmo GLM es el más eficiente, sin embargo para las

vacaciones, el algoritmo con el error más bajo es la SVM con Kernel lineal.

- **Parte II. Determinar las variables espaciales locales que influyen en la afluencia de pacientes a los centros de salud**

  El objetivo principal en esta parte del trabajo es determinar los factores espaciales que afectan a la afluencia de pacientes en un determinado centro de salud. Para ello analizamos datos locales de cada Centro de Salud: como el nivel económico y el tipo de población que atiende (pediátrica, geriátrica, etc.).

  Se encontró que el número de visitas al médico en un centro de salud se relaciona con variables como el nivel económico de la zona y sobre todo con el tipo de población atendida por el centro de salud, la población pediátrica es la que más influye en la afluencia de pacientes.

- **Parte III. Diseño de un nuevo sistema de cita médica para optimizar los sistemas de salud.**

  Una gestión óptima de los recursos en los centros de salud implica el uso de unas agendas apropiados para programar las citas. Los horarios de los centros de salud se suelen dividir en intervalos de tiempo cuya duración es igual al tiempo necesario para la asistencia clínica (5 minutos en el caso de Jaén). Sin embargo los médicos realizan una serie de tareas que no siempre son de naturaleza clínica: la emisión de recetas o la emisión de certificados médicos. En este sentido, el tiempo empleado en asistir una demanda clínica o una demanda administrativa es diferente. Esta última requiere menos tiempo para asistir a la paciente. Esta parte del estudio se centra en la tarea administrativa, para lo cual creamos un modelo predictivo que proporcione información diaria sobre el número de pacientes que irán al centro de salud para una cuestión administrativa. El error absoluto del modelo es inferior a 4,6%, este dato del error es interesante si tenemos en cuenta que de unas épocas del año a otras la demanda administrativa varia hasta un 350%. Finalmente proponemos un cambio en el sistema de agendas que puede suponer un ahorro de tiempo del 21,73%.

- **Parte IV. Desarrollo de un sistema experto que implementan los modelos generados en este estudio, que permita la predecir los pacientes que demandaran una atención médica.**

  Hay una multitud de sistemas expertos que los líderes de la organización utilizan para tomar mejores decisiones. El objetivo principal en esta parte del trabajo es el Diseño y Desarrollo de este sistema experto, que implemente los algoritmos desarrollados en puntos descritos anteriormente. Por otra parte, queremos demostrar que estos sistemas pueden ser desarrollados rápidamente con las herramientas que se encuentran actualmente en el mercado. El sistema experto será una herramienta que proporcione información clave para los administradores de recursos de los centros de salud.

ABSTRACT

Nowadays information systems are essential in the health realm, both from the point of view of resource management, and from a purely health perspective. In this paper we advance in the optimization of healthcare resources based on data mining and obtaining spatial attributes of health centers, the main objective is to create models to predict the influx of patients to the health centers and determine the variables affecting the inflows. Our work is divided into four parts:

- **Part I design a model that is able to predict the number of patients who need medical attention.**

  Patient delays and long waits in primary health care imply great difficulties for medical staff to work and attend patients in a correctly manner. This study has been particularized in centers of a city of southern Spain, Jaen. Here the use of health services is highly variable because there are many influential factors. The objective is to create a model to predict in advance the number of patients who are going to visit the care center. This will be useful for carrying out an optimal clinic resource management and making the health system more sustainable.

  This study is based on the number of patients who have come to the primary health care centers of Jaen from 2007 to 2011. To generate the predictive data model from 2007 to 2010 are used, and the validation model is performed with the year 2011. It is a supervised study. To generate these predictive models techniques of data mining have been used. The MDL algorithm is used to analyze the relation of each attribute considered with the attribute target. The algorithms SVM (with Gaussian and Linear kernel) and GLM are applied. Finally, this prediction is compared with the real data and the mean error by each of the algorithms is calculated.

  We have obtained a predictive model with absolute error of 2.29%. In the prediction of the weekdays, the GLM algorithm is the most efficient, however for holidays, the algorithm with the lowest error is the SVM with Linear kernel.

- **Part II. determine the local spatial variables that influence the influx of patients to health centers**

  The main objective in this part of the study is to determine the spatial factors that affect the number of patients in a particular clinic. We analyzed data from each local health center: as the economic level and the type of population served (pediatric, geriatric, etc.).

  It was found that the number of visits to the doctor in a health center is related to variables such as the economic status of the area and especially with the type of population served by the health center.

- **Part III. Design of a new appointment system medical to optimize healthcare systems.**

  An optimal resource management in health care centers implies the use of an appropriate timetabling scheme to schedule appointments. Timetables of health centers are usually divided into time slots whose duration is equal to time required for clinical attendance. However doctors perform a series of tasks that are not always clinical in nature: issuing prescriptions or prescribing sick leave certificates. In this sense the time spent in attending a clinical or an administrative matter is different. This last required less time to attend the patient. This study is focused in the administrative task. A predictive model is generated to provide daily information on how many patients will go to the health center for an administrative issue. The accuracy of the model is less than 4,6% absolute error and the improvement in scheduling appointments is a time saving of 21,73%.

- **Part IV. Rapid development of an expert system**

  There are a multitude of expert systems that organisational leaders use to make better decisions. The main objective of this paper is to Desing and Develop this expert system, basing it on the previous study. Furthermore, we want to demonstrate that such systems can quickly be developed with the tools that are currently on the market. The expert system will be a tool that provides key information to healthcare centre resource managers so that they can correctly design their services.

INDICE:

INDICE DE FIGURAS :

INDICE DE TABLAS :

1. INTRODUCCIÓN

Hoy en día hay números sistemas de información que contribuyen de forma decisiva a mejorar los sistemas de Salud, tanto desde el punto de vista de gestión de recursos, como desde el punto de vista puramente sanitario. En Andalucía contamos con dos de los mayores sistemas de información sanitarios del mundo: Diraya [1] y Salud Responde.

Diraya se utiliza en el sistema sanitario público de Andalucía dando soporte a la historia clínica electrónica. Integra toda la información de salud de cada ciudadano, para que esté disponible en el lugar y momento en que sea necesario, y sirve también para la gestión del sistema sanitario. A continuación podemos ver algunas cifras significativas del sistema:

- Implantado en 735 centros de Atención Primaria y 27 Áreas Hospitalarias.
- Cubre a 7.687.399 ciudadanos.
- Más de 200.000 episodios de urgencias recogidos mensualmente en el sistema.
- Más de 6 millones de prescripciones mensuales y 8 millones de dispensaciones de receta electrónica
- Más de 300 millones de citas gestionadas, mensualmente más de 7 millones entre atención primaria y especializada a través de los diferentes canales: presencial, telefónico, internet y SMS.

Esto es un claro ejemplo de cómo los sistemas de información son claves en las organizaciones sanitarias, no solo desde el punto de vista clínico sino también desde el punto de vista económico. Diraya contribuye a mejorar el sistema desde varios puntos de vista:

- Establecer reglas de negocio en prescripción electrónica orientadas al control del gasto farmacéutico.
- Reducción de consultas de atención primaria gracias a la implantación de la receta electrónica.
- Ahorro debido a la relación con el ciudadano a través de los canales telefónicos e Internet.

- Reducción de pruebas y exámenes duplicados.
- Ahorro en formatos papel (adquisición, almacenamiento, destrucción).
- Reducción del fraude (suplantación de identidad con objeto de obtener prestaciones).

También desde el punto de vista puramente clínico Diraya es clave ya que dispone de un modelo de historia clínica que permite la **consulta y la anotación de datos en todos los dispositivos y niveles asistenciales**: atención primaria, atención especializada, urgencias y hospitalización. Gracias a este sistema los profesionales sanitarios que asisten a un mismo paciente, tienen acceso a la información clínica en cualquier centro sanitario de la geografía andaluza.

Otro claro ejemplo de éxito en Andalucía es Salud Responde. Dispone de un Centro de atención al ciudadano para temas relacionados con la Salud, CRM Sanitario (Customer Relationship Management) [2,3,4]. Actualmente cuenta con una amplia cartera de servicios como: Consejos sanitarios, Seguimiento de cuidados paliativos, Cita previa, Información sanitaria, etc. Un claro ejemplo de cómo estos sistemas son claves tanto para el paciente como para la administración pública lo podemos encontrar en el papel que jugó Salud Responde durante la pandemia de la Gripe A en el año 2009 y 2010, contribuyendo a mejorar la atención del ciudadano y minimizando el gasto de la atención sanitaria. El principal objetivo era evitar la presencia de personas con sintomatología compatible con la gripe A cuya afectación era leve en los centros sanitarios. Esta mejora para el enfermo se centraba en varios aspectos:

- informar al ciudadano del alcance de la Gripe A y darle consejos para su prevención.
- En coordinación con los centros de Atención Primaria y Emergencias Sanitarias se establecieron unos protocolos para valorar a los pacientes mediante un triaje telefónico, derivando a los pacientes a los sistemas sanitarios en caso de gravedad, o recomendarles su permanencia en casa a los pacientes que no revestían gravedad. Para clasificar la gravedad de los pacientes se uso un triaje [5,6,].

Lo más importante por parte de Salud Responde era establecer y coordinar los protocolos de actuación de los participantes en el dispositivo: operadores de Salud Responde, enfermeros de Salud Responde, médicos de Salud Responde, servicios de urgencias, servicios de emergencias y médicos de atención primaria. Su función en el control de la pandemia era la siguiente:

- Proporcionar todo tipo de información de la Gripe A. síntomas, consejos para evitar contagios, etc.
- Asistencia de los pacientes que llamaban a Salud Responde.
- Seguimiento de los pacientes que una vez clasificadas su gravedad.
- Prescripción en aquellos pacientes que lo requerían.

La atención al paciente se basó en un triaje, éste tenía en cuenta: síntomas, complicaciones y factores de riesgo. El triaje estuvo estructurado en una entrevista sobre el CRM de Salud Responde. Se establecieron los siguientes estados:

- Sintomatología inherente a un Síndrome Gripal no complicado: tos, rinorrea, fiebre mayor de 37,5ºC, dolor de garganta, diarrea, vómitos, cefalea, mialgias, malestar general.

- Existencia de complicaciones moderadas: Empeoramiento del estado previo sin complicaciones graves, fiebre persistente de más de 3 días que no cede con antitérmicos, dolor torácico de características pleuríticas, incapacidad para la ingesta oral, descompensación no grave de patología previa.

- Existencia de complicaciones graves: disnea, afectación hemodinámica, disminución del nivel de conciencia, descompensación grave de patología previa.

- Existencia de Factores de Riesgo: edad de 65 años o mayor, Embarazo, Enfermedad, cardiovascular previa (excepto hipertensión), enfermedad pulmonar crónica previa (incluye

EPOC, fibrosis quística y asma), enfermedades metabólicas (incluye diabetes mellitus y obesidad), insuficiencia renal crónica, anemias, hemoglobinopatías, hepatopatía crónica, convive en residencias o centros con pacientes crónicos de cualquier edad, hepatopatía crónica, menores de 18 años en tratamiento prolongado con ácido acetil salicílico, inmunodeficiencia o inmunosupresión, enfermedades neuromusculares graves, Asplenia.

Según el resultado de la entrevista, el sistema informático de Salud Responde proporcionaba el nivel de gravedad:

- **Nivel Verde.** Síntomas sin existencia de Factores de Riesgo ni de complicaciones. Recomendaciones:

  1. higiénicas sanitarias y dosificación de antitérmicos.

- **Nivel Amarillo.** Presencia de complicaciones moderadas sin existencia de factores de riesgo o síntomas sin complicaciones con existencia de Factores de Riesgo. Recomendaciones:

  1. higiénicas sanitarias y dosificación de antitérmicos.
  2. Revisión proactiva a las 24 horas y nuevo triaje.

- **Nivel Naranja.** Presencia de complicaciones moderadas con existencia de factores de riesgo y/o incapacidad para la ingesta oral y/o expectoración hemoptoica:

  1. Recomendaciones higiénicas sanitarias y dosificación de antitérmicos.
  2. Derivación a Atención Primaria.
  3. Revisión proactiva a las 12 horas o hasta ser vista por Atención Primaria y nuevo triaje.

- **Nivel Rojo.** Presencia de complicaciones graves. Recomendaciones:

1. Derivación a Centro Coordinador de Urgencias de su provincia para asistencia.

Los resultados de este trabajo fueron los siguientes. El dispositivo de Salud Responde estuvo operativo desde Octubre de 2009 hasta Junio de 2010, en estos 9 meses Salud Responde atendió a 56.497 pacientes, de los cuales 32.854 fueron atendidos y sus solicitudes resueltas por los teleoperadores de Salud Responde.

De este trabajo, 23.643 pacientes fueron derivados por los operadores de Salud Responde a los Enfermeros de Salud Responde. Los enfermeros de Salud Responde atendían los casos más graves. De todos los casos atendidos por ellos, resolvieron 15.433 casos. De los casos más graves, fueron derivados a urgencias 3.260 pacientes y a su médico de Cabecera fueron derivados 4.550 pacientes. Por tanto de los 56.497 pacientes atendidos por Salud Responde, un total de 48.287 casos que no requirieron atención médica y se evito el desplazamiento a su Centro de Salud o Urgencias del Hospital. Esto supuso que el 85,46% de los usuarios que requirieron atención a Salud Responde fueron resueltos directamente por el Centro.

Los CRM Sanitarios pueden jugar un papel clave en las pandemias y las alertas sanitarias. Desde ellos se pueden informar al usuario contribuyendo así a tranquilizar a la sociedad en caso de alarma social y con la colaboración de otras instituciones médicas a establecer un filtro de pacientes para ayudar a que se haga un uso más eficiente de los recursos sanitarios [4].

Estos dos casos de existo que hemos presentado en la introducción de este trabajo, son actualmente en Sanidad los dos sistemas de Información más importantes en Andalucía y como hemos visto este tipo de sistemas contribuyen enormemente a mejorar los sistemas de salud en distintas vertientes.

Nuestro trabajo pretende continuar avanzando en este campo, sobre todo en la mejora de los sistemas basado en el conocimiento existente y oculto en los datos almacenados. Tanto Diraya como el sistema de Salud Responde cuenta con información fiable desde el año 2003 aproximadamente y a toda esta información se le puede aplicar técnicas de

minería de datos para extraer el conocimiento oculto y plantear mejoras en el ámbito Sanitario.

Hoy en día la optimización de los recursos es esencial en cualquier campo y organización. El caso de la gestión de los recursos en los centros de atención primaria es particularmente interesante y crítico [8], ya que la mejora de la gestión repercute directamente en un incremento de la satisfacción del paciente y también reduce los riesgos de salud de los pacientes [9].

Como hemos comentado anteriormente en Diraya se almacenan todos datos de los usuarios que son atendidos en un Centro de Salud y el motivo por el que ha sido atendido, esto nos permite poder explotar esta información para buscar patrones y relacionarlos con factores ambientales, meteorológicos, económicos, etc.

Partimos de la Hipótesis de que se pueden relacionar los factores ambientales y meteorológicos de una zona geográfica con ciertas patologías padecidas por los usuarios y por tanto existe una relación entre estos factores y el número de pacientes que van diariamente a los Centros sanitarios a buscar una demanda clínica o una demanda administrativa (receta, parte de baja, etc.).

El estudio de este trabajo se centra en los Centros de Salud de Jaén: Belén, La Magdalena, El Valle, Federico Castillo, San Felipe, Virgen de la Capilla y Fuentezuelas, es decir todos los centros de Salud de Jaén capital. El periodo del estudio es desde 2007 a 2011 (ambos inclusive).

En la primera parte de la tesis, se creará una Base de Datos que contendrá de forma estructurada y normalizada la información de asistencia a los centros de salud, información meteorológica de las estaciones de Jaén, calidad del aire y datos espaciales locales a cada centro de Salud, como nivel económico y tipo de población atendida por el centro de salud (pediátrica, geriátrica, etc.). Los factores ambientales que se usaran en el estudio son: Temperatura Media, Temperatura Máxima, Temperatura Mínima, Precipitaciones, Humedad Relativa y la calidad del aire. La calidad del aire se calcula en fusión de los siguientes contaminantes: Dióxido de Azufre, Ozono, Monóxido de Carbono, Dióxido de

Nitrógeno, Benceno, Sulfuro de Hidrogeno, Partículas en Suspensión, Acido Sulfhídrico, Partículas de Polen.

En la segunda parte de este trabajo se probarán varios algoritmos de Minería de datos para determinar el que mejor se adapta a la naturaleza de nuestro estudio y se creara unos modelos predictivo que sea capaces de predecir el número de usuarios que requerirán una atención sanitaria y qué tipo de atención solicitarán (atención médica o administrativa).

Finalmente, se desarrollara un prototipo de sistema experto que implementará los modelos desarrollados en este trabajo. El objetivo de esta parte de la tesis es demostrar que este tipo de trabajos se pueden llevar a la realidad sin realizar un gran esfuerzo de desarrollo, evitando que este tipo de trabajos se quede simplemente en un trabajo meramente teórico.

1.1. Justificación de la Investigación

La puntualidad en la atención al ciudadano es una característica fundamental en la prestación de los servicio de Salud. Algunos autores hacen hincapié en la importancia del tiempo y la eficiencia de los recursos en la medicina [10]. Desde el punto de vista del paciente, un ejemplo de la falta de eficiencia es la dificultad para obtener una cita para que sea visto por un médico, sobre todo en determinadas épocas del año. Este es un problema generalizado en la atención primaria de salud. Los retrasos de los pacientes y las largas esperas son uno de los principales problemas de los profesionales médicos y del resto del personal sanitario y administrativo. Esta situación implica grandes dificultades para trabajar y asistir a los pacientes de una manera correcta. Ya en 1996, la obra de Starfield [11] indica que con una gestión adecuada de los recursos en atención primaria se obtiene mayores puntuaciones de satisfacción de los pacientes y menor gasto sanitarios, reduciéndose además la prescripción farmacológica y reduciéndose el riesgo de salud del paciente al ser atendido antes.

Normalmente el dimensionamiento de los recursos clínicos se estima a partir de los promedios de datos generalizados de años anteriores. Esto lleva a una atención satisfactoria para los pacientes, pero a veces, si hay una entrada masiva de pacientes al centro Salud, tiene un impacto

negativo en la calidad de la atención al paciente. El centro de salud debe tener en cada momento dimensionado los recursos en función de la demanda de pacientes. En este sentido, sabiendo de antemano el número de pacientes que va a ir al centro de salud supondría una información fundamental para poder gestionar adecuadamente los recursos técnicos y humanos necesarios para ofrecer una sanidad de calidad.

A continuación en la figura 1 se muestra datos de asistencia a los centros de salud de Jaén de atención primaria durante el año 2011. Como puede verse hay una gran variabilidad en función de la estación del año, del tipo de día (Laborable o festivo) y día de la semana.

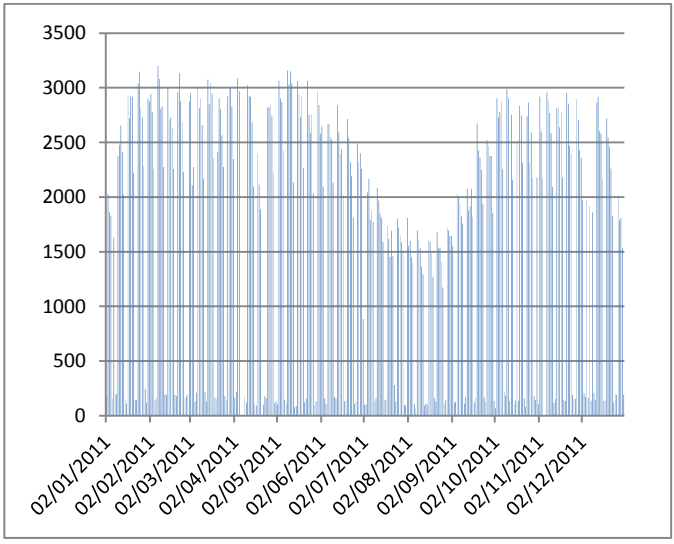


Figura 1. Numero de vistas de pacientes a los centros de salud de Jaén durante el año 2011.

Si analizamos el número de visitas a los centros de Salud de Jaén por día de la semana (figura 2), puede observarse cómo lunes y martes son los días de más actividad para bajar los miércoles, jueves y viernes. Si

comparamos los datos de un martes con un viernes, la demanda cae de 622.319 a 489.391, esto supone una diferencia del 21,36%.

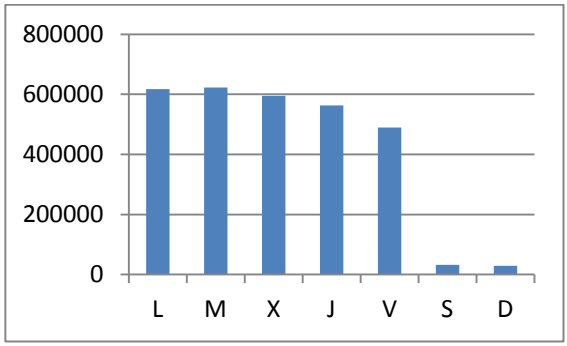


Figura 2. Número de visitas a los centros de salud de Jaén en el año 2011 por día de la semana.

Por otro lado si analizamos los datos de visitas al médico agrupados por meses (figura 3), también podemos encontrar importantes diferencias de unos meses a otros.

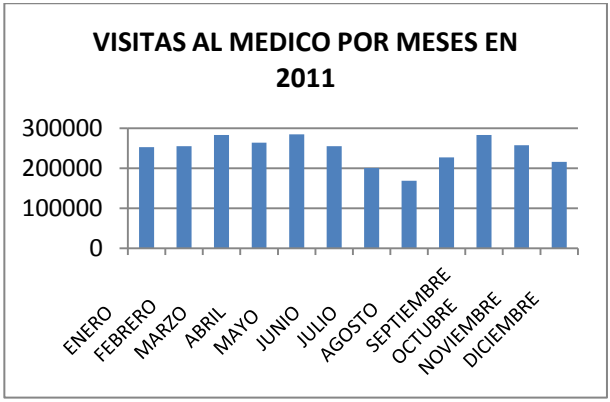


Figura 3. Número de visitas a los centros de salud de Jaén durante el año 2011 agrupada por meses.

En Mayo tenemos un mes de máxima actividad mientras que Agosto como era de esperar es el mes con menos visitas, si comparamos estos meses (Mayo y Agosto) la demanda baja un 40,7%. Este comportamiento puede ser previsible debido al periodo estival en agosto, pero hay otras diferencias que no son tan evidentes. Si comparamos dos meses como pueden ser Octubre y Noviembre, podemos observar que hay también importantes diferencias, la demanda baja en Noviembre un 9% con respecto a Octubre.

Otro factor a tener en cuenta son los días festivos, estos días como es de esperar baja la demanda ya que solo se atienden las urgencias. Pero también encontramos importantes diferencias en función de la estación del año. En la figura 4 podemos ver el número de pacientes que asistieron al centro de salud durante los festivos de 2011.

Figura 4. Afluencia de pacientes en días festivos a los centros de salud de Jaén durante el año 2011 agrupado por estaciones.

Como puede observarse en la figura 4 se duplica la demanda el lunes 26 de diciembre que fue Festivo en Jaén, con respecto al viernes 22 de Marzo.

Con toda esta variación de datos y teniendo en cuenta que detrás de esta demanda asistencial de los ciudadanos está un equipo técnico y humano de los Centros de Salud, es muy importante poder predecir con antelación la demanda que tendremos para poder dimensionar correctamente dichos servicios y poder hacer que la sanidad sea más eficaz, eficiente y sostenible.

Otra cuestión importante a tener en cuenta, si analizamos la tabla 1 donde se relaciona el porcentaje de adscripciones de pacientes y el porcentaje de visitas reales de los distintos centros de Salud de Jaén, observaremos que en la mayoría de los centros no coinciden estos porcentajes. El caso más llamativo es el del centro de Salud Virgen de la Capilla donde hay adscritos más de un 18% de la población de Jaén y sin embargo sólo recibe el 16% de visitas de los pacientes de Jaén, esto es indicativo de que hay factores locales al centro de Salud que influyen en que los pacientes asistan con más o menos frecuencia a dichos centros de Salud.

CENTRO DE SALUD	% DE ADSCRIPCIONES	% de visitas totales
Virgen de la Capilla	18,55%	16,40%
Belén	8,17%	8,16%
San Felipe	17,49%	17,84%
La Magdalena	11,11%	11,33%
Fuentezuelas	7,30%	7,32%
El Valle	14,68%	15,55%
Federico Castillo	22,69%	23,41%

Tabla 1. Pacientes adscritos a los centros de salud y porcentaje de visitas reales durante el año 2011.

Otro gran problema que tratamos de mejorar con este trabajo es que para atender adecuadamente a un paciente en un centro de salud se requiere que las agendas estén en equilibrio entre número de pacientes que necesita una cita y los huecos que ofertan los centros de salud. Esto implica utilizar unas configuraciones correctas de agenda que garanticen un buen

servicio. Como se ha indicado anteriormente en este sentido hay estudios que analizan la programación de citas médicas óptimas [12,13,14] y los factores que afectan al tiempo de espera en las consultas médicas [15]. Hay que tener en cuenta que en los Centros de Atención Primaria los médicos también realizan una serie de tareas que no siempre son de naturaleza clínica. A veces los pacientes acuden al centro para solicitar un servicio administrativo. En Jaén en 2011 casi el 33,6% de las visitas al médico fueron por la emisión de una receta repetitiva o un certificado médico. El tiempo requerido para una demanda administrativa es muy inferior que el tiempo necesario para una demanda clínica. Normalmente las agendas de citas de los centros de salud se dividen en intervalos regulares de tiempo cuya duración es igual a la demanda clínica, en Jaén es de 5 minutos, que es el tiempo que por lo general requiere un médico para asistir a un paciente. Después de consultar con varios profesionales de los centros de Salud de Jaén, nos indicaron que el tiempo necesario para una demanda administrativa es inferior que el de una demanda clínica, por ejemplo, un médico sólo necesita 1 minuto para renovar una receta ya que en Andalucía la receta electrónica [1] está integrada con Diraya y un médico tiene que hacer muy pocos "clics" para prescribir o renovar una receta. Por el contrario un certificado médico requiere 3 minutos, ya que el profesional tiene que escribir parte del informe e imprimirlo, sin embargo para estas tareas actualmente se reservan huecos de 5 minutos. Esta situación produce desequilibrios importantes en el sistema de citas del centro de salud. Un sistema de citas eficaz seria aquel que nos permitiera conocer de antemano los usuarios que asistirán a nuestro centro de Salud y el tipo de demanda que nos solicitara, esto nos permitiría generar unas agendas optimas que reducirían el tiempo de espera de los pacientes y optimizara los sistemas sanitarios.

Con la problemática comentada anteriormente otro de los objetivos de este estudio es diseñar un sistema que ofrezca información diaria sobre el número de pacientes que acudirán al centro de salud para una cuestión administrativa y cuántos para un problema de salud. Para ellos se aplicaran técnicas de minería de datos con el fin de predecir con exactitud dicho número de pacientes, permitiendo así crear unas agendas optimas con unas asignación de tiempo correcta, con ello se puede conseguir una mejora significativa en la programación de citas y, por tanto, una optimización adecuada de los recursos en los centros de atención primaria. En la figura 5 podemos ver gráficamente los tiempos que pueden ser optimizados actualmente.

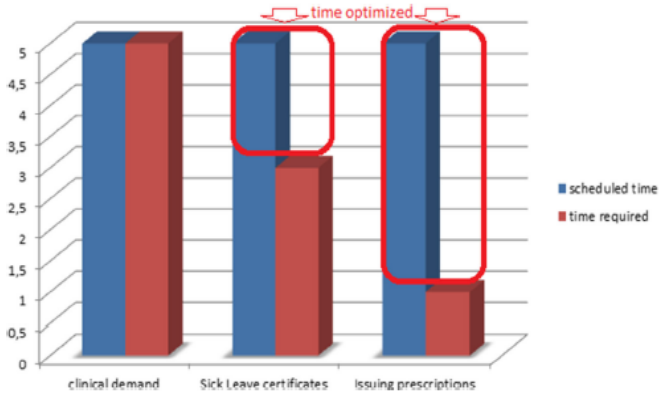


Figura 5. Comparativa de los tiempos programados y requeridos por tarea.

1.2. Objetivos de la investigación y estado del arte

Parte de los objetivos de este trabajo ya han sido comentados anteriormente, pero en este punto vamos a hacer una descripción más profunda.

En la primera fase del estudio nuestro objetivo es determinar qué factores externos tienen influencia en la asistencia de pacientes a los centros de salud. En otras palabras, ¿qué factores influyen en las patologías más comunes que se tratan en los centros de salud de atención primaria? Respecto a este tema, hay muchos estudios que llevan a cabo una revisión sistemática de los factores influyentes en las enfermedades. Por ejemplo, en Dawson et al. 2007 [16] se analiza la influencia de los factores meteorológicos sobre la incidencia de accidente cerebrovascular. Oiamo et al. 2011 [17] llega a la conclusión de que algunos niveles de exposición a la contaminación parecen influir en la utilización de servicios de atención de salud. Según Donaldson et al. 2012 [18] ciertas estaciones tienen influencia de exacerbaciones en pacientes con EPOC. Otras investigaciones también están trabajando en la misma dirección, Ferrari et

al. 2012 [19] o Tseng et al 2013 [20] son otros ejemplos. La búsqueda de una relación entre los factores externos y la aparición de determinadas enfermedades es generalizado en toda la comunidad científica. Todos estos estudios tienen como objetivo la posibilidad de ser capaz de anticipar la visita del paciente al médico y también llevar a cabo una mejor gestión de los recursos médicos. Sin embargo las variables estacionales, ambientales y meteorológicas influyen en las enfermedades de forma diferente según la zona geográfica considerada. Los primeros estudios como Keatinge et al. 1997 [21] enfocado en este sentido concluyen que estos estudios son aplicables a la zona sobre la que se realizó el análisis. Dawson et al. 2008 [22] sitúa su investigación en Escocia, donde se analizan las asociaciones entre las variables meteorológicas y de los ingresos hospitalarios con accidente cerebrovascular agudo. Anteriormente, en 1996, Rothwellet al. [23] ya había buscado la relación de incidencia de accidente cerebrovascular a la estación del año o a la temperatura. Llegaron a la conclusión de que se necesitan más estudios para determinar estas relaciones para ciertas zonas.

Otro objetivo de nuestro trabajo es proponer un cambio en el modelo de agenda de citación de los centros de salud para que sean más eficiente, ya que como se ha visto en la introducción actualmente hay tareas a las que se destina más tiempo del necesario, para ello estudiaremos realizar una separación de tareas en las agendas. Para proponer este cambio es fundamental disponer de un modelo predictivo ya que en Jaén en 2011, hubo una gran variación en la demanda administrativa de un mes a otro. La demanda de 29 de julio fue de 341 pacientes, mientras que el 22 de noviembre esta demanda se elevó a 1.190 pacientes, una variación de casi el 350%. Un mal diseño de este modelo podría provoca desequilibrios y permanecer desocupados huecos de agenda en el horario destinado a la demanda administrativo mientras que en la demanda clínica no hay huecos suficientes, o viceversa. Es por eso que necesitamos un modelo fiable que sea capaz de predecir con exactitud la demanda de servicios administrativos.

A continuación se presenta una reseña de los trabajos de investigación previos en que se aplicaron diferentes técnicas para alcanzar la optimización en la programación de citas. Desde 2003-2009 se han usado técnicas de simulación y heurística, y en 2010 se usaron técnica mixta de simulación y la minería de datos. Tabla 3 muestra un resumen de las técnicas utilizadas en trabajos anteriores.

año	Titulo	técnica
2003	Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach[24]	Simulación
2003	A Constraint Programming Application to Staff Scheduling in Health Care[25]	Heurística
2006	Designing appointment scheduling systems for ambulatory care services[26]	Simulación
2007	Optimal outpatient appointment scheduling[27]	Simulación
2009	A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling[28]	Simulación
2010	Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations[29]	Minería de Datos y Simulación

Tabla 2. Trabajos previos para la optimización de citas médicas.

Las principales diferencias entre nuestro trabajo y las contribuciones de los trabajos de investigación mencionados anteriormente es que en estos trabajos anteriores no se utiliza una combinación de datos históricos de asistencia, datos ambientales y meteorológicos. Además, no se usan técnicas puras de minería de datos como en nuestro trabajo donde se realiza un análisis, utilizando técnicas y algoritmos para detectar anomalías (descartando los datos erróneos que podrían afectar el modelo), algoritmos de importancia de atributos para buscar los atributos con relación con el atributo target y, finalmente, se genera varios modelos con diferentes algoritmos de regresión con el fin de conseguir el modelo más eficiente.

Otro objetivo importante es buscar los factores locales que influyen en que los pacientes acudan con mayor o menor frecuencia a los centros de salud. Como hemos visto en el punto anterior en la mayoría de los centros de salud no coinciden las personas adscritas al centro con el número de visitas. En la literatura encontramos varios artículos donde se indica el impacto de la edad en los tratamientos y en el riesgo de padecer ciertas

patologías [30-33]. Con estos datos podemos suponer que uno de los factores más importantes para que un paciente vaya más o menos a su centro de salud depende de su edad, precisamente por la probabilidad de padecer ciertas patologías, ir a por recetas, controles de salud, etc.

Otra variable que vamos a considerar en nuestro estudio es el nivel económico de la zona, hasta ahora no hemos encontrado esta relación en la literatura, pero en nuestro estudio partimos de la hipótesis de que a mayor nivel económico es posible que los pacientes dispongan de mayor numero de seguros privados y esto influyan en las visitas a los centros de salud públicos.

Finalmente como último objetivo de este trabajo se desarrollará un sistema experto que implemente los algoritmos generados. Hoy en día, hay multitud de sistemas expertos que los líderes de la organización utilizan para tomar las mejores decisiones. Estos sistemas se utilizan en numerosos sectores, y muchas de estas aplicaciones se han documentado en la literatura [34-37]. El motivo de realizar este desarrollo es evitar que un estudio o una investigación que supone un importante esfuerzo, se quede en un mero trabajo teórico ya que para generar los modelo predictivos, se suelen usar distintas herramientas de software libre como R, Weka, MySQL, etc. [38-40], por lo tanto, el uso del sistema es complejo y limitado a personal altamente especializado. El objetivo en esta parte del trabajo es aislar al usuario final de la complejidad del uso de estos sistemas. Sin embargo el desarrollo de este tipo de herramientas suele ser complejo y costoso, por un lado hay que desarrollar la parte de Base de Datos, carga de datos, implementación de los algoritmos de minería de datos y finalmente una aplicación que gestione datos y algoritmos. Cada de vez es más común que las Bases de Datos integren módulos de Minería de Datos y módulos de desarrollo [41]. Este tipo de herramientas facilita enormemente el desarrollo de este tipo de productos, permitiendo crear un sistema experto complejo en pocas jornadas de trabajo y sin necesidad de tener un alto conocimiento en Minería de Datos, Base de Datos o Desarrollo de aplicaciones.

1.3. Aportaciones

Como hemos visto en el punto anterior donde se desarrollan los objetivos y se hace un repaso del estado del arte, podemos observar cómo la mayoría de las investigaciones que hay actualmente buscan la relación entre algunos factores ambientales, meteorológicos y temporales con ciertas patologías concretas, nuestro trabajo busca establecer una relación global entre estos factores para poder predecir el número de pacientes que necesitarán una atención médica.

Por otro lado y basado en la predicción de pacientes que asistirán al centro de salud y el motivo de la visita, propondremos una optimización del sistema sanitario a través de un cambio en la configuración del sistema agendas de los centros de salud. El punto anterior hemos realizado un repaso de la actual configuración de las agendas y hemos visto que desde un punto de vista teórico no tienen la configuración mas optima ya que se gastan 5 minutos en actividades que teóricamente requieren menos tiempo. Al repasar el estado del arte vemos como muchos autores han trabajado en la optimización de las agendas, pero ellos usan técnicas distintas a las usadas en este trabajo como: simulación, heurística, etc. y lo hacen para mejorar agendas concretas. En nuestro trabajo hacemos un análisis global y buscamos la optimización global del sistema sanitario proponiendo un nuevo modelo de agendas, donde tendremos una separación física de huecos de demanda clínica, recetas y parte médico.

Otra contribución importante de nuestro trabajo es determinar los factores locales que afectan a cada centro de salud y que provocan que acudan más o menos pacientes a los centros de salud. Se ha comentado anteriormente como no coinciden en Jaén la población adscrita con el número de visitas reales. En la literatura si hay documentada como la edad influyen en ciertas patologías y esto puede influir en que un centro de salud a pesar de tener menos usuarios adscritos tenga que atender más pacientes que otro con los mismos pacientes adscritos. En nuestro trabajo vamos a analizar cómo influye este factor y vamos a cuantificar qué edades son las más influyentes en la suma o resta pacientes a los centros de salud.

Por otro lado en nuestro trabajo no solo trabajaremos con la edad como factor local del centro de salud, vamos a analizar otro factor que puede ser clave como es el nivel económico de la zona, en este sentido no hay nada

actualmente descrito en la literatura. En nuestro trabajo analizaremos este factor y determinaremos si afecta al número de pacientes que asisten a los centros de salud y en qué medida lo hace.

Finalmente pretendemos demostrar que a pesar de que este tipo Sistemas Expertos requieren un importante esfuerzo para ser desarrollados, hoy en día existen varias bases de datos del mercado que integran todas las herramientas necesarias para desarrollar este trabajo en pocas horas y sin necesidad de ser un experto en minería de datos ni en desarrollo de aplicaciones. Es esencial que este tipo de trabajos se terminen materializando en herramientas explotables por los gestores sanitarios para que no se queden como meros trabajos teóricos. En la literatura no hemos encontrado información de este tipo de desarrollos para sistemas expertos.

2. METODOLOGÍA

En nuestro estudio usaremos distintas técnicas de minería de datos para generar los modelos predictivos. La minería de datos es un área de estudio que surge de la convergencia de otras disciplinas: Ciencias de la Computación, Estadística, Inteligencia Artificial, Tecnología de Bases de Datos y Reconocimiento de Patrones, entre otras. Comprende el análisis de grandes conjuntos de datos y la búsqueda de relaciones entre variables, a través de métodos computacionalmente intensivos. Muchas veces se encuentran relaciones o coincidencia no esperadas y, por lo general, los métodos involucran el análisis de enormes cantidades de datos multidimensionales, por tanto, la minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque el alto número de información.

La generación de un modelo de minería de datos incluye desde la formulación de preguntas acerca de los datos, la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. Definir el problema
2. Preparar los datos
3. Explorar los datos
4. Generar los modelos
5. Explorar y validar los modelos
6. Implementar y actualizar los modelos

El la Figura 6 mostramos un diagrama donde se describe las relaciones existentes entre cada paso del proceso de Minería de Datos.

Figura 6. Ciclo de vida de un proyecto de minería de datos.

El proceso que se muestra en la figura 6 es cíclico, lo que significa que la creación de un modelo de minería de datos es un proceso dinámico e iterativo.

En nuestro estudio buscamos el número de pacientes que necesitan una atención en su centro de salud, esta es una variable desconocida, pero que puede ser deducido a partir de variables conocidas y que en teoría están relacionadas de forma directa o inversa. Por tanto nuestro estudio es un análisis supervisado.

El sistema ha sido desarrollado con Oracle data mining [42].Oracle Data Mining ofrece una potente funcionalidad de minería de datos como funciones nativas de SQL dentro de la base de datos. Directamente en la base de datos se pueden generar modelos predictivos, su interfaz es una extensión de SQL Developer, que permite explorar los datos, construir y evaluar los modelos, aplicarlos a nuevos datos, guardar y compartir sus metodologías analíticas. Esta herramienta permite a los investigadores y desarrolladores de aplicaciones poder utilizar su API de SQL para crear aplicaciones de última generación de forma rápida. Debido a que los datos que utilizan los modelos y resultados permanecen dentro de la base

de datos, se eliminan los tiempos de latencia de envío de información de unos sistemas a otros. Además, los modelos de minería de datos generados con Oracle Data Mining se pueden incluir en las consultas SQL y embebidos en las aplicaciones para ofrecer una mejor inteligencia de negocios.

Oracle Data Mining ofrece una colección de algoritmos de minería de datos que resuelven una amplia gama de problemas. La mayoría de los algoritmos de data mining pueden separarse en técnicas de data mining con "aprendizaje supervisado" y "aprendizaje no supervisado". El aprendizaje supervisado requiere que el analista de datos identifique un atributo objetivo o una variable dependiente. La técnica de aprendizaje supervisado examina cuidadosamente los datos para buscar patrones y relaciones entre otros atributos y el atributo objetivo, en nuestro caso ya hemos comentado que es un aprendizaje supervisado y para ello Oracle nos ofrece los siguientes algoritmos: Naïve Bayes, Árbol de Decisión, Modelos Lineales Generalizados y Máquinas de Vectores Soporte.

A continuación detallamos los algoritmos y técnicas que vamos a utilizar en nuestro estudio:

1. **Extracción de la información no espacial.-** En el estudio se utilizaran las siguientes variables de entrada para predecir nuestro atributo target que es el número de pacientes:

   - **Datos históricos de asistencia**: Este conjunto de datos contiene la fecha, código de centro de salud y el número de visitas a dicho centro de atención primaria y el tipo de asistencia (clínica, administrativa por receta, administrativa por parte médico). Estos datos han sido aportados por el Distrito Sanitario de Jaén en un fichero plano (txt) que debemos cargar en nuestra Base de Datos.

   - **Datos meteorológicos**: Los datos meteorológicos utilizados provienen de dos estaciones meteorológicas en la ciudad de Jaén. Los datos contiene la fecha, temperatura máxima, media y mínima, humedad relativa y precipitaciones. Esta información es extraída de la página web de la consejería de medio ambiente de la Junta de Andalucía.

   - **Niveles de contaminación**: Esta información es proporcionada por la Consejería de Medio Ambiente y

Ordenación del Territorio, los datos son servidos por este organismo como: Buena, Admisible, Mala o Muy Mala, dependiendo del índice parcial para cada contaminante: SO2 dióxido de sulfuro, partículas, dióxido de nitrógeno, NO2, monóxido de carbono CO y O3 ozono. (REDIAM, 2014) [43].

2. **Extracción de la información espacial.-** Como se ha indicado anteriormente uno de los objetivos es determinar los factores locales a los centros de Salud que afectan a la afluencia de pacientes. Para ello hemos considerado dos datos que pueden influir:

   - **Tipo de población adscrita al centro de salud**: El objetivo es obtener para cada centro de salud el porcentaje de pacientes adscritos por rango de edad. Este dato no ha sido proporcionado por el distrito Sanitario de Jaén, por lo que será obtenido a partir de un Sistema de Información Geográfico (GIS) [44] construido con MapInfo [45], en este GIS cargaremos el callejero de Jaén y delimitaremos las zonas de influencia de cada centro de Salud, finalmente se cargarán los datos por unidad censal del INE (Instituto Nacional de Estadística). Finalmente realizaremos una consulta SQL al GIS podremos saber porcentaje de pacientes por rango etario que están adscritos a los distintos centros de salud de Jaén.
   - **Nivel económico de la zona de influencia del Centro de Salud**: En este punto, necesitamos un indicador global económico de la zona de influencia de cada centro de salud. Actualmente en Jaén no hay datos oficiales, por lo que buscaremos un indicador externo, este indicador lo calcularemos con el precio medio de los pisos de las zonas de influencia de los centros de salud y lo cargaremos en el SIG construido para nuestro proyecto. Finalmente con una consulta SQL obtendremos un dato global aproximado del nivel económico de la zona que atiende cada centro de salud.

3. **Exploración de la Información.-** Todos los datos serán explorados estudiando su distribución mediante Histogramas. El histograma es una técnica gráfica utilizada para presentar gran cantidad de datos. Para la construcción del histograma se requiere elaborar una tabla de distribución de frecuencias.

El histograma de frecuencias es una representación visual de los datos en donde se evidencian fundamentalmente tres características: forma, acumulación o tendencia posicional y dispersión o variabilidad. El objetivo de esta técnica es revisar los datos y limpiar aquellos que claramente son erróneos y pueden provocar ruido en los modelos.

4. **Detección de Anomalías.-** Para la detección de Anomalías usaremos el algoritmo One-Class Support Vector Machine (SVM) [46]. La detección de anomalías se implementa como una clasificación de clase, ya que sólo una clase se representa en los datos de entrenamiento. Un modelo de detección de anomalías predice si un punto de datos es típico de una distribución o no.

   El objetivo de la detección de anomalías es identificar los casos que no son habituales dentro de los datos que aparentemente son homogéneos. La detección de anomalías es una herramienta importante para la detección de fraudes, intrusiones en la red, y otros eventos raros que pueden tener una gran importancia, pero que son difíciles de encontrar.

5. **Análisis de la Información**.- Se realizará un profundo análisis de la información para homogeneizarla, hay que tener en cuenta que tenemos datos con orígenes muy diversos (datos del Distrito Sanitario de Jaén, datos meteorológicos provenientes de una página web, datos ambientales provenientes de la red REDIAM y datos espaciales provenientes de un SIG).

6. **Trasformaciones de la información**.- En función del análisis del punto 5 se realizarán una serie de transformaciones que faciliten la integración de la información proveniente de las distintas fuentes.

7. **Agrupación de la información**.- En esta fase realizaremos las agrupaciones necesarias para preparar de forma óptima los datos para que puedan ser utilizados por los modelos de minería de datos. Es fundamental agrupar la información con el objetivo de maximizar los datos de entrenamiento para generar los modelos, sin perjudicar la calidad de la predicción. Cuantos más datos de entrenamiento

tengamos en un periodo de tiempo concreto, mucho más preciso será nuestro modelo generado.

8. **Integración de los datos**.- Para nuestro trabajo contamos con información heterogénea que proviene de distintas fuentes, para poder realizar el estudio de minería de datos es necesario integrar toda la información en una única fuente que sirva como origen a los modelos predictivos.

9. **Detección del peso de los atributos de entrada sobre el atributo objetivo**.- El algoritmo Longitud mínima de la descripción (LMD) [47] ayuda a identificar los atributos con mayor influencia sobre el atributo objetivo. A menudo, conocer los atributos con mayor influencia ayuda a comprender y gestionar mejor el negocio y a simplificar las actividades de modelado. Además, estos atributos pueden indicar los tipos de datos que se desean añadir para argumentar los modelos.

   El algoritmo Mínimum Description Length (MDL) es un modelo teórico que se basa en el principio de selección. MDL considera cada atributo como un simple modelo predictivo de la clase de objetivo. Estos modelos de predicción individuales se comparan y clasifican con respecto a la métrica MDL (compresión en bits). MDL penaliza la complejidad del modelo para evitar el sobre ajuste. Se trata de un enfoque basado en principios, que tenga en cuenta la complejidad de los factores de predicción (según modelos) para hacer las comparaciones justas.

   El objetivo de esta fase del estudio es descartar aquellos atributos que no tiene relación con el objetivo de nuestro estudio.

10. **Uso de varios algoritmos de regresión** [48,49,50].- Para realizar la predicción de pacientes que acuden al centro de Salud probaremos varios algoritmos de regresión. El objetivo del análisis de regresión consiste en determinar los valores de parámetros para una función que hacen que se adapte mejor a un conjunto de observaciones. La regresión es el proceso de estimación del valor de un objetivo continuo (y) como una función (F) de uno o más predictores (X_1, X_2,

..., X_n), un conjunto de parámetros ($\theta 1$, $\theta 2$, ..., θn), y una medida de error (e).

$$y = F(x, \theta) + e$$

El proceso de formación de un modelo de regresión consiste en encontrar los mejores valores de los parámetros para que la función minimice la suma de errores cuadráticos.

Existen diferentes familias de funciones de regresión y las diferentes formas de medir el error.

- **Regresión Lineal-.** La forma más simple de regresión con un único predictor. Una técnica de regresión lineal se puede utilizar si la relación entre *x* e *y* puede ser aproximada con una línea recta. En la figura 7 podemos ver un ejemplo.

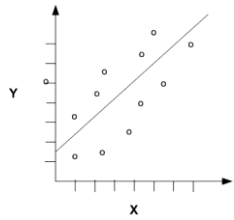


Figura 7. Ejemplo de regresión lineal con relación entre x e y.

- **Regresión no lineal.-** A menudo la relación entre *x* e *y* no se puede aproximar con una línea recta. En este caso, tenemos que usar una técnica de regresión no lineal.

  En la Figura 8 podemos ver como x e y tienen una relación no lineal. Oracle Data Mining desarrolla este

tipo de regresión con el algoritmo SVM con Kernel Gaussiano.

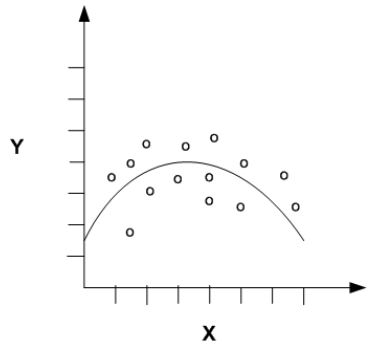


Figura 8. Ejemplo de regresión no lineal con relación entre x e y.

- **Regresión Multivariante.-** Es una regresión con varios predictores (X_1, X_2, ..., X_n). Este tipo de regresión también se conoce como regresión múltiple.

Los algoritmos que usaremos en el estudio son: **SVM**: Máquinas de Vectores Soporte (con Kernel Lineal y Gausiano)[51,52,53] y GLM: Modelo lineal generalizado [54].

SVM es un algoritmo de clasificación y regresión que utiliza la teoría de aprendizaje de las máquinas para maximizar la precisión de los pronósticos sin ajustar excesivamente los datos. SVM utiliza una transformación no lineal opcional de los datos de entrenamiento, seguida de la búsqueda de las ecuaciones de regresión en los datos transformados para separar las clases (para objetivos categóricos) o ajustar el objetivo (para los objetivos continuos).

La implementación de SVM de Oracle permite que se generen modelos mediante el uso de los dos kernels: lineal o gaussiano. El kernel lineal omite la transformación no lineal de una vez, de tal

forma que el modelo resultante sea, en esencia, un modelo de regresión.

SVM es un algoritmo que funciona muy bien en conjuntos de datos que tienen muchos atributos, incluso si hay muy pocos datos para entrenar el modelo. No hay límite superior en el número de atributos; las únicas limitaciones son las impuestas por el hardware. Las redes neuronales tradicionales no funcionan bien en estas circunstancias. Este algoritmo es óptimo en teoría en nuestro estudio para predecir el número de pacientes que necesitarán una atención médica en días festivos que caen entre semana, ahí contamos con muy pocos datos para el entrenamiento.

En nuestro problema usaremos SVM con las dos configuraciones de Kernel y los siguientes parámetros:

- **Tamaño de caché de Kernel.** Se especifica el tamaño de caché (en bytes), que se utiliza para almacenar los kernels calculados durante la operación de generación. Como es de esperar, las cachés de mayor tamaño generalmente originan construcciones más rápidas. En nuestro caso lo configuraremos con el valor predeterminado de 50 MB.

- **Tolerancia de convergencia**. Especifica el valor de tolerancia permitido para la generación del modelo antes de terminar. El valor debe estar comprendido entre 0 y 1. El valor configurado en nuestro problema es de 0,001. Los valores mayores tienden a originar generaciones más rápidas pero modelos menos exactos. En nuestro estudio buscamos un valor muy próximo a 0 para garantizar la máxima precisión sin penalizar el tiempo de cómputo.

- **Especificar desviación estándar**. Permite especificar el parámetro de desviación estándar que el kernel gaussiano utiliza. Este parámetro afecta al equilibrio entre la complejidad del modelo y la capacidad para generalizar a otros conjuntos de datos (sobre ajustando y sub ajuste de los datos). Los valores de desviación estándar mayores favorecen el sub ajuste. Este parámetro lo dejaremos con

configuración por defecto, con esta configuración el sistema estima automáticamente este parámetro a partir de los datos de entrenamiento.

- **Especificar épsilon**. Especifica el valor del intervalo del error permitido en la generación de modelos no sensibles a épsilon. En otras palabras, distingue pequeños errores (que se ignoran) de los grandes (que no se ignoran). El valor debe estar comprendido entre 0 y 1. Al igual que el parámetro anterior lo configuramos como automático, de tal forma que es calculado por el sistema en función de los datos de entrenamiento.

- **Especificar factor de complejidad**. Permite determinar el factor de complejidad, que equilibra el error del modelo (como se mide con respecto a los datos de entrenamiento) y la complejidad del modelo a fin de evitar el sobre ajuste o el sub ajuste de los datos. Los valores mayores proporcionan una penalización mayor a los errores, lo que supone un mayor riesgo de sobre ajuste de los datos; los valores menores proporcionan una penalización menor en los errores y pueden originar sub ajustes. En nuestro problema este valor se configura como automático y el sistema lo determina en función de los datos de entrenamiento.

- **Método de normalización**. Especifica el método de normalización utilizado para la entrada continua y el atributo objetivo. Es posible seleccionar puntuaciones Z, Mín.-Máx. o Ninguna. Oracle realiza la normalización automáticamente si no se indica ninguna. En nuestro caso lo dejamos sin seleccionar.

- **Aprendizaje activo.** Proporciona un método para gestionar los conjuntos generados de gran tamaño. Con el aprendizaje activo, el algoritmo crea un modelo inicial en base a una pequeña muestra antes de aplicarlo a todo el conjunto de datos de entrenamiento y, a continuación, actualiza la muestra y el modelo de forma gradual en función de los resultados. Este ciclo se repite hasta que el modelo converge en los datos de entrenamiento o hasta

que se alcanza el número máximo de vectores de soporte permitidos. En nuestro caso para agilizar los cálculos activamos este parámetro.

Por otra parte también se utilizará Modelo lineal generalizado (GLM). Los modelos lineales hacen una serie de suposiciones restrictivas, sobre todo, que el objetivo (variable dependiente) se distribuye normalmente condicionada a que el valor de los predictores tenga una varianza constante, independientemente del valor de respuesta predicho. La ventaja de los modelos lineales es la simplicidad computacional. Los modelos lineales generalizados relajan las restricciones, que a menudo son violados en la práctica. Por ejemplo, binaria (sí/no o 0/1) las respuestas no tienen varianza igual en todas las clases. Por otra parte, la suma de los términos de un modelo lineal general pueden tener rangos muy grandes que abarcan valores muy negativos y muy positivos.

GLM acomodar las respuestas que violan los supuestos del modelo lineal a través de dos mecanismos: una función de enlace y una función de la varianza. La función de enlace transforma el rango objetivo para potencialmente tender a infinito para que la forma simple de los modelos lineales se pueda mantener. La función de varianza expresa la varianza como una función de la respuesta prevista, acomodando las respuestas con varianzas no constantes (tales como las respuestas binarias).

Un modelo lineal generalizado es adecuado para las predicciones en las que es probable que tenga una distribución no normal, como una multinomial o una distribución de Poisson. Del mismo modo, un modelo lineal generalizado es útil en los casos en que es probable que sea no lineal la relación o vínculo, entre los predictores y el atributo target.

En nuestro estudio usaremos el modelo GML con la siguiente configuración:

- **Nivel de confianza de coeficiente.** El grado de certidumbre, entre 0 y 1, del valor predicho para el objetivo en un intervalo de confianza calculado para el modelo. Los límites de confianza se devuelven con los estadísticos de coeficientes. En nuestro problema configuramos este parámetro a 0.95, con esta

configuración buscamos el equilibrio entre tiempo de computo y calidad del modelo.

- **Preparación automática de datos**. Con esta opción Oracle Data Mining realiza automáticamente las transformaciones de datos requeridas por el algoritmo. En nuestro estudio activaremos esta opción.

- **Método de normalización.** Especifica el método de normalización utilizado para la entrada continua y los campos objetivo. Es posible seleccionar Puntuaciones Z, Mín.-Máx. o Ninguna. Oracle realiza la normalización automáticamente si la casilla de verificación "preparación de datos" esta seleccionada como automática. En nuestro caso lo configuraremos como automático.

- **Gestión de valores perdidos**. Especifica cómo se procesarán los valores perdidos en los datos de entrada:
    - *Sustituir con media o modo*. Sustituye los valores perdidos de los atributos numéricos con el valor de la media y sustituye los valores perdidos de los atributos categóricos con el modo.
    - *Solamente utilizar registros completos*. Ignora los registros con valores perdidos.

  En nuestro estudio indicaremos que ignore los valores perdidos.

11. **Estudio de los coeficientes estandarizados de la regresión.-** Los coeficientes estandarizados remiten a una escala única (en desviaciones típicas respecto al 0) en que se miden las diferentes variables y por tanto pueden constituir la base para conocer exactamente en cuántos puntos se modifica la variable Y por cuenta de cada regresor comparativamente. Esto nos sirve desde un punto de vista práctico a entender mejor la logia de nuestro negocio, determinando qué valores de las variables afectan en mayor o menor medida a la asistencia de pacientes a los Centros de Salud.

3. DISEÑO DEL ESTUDIO

El eje principal de nuestro estudio son los datos proporcionados por el distrito sanitario de Jaén con el número de pacientes que han utilizado el Sistema Sanitario Público de atención primaria en Jaén durante los años: 2007, 2008, 2009, 2010 y 2011 de todos los Centros de Salud de Jaén Capital:

- C.S. Belén
- C.S. La Magdalena
- C.S. El Valle
- C.S. Federico Castillo
- C.S. San Felipe
- C.S. Virgen de la Capilla
- C.S. Fuentezuelas

Toda esta información ha sido extraída de la Base de Datos Diraya. Para crear el modelo de Data Mining trabajaremos con dos subconjuntos de datos (Subset): el primero contendrá los datos de 2007, 2008, 2009 y 2010 y será utilizado para crear el modelo predictivo. El segundo Subset contendrá los datos de 2011 y lo usaremos para validar los Modelos. En la figura 9 puede verse como se utilizarán los datos.

Figura 9. Diseño del uso de los datos para la generación de los modelos y validación de los mismos.

Al sistema se le añadirá otro tipo de información como: información Meteorológica, calidad ambiental y otro tipo de información particular de cada centro de salud. Esta información se analizará en profundidad para determinar el peso estadístico que tienen estos elementos sobre el aumento o disminución del uso de los servicios Sanitarios. La información Meteorológica y de calidad del aire en Jaén la obtendremos de Red de Información Ambiental de Andalucía (REDIAM). En Jaén hay actualmente dos nodos de esta Red que recoge numerosa información y que es publicada a Diario en la WEB de la Consejería de Medio Ambiente de Andalucía. A continuación se muestran los datos de las estaciones de Jaén de las que hemos obtenido los datos de nuestro estudio.

1. Datos de la estación: SIVA-RONDA DEL VALLE

   Red: SIVA
   Código Estación: SIVA59
   Denominación: RONDA DEL VALLE
   Provincia: JAEN
   Municipio: JAEN
   Área Climática: Alto y Medio Guadalquivir
   Coordenada X: 431294
   Coordenada Y: 4182158
   Coordenada Z: 460
   Tipo: A

2. Datos de la estación: SIVA-LAS FUENTEZUELAS

   Red: SIVA
   Código Estación: SIVA63
   Denominación: LAS FUENTEZUELAS
   Provincia: JAEN
   Municipio: JAEN
   Área Climática: Alto y Medio Guadalquivir

Coordenada X: 428759

Coordenada Y: 4182414

Coordenada Z: 520

Tipo: A

En la figura 10 se muestra un mapa de las distintas fuentes de datos que usaremos para nuestro estudio.

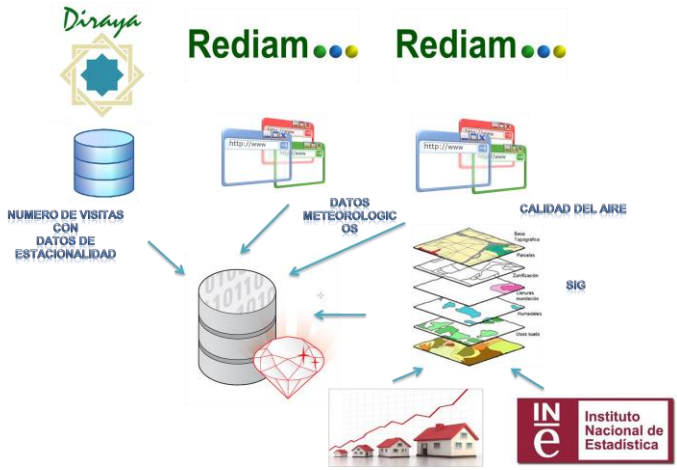


Figura 10. Esquema de las fuentes de datos para la generación de los modelos predictivos.

Otra fuente de información importante para nuestro estudio es obtener datos particulares de cada centro de salud, como el tipo de población atendida, para ello construiremos un SIG donde delimitaremos la zona de influencia de cada centro de salud y con los datos de las unidades Censales obtenidas del Instituto Nacional de Estadística, obtendremos el porcentaje de usuarios atendidos por rango etario.

Otro dato particular de los centros de salud será el nivel económico de su zona de influencia. Para obtener este dato, usaremos un indicador externo como es el precio medio de un piso, ya que no existe en Jaén datos oficiales del nivel económico de una zona concreta. Este indicador es calculado al obtener de algunas inmobiliarias de Jaén los precios de venta de un pisos estándar de la zona de influencia de Centro de Salud, este datos es almacenado también el SIG construido para este trabajo.

A pesar de que la información con el número de visitas a los centros de salud son servidos por parte del distrito sanitario de Jaén separados por centro de Salud, en el estudio no vamos a tener en cuenta el centro de salud ni el tipo de demanda, con lo cual se generará un modelo global de la ciudad de Jaén. En figura 11 se muestra la integración de los datos:

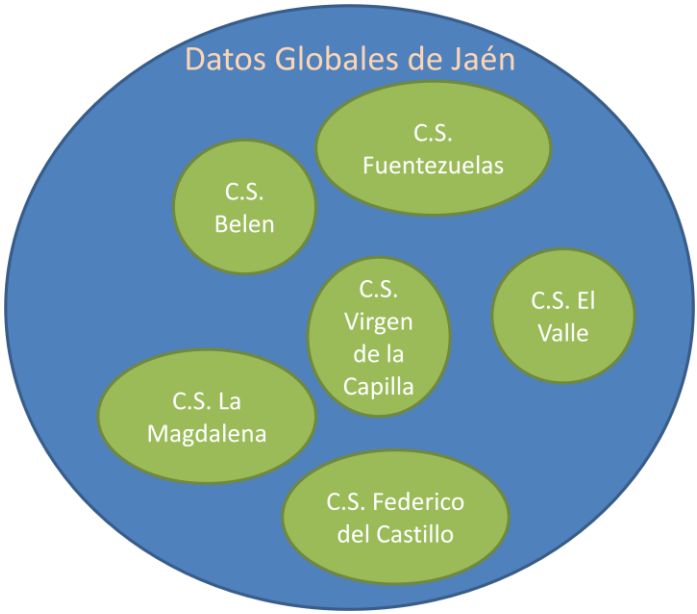


Figura 11. Integración de los datos para la generación de los modelos.

Con esta integración pretendemos evitar que los modelos predictivos tenga en cuenta de forma implícita las características particulares de cada centro de salud, es por ello que generarnos un modelo global para toda la ciudad de Jaén. Una vez generado dicho modelo lo aplicaremos a los distintos centros de salud, como es de esperar al aplicar el modelo a nivel de centro de salud (un centro concreto) el porcentaje de errores debe de aumentar, ya que como hemos comentado anteriormente partimos de la hipótesis de que hay ciertos factores locales que afectan a la afluencia de pacientes a cada centros. A continuación añadiremos al modelo las variables espaciales generadas por el SIG y analizaremos si el error particular de cada centro de salud disminuye, demostrándose así la eficacia e importancia de las variables locales de los centros de salud para la predicción de pacientes que requerirán una atención sanitaria.

3.1. Información no espacial para la generación de los modelos predictivos

Los primeros modelos predictivos los generamos con información no espacial, como datos históricos de asistencia de pacientes a los centros de salud (datos de estacionalidad), datos meteorológicos (temperatura, humedad y precipitaciones) y datos de calidad de aire, donde se tienen en cuenta los principales contaminantes ambientales.

3.1.1. Datos de estacionalidad

De la base de datos de Diraya obtenemos los datos de estacionalidad. Esta información la ha extraído el personal del Distrito Sanitario de Jaén y nos la proporcionan en un fichero plano. Sólo tendremos datos globales de visitas a los centros de salud agrupados por fecha, centro de Salud (sólo los centros de Salud de Jaén) y tipo de demanda (solo 3 valores: clínica, receta, certificado médico). A continuación en la tabla 3 podemos ver una muestra de los datos.

FECHA DE CONSULTA	NUMERO TOTAL DE VISITAS	CODIGO DEL CENTRO DE SALUD	TIPO DE ASISTENCIA
12/04/2011	320	22126	clínica
31/05/2011	330	22126	clínica
14/07/2011	184	22126	receta
17/08/2011	189	22126	receta
17/11/2011	369	22126	clínica
21/02/2011	428	22566	clínica
13/06/2011	442	22566	clínica
20/10/2011	35	22566	certificado

Tabla 3. Ejemplo de los datos proporcionados por el distrito sanitario de Jaén.

3.1.2. Datos Meteorológicos

De la página Web de la Consejería de Medio Ambiente de la Junta de Andalucía podemos descargar los ficheros en formato "CSV" con la información meteorológicos histórica registradas en las estaciones de Jaén. En la tabla 4 podemos ver un ejemplo de la información obtenida.

Est	Día	T.Med	T.Máx	T.Mín	Prec.	Hum.R	V.Med
SIVA63	01/01/2007	9.3 ºC	14 ºC	6 ºC	0 mm	66.8 %	n/a
SIVA63	02/01/2007	8.8 ºC	14 ºC	4 ºC	0 mm	79%	n/a
SIVA63	03/01/2007	9.5 ºC	17 ºC	3 ºC	0.2 mm	50.7 %	n/a
SIVA63	04/01/2007	10 ºC	15 ºC	7 ºC	0 mm	47.2 %	n/a
SIVA63	05/01/2007	8.8 ºC	14 ºC	3 ºC	0.2 mm	69.1 %	n/a
SIVA63	06/01/2007	9.5 ºC	14 ºC	4 ºC	0.2 mm	66%	n/a
SIVA63	07/01/2007	8.9 ºC	14 ºC	4 ºC	0.2 mm	77.2 %	n/a
SIVA63	08/01/2007	7.7 ºC	11 ºC	4 ºC	0.2 mm	87.2 %	n/a
SIVA63	09/01/2007	8.1 ºC	15 ºC	3 ºC	0 mm	76.3 %	n/a
SIVA63	10/01/2007	7.9 ºC	n/a	n/a	n/a	73.8 %	n/a
SIVA63	11/01/2007	8.1 ºC	16 ºC	1 ºC	0 mm	56.1 %	n/a
SIVA63	12/01/2007	9.1 ºC	17 ºC	3 ºC	0 mm	42.2 %	n/a

Tabla 4. Ejemplo de los datos meteorológicos obtenidos de REDIAM.

3.1.3. Datos de Calidad ambiental

Los datos de calidad de Aire los obtenemos también de la página web de la Consejería de Medio Ambiente de la Junta de Andalucía, ahí podemos encontrar información histórica de los días con calidad de aire: Muy buena, Aceptable, Mala y Muy Mala. Los datos son globales agrupados por meses. En la tabla 5 se muestra un ejemplo de la información obtenida.

MES	AÑO	ESTACION DE RECOGIDA	NUM. DE DIAS	NUM. DE DIAS CON CALIDAD ACEPTABLE	NUM. DE DIAS CON CALIDAD MALA	NUM. DE DIAS CON CALIDAD MUY MALA
01	2007	LAS FUENTEZUELAS	21	10	0	0
01	2007	RONDA DEL VALLE	3	12	5	11
02	2007	LAS FUENTEZUELAS	4	24	0	0
02	2007	RONDA DEL VALLE	2	21	5	0
03	2007	LAS FUENTEZUELAS	0	31	0	0

Tabla 5. Ejemplo de los datos de calidad de aire obtenidos de REDIAM.

Como hemos comentado anteriormente hay dos estaciones en Jaén que recogen esta información. Estas estaciones miden los siguientes niveles de contaminantes ambientales:

- SO2: Dióxido de Azufre.
- Partículas:
    - NO2: Dióxido de nitrógeno
    - CO: Monóxido de carbono
    - O3: Ozono

Como puede observarse en los datos, tenemos por cada día la calidad ambiental clasificada en Buena, Admisible, Mala o Muy Mala. Esta clasificación la realiza la Consejería de Medio Ambiente de la Junta de Andancia según el siguiente criterio:

En cada estación se calcula un índice individual para cada contaminante, conocido como índice parcial. A partir de ellos se obtendrá el índice global que coincidirá con el índice parcial del contaminante que presente el peor comportamiento. De este modo, existirá un índice global para cada estación.

- **Rango cualitativo:** el índice estará dividido en cuatro tramos, que definirán los Principales estados de calidad de aire; estos serán buena, admisible, mala o muy mala. En la tabla 6 podemos ver el rango de valores de los contaminantes.

Valor del índice	Calidad del aire
0-50	Buena
51-100	Admisible
101-150	Mala
>150	Muy mala

Tabla 6. Rango de los valores de los contaminantes para clasificar la calidad del aire.

- **Rango cuantitativo.-** En la tabla 7 podemos ver el índice parcial de cada contaminante.

ÍNDICE PARCIAL PARA CADA CONTAMINANTE.					
INDICE	SO2 (24H)	PARTICULAS (24 H)	NO2 (1H MÁX)	CO (8H MÓVIL MÁX)	O3 (8H MÓVIL MÁX)
0-50	63	25	120	5000	60
51-100	125	50	240	10000	120
101-150	187	75	360	15000	180
>150	>187	>75	>360	>15000	>180

Tabla 7. Índice parcial de cada contaminante.

- En el caso del **SO2** siempre que se supere el valor límite horario (350 µg/m3) fijado en el R.D. 1073/2002, la calidad del aire será considerada "mala" y siempre que se supere el umbral de alerta (500 µg/m3) registrados durante tres horas consecutivas la calidad del aire será considerada "muy mala".

- En el caso del **NO2** se tiene en cuenta para el cálculo del índice, el valor límite medido en 1 hora que establece el R.D. 1073/2002. Sin embargo, siempre que se supere el

umbral de alerta (400 µg/m3) registrados durante tres horas consecutivas la calidad del aire será considerada "muy mala".

- En el caso del **O3** siempre que se supere el valor de información a la población, valor horario (180µg/m3) fijado en el R.D.1796/2003, la calidad del aire será considerada "mala" y si se supera el umbral de alerta para la población, valor horario (240 µg/m3) la calidad del aire se considerará "muy mala".

3.2. Información Espacial para la generación de los modelos predictivos

Todos los modelos predictivos que afectan a un centros de salud concretos o a una farmacias, serán generados añadiendo a las variables no espaciales otras espaciales. En el punto anterior se hizo un análisis de la información que usaremos para generar los modelos globales para toda la ciudad de Jaén, pero estos modelos como veremos más adelante, no se adaptan adecuadamente al ser aplicados a un centro de salud concreto, es por ello que tenemos que añadir a nuestro sistema variables espaciales a cada centro de salud que modele sus particularidades.

3.2.1. Tipo de población atendida

Desde el distrito de Jaén no nos proporcionan información sobre el tipo de población adscrita a cada centro de salud. El objetivo es obtener el porcentaje de usuarios atendidos por cada centro de salud por edad, es decir porcentaje de pacientes de 0-14 años (pacientes pediátricos), de 15 – 24 años de 25 – 34 y así sucesivamente hasta llegar a los mayor de 85 años.

Al no disponer de esta información directamente desde el distrito de Jaén, construimos un Sistema de Información Geográfico (SIG). Un SIG se define como conjunto de procedimientos con capacidad de construir modelos o representaciones del mundo real, a partir de datos geográficos de localización. Estos sistemas, utilizan herramientas de gran capacidad de administración de datos y procesamiento gráfico que logran capturar, almacenar, visualizar y analizar información georeferenciada.

El SIG es construido con la herramienta de MapInfo, MapInfo Professional es una herramienta para la creación de mapas que permite llevar a cabo análisis geográficos complejos: zonificación, acceso a datos remotos, arrastrar objetos de mapa y soltarlos en aplicaciones, creación de mapas temáticos que revelen patrones en los datos y muchas otras funciones.

Para nuestro problema cargamos en MapInfo el callejero de Jaén y delimitamos las calles que son atendidas por cada centro de salud, en la figura 12 podemos ver el mapa con el posicionamiento de los centros de salud de Jaén y su ámbito de influencia.

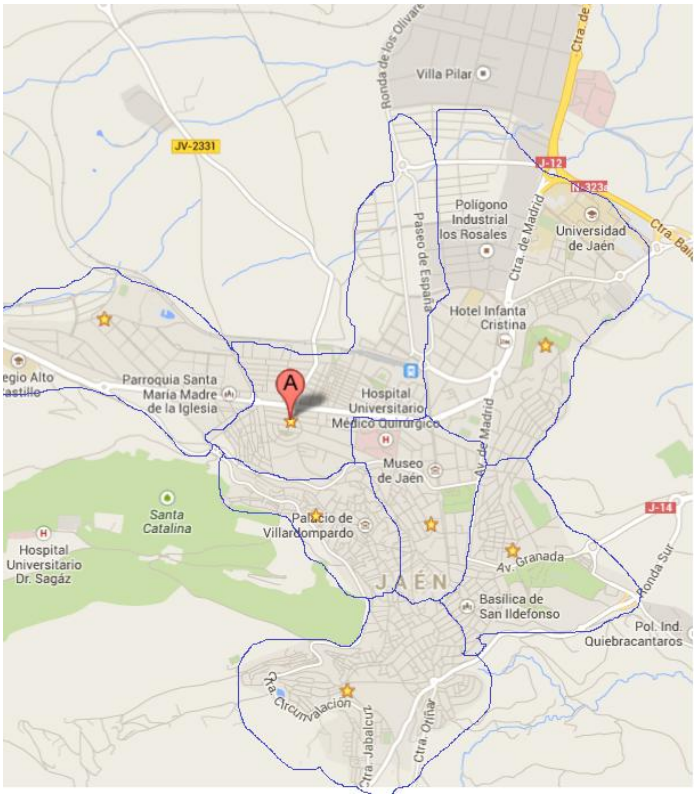


Figura 12. Posicionamiento geográfico de los centros de salud y delimitación de su zona de influencia.

Finalmente cargamos los datos de las unidades censales obtenidas desde el Instituto Nacional de Estadística (INE). Con esta información tenemos por unidad censal la edad y el número personas. A continuación realizamos una consulta SQL sobre la base de datos de MapInfo para obtener los datos de la tabla 8.

	San Felipe	Belén	El valle	La Magdalena	Virgen de la Capilla	Fuentezuelas	Federico del Castillo
0 -14 años	14,85%	12,70%	15,66%	16,74%	12,06%	18,04%	18,40%
14 -24 años	12,60%	10,90%	11,84%	12,66%	8,55%	13,17%	9,77%
25 -34 años	16,04%	17,69%	17,04%	17,53%	15,80%	16,17%	15,64%
35 -44 años	14,99%	15,72%	17,38%	15,39%	17,47%	18,84%	18,66%
45 -54 años	15,55%	14,23%	14,20%	14,21%	14,10%	15,84%	13,79%
55 -64 años	9,52%	11,41%	9,42%	8,92%	10,92%	8,82%	8,21%
65 -74 años	6,51%	7,88%	6,43%	6,12%	8,80%	4,34%	6,31%
75 -84 años	6,48%	5,92%	4,67%	5,40%	7,27%	2,91%	5,59%
Más de 85 años	3,45%	3,54%	3,37%	3,03%	5,04%	1,87%	3,64%

Tabla 8. Población atendida por cada centro de salud por rango etario.

Al analizar los datos vemos claramente las importantes diferencias que existen entre los distintos centros de Salud en función de la población que atienden, por ejemplo el centro Federico del Castillo tiene adscrito un 18,40% de población pediátrica frente al 12,6% que tiene el centro Virgen de la Capilla. Otra importante diferencia es el bajo porcentaje de personas mayores (mayores de 65 años) que atiende el Centro de salud de las Fuentezuelas frente a otros centros como por ejemplo Belén o Federico del castillo. Estas importantes diferencias hacen que cada Centro de Salud tenga un comportamiento diferente de visitas diferente.

3.2.2. Nivel económico de la zona

Como hemos comentado anteriormente en varias ocasiones, uno de los objetivos de nuestro estudio es determinar la importancia del nivel económico sobre el número de vistas a un centro de salud concreto. Al no disponer de datos económicos oficiales de las zonas de Jaén, buscamos un indicador externo que puede aproximar el nivel económico. Un indicador de nivel económico de una zona puede ser el precio de compra de los pisos, para ellos buscamos en las inmobiliarias de Jaén los precios de un piso estándar medio con las siguientes características: 3 dormitorios, entre 80-90 metros cuadrados con una plaza de garaje y un trastero. Todos estos datos son insertados también en el SIG construido para nuestro trabajo en MapInfo. En la Figura 13 podemos ver la representación grafica del posicionamiento geográfico de los pisos insertados en el SIG.

Figura 13. Mapa de inmuebles analizados para la generación de un indicador económico de la zona.

Finamente realizamos una consulta SQL sobre la base de datos para obtener la media de los precios de la zona. En la tabla 9 podemos observar los resultados:

Centro de Salud	PRECIO MEDIO
FEDERICO DEL CASTILLO	161000
BELEN	139125
VIRGEN DE LA CAPILLA	244750
FUENTEZUELAS	137000
SAN FELIPE	101500
EL VALLE	109250
LA MAGDALENA	94500

Tabla 9. Precio medio de un piso estándar por la zona de influencia de los centros de salud e Jaén.

3.3. Exploración de la Información

A continuación se realiza un profundo análisis de la información que componen nuestro estudio con el objetivo de garantizar la calidad de la misma.

3.3.1. Análisis de los datos.

Se analizan en profundidad los datos no espaciales que servirán como entrada para la generación de los modelos: datos de visitas, datos meteorológicos y ambientales

3.3.1.1 Datos con el número de vistas a los Centros de Salud.

Realizamos un análisis profundo de la distribución de los datos mediante la generación de Histogramas. En primer lugar estudiamos la distribución del número de visitas los centros de Salud que nos ha proporcionado el distrito Sanitario de Jaén. En la distribución que mostramos en la figura 14 podemos observar los porcentajes del número de visitas en las diferentes horquillas, como puede ver en la horquilla de 1 a 86 pacientes representa un porcentaje importante del total (un 10%).

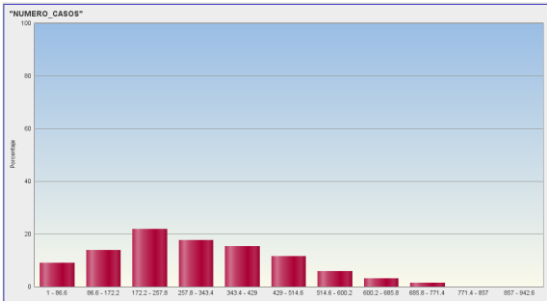


Figura 14. Histograma con la distribución del número de visitas.

Analizamos esta horquilla de datos y vemos que hay muchos días con sólo 1, 2 o 3 visitas. Esto evidentemente es un dato anómalo. Tras contactar con los profesionales médicos para conocer la precedencia de esta información, nos indican que es muy probable que esos datos se deban a paradas de mantenimiento o incidencias de Diraya que impidieron registrar todos los pacientes atendidos. También nos aseguran estos mismos profesionales que ningún día del año hay menos de 50 pacientes en los centros de atención Primaria de Jaén. Teniendo en cuenta esta información eliminamos del estudio todos estos datos anómalos.

Tras su eliminación, el Histograma con la distribución de pacientes atendidos en los centros de Salud durante los años 2007, 2008, 2009 y 2010 queda como se muestra en la Figura 15.

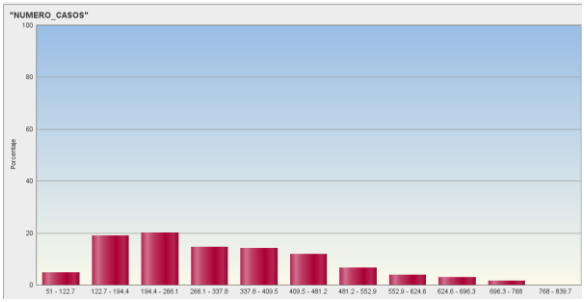


Figura 15. Histograma con la distribución de pacientes atendidos por los centros de salud tras la eliminación de registros ruidosos.

3.3.1.2 Datos meteorológicos.

En el siguiente paso, realizamos el mismo análisis para los datos meteorológicos. Al estudiar los histogramas de los datos climáticos: temperatura mínima, máxima, media, humedad relativa y precipitaciones de las dos estaciones que podemos observar en la Figuras 16, 17, 18, 19 y 20, vernos que hay unos porcentaje muy alto de datos nulos.

- Temperatura Mínima

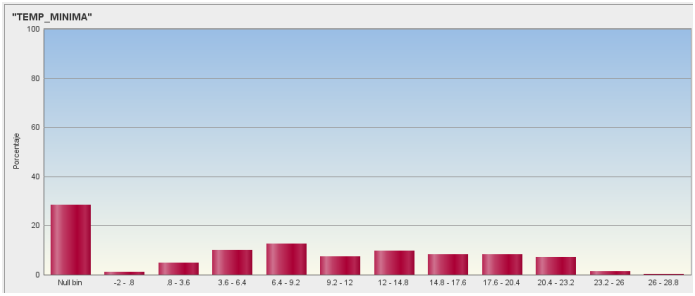


Figura 16. Histogramas con la distribución de temperatura mínima.

- Temperatura Media

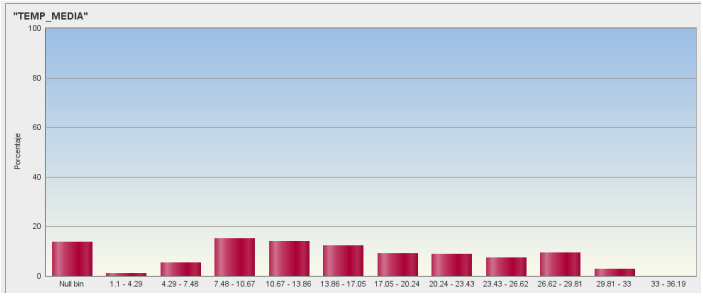


Figura 17. Histogramas con la distribución de temperatura media.

- Temperatura Máxima

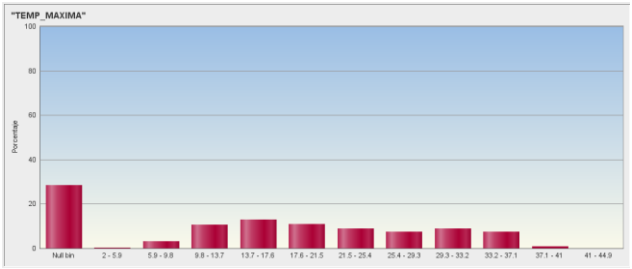


Figura 18. Histogramas con la distribución de temperatura máxima.

- Precipitaciones

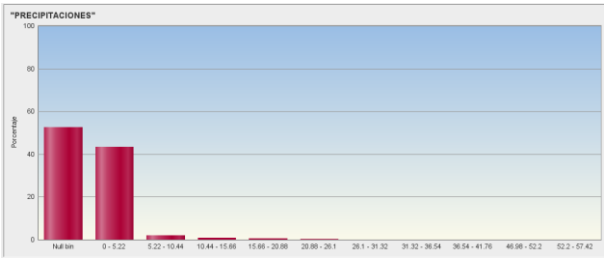


Figura 19. Histogramas con la distribución de precipitaciones.

- Humedad Relativa

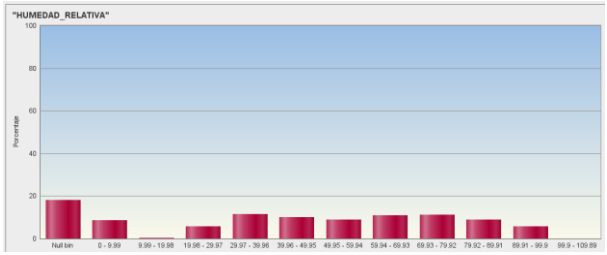


Figura 20. Histogramas con la distribución de la humedad relativa.

Al analizar en más profundidad la procedencia de estos datos nulos, comprobamos que la mayor parte provienen de una de las estaciones que no registró información durante el año 2007 (tabla 10).

AÑO	ESTACION	NUMERO DE NULOS
2007	EL VALLE	7
2007	FUENTEZUELAS	365
2008	EL VALLE	3
2008	FUENTEZUELAS	16
2009	EL VALLE	42
2009	FUENTEZUELAS	25
2010	EL VALLE	12
2010	FUENTEZUELAS	11
2011	FUENTEZUELAS	12
2011	EL VALLE	7

Tabla 10. Estudio de valores nulos de las extradiciones meteorológicas de REDIAM en Jaén.

Todos los registros Nulos son eliminados del estudio. Esto no supone ninguna merma en la información ya que al existir dos estaciones en Jaén nosotros trabajaremos con la media de los datos de las dos estaciones y en

el caso de que algún dato falte en una de ellas, consideremos el dato de la otra estación.

3.3.1.3 Datos ambientales.

Finalmente analizamos los datos de Calidad del Aire en Jaén. En este caso no detectamos ninguna anomalía en ellos. En la Figura 21, 22, 23 y 24 pueden verse los histogramas con la distribución de la Información:

- Número de días con una Calidad ambiental buena

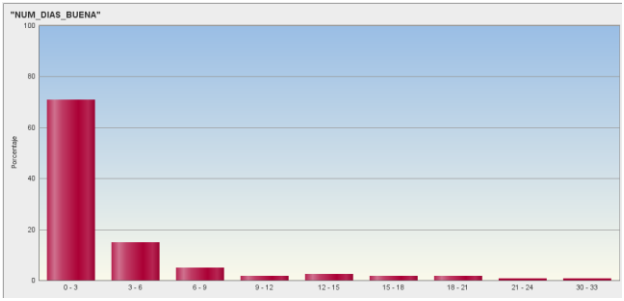


Figura 21. Histogramas con la distribución del número de días con calidad ambiental buena.

- Número de días con una Calidad ambiental Admisible

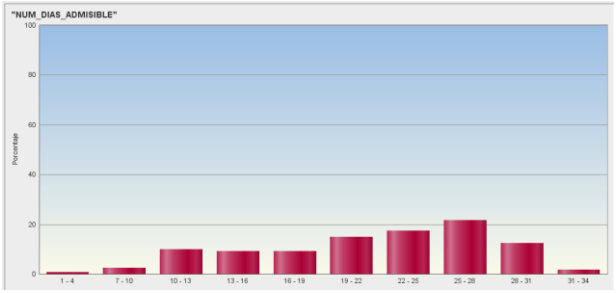


Figura 22. Histogramas con la distribución del número de días con calidad ambiental admisible.

- Número de días con una Calidad ambiental Mala

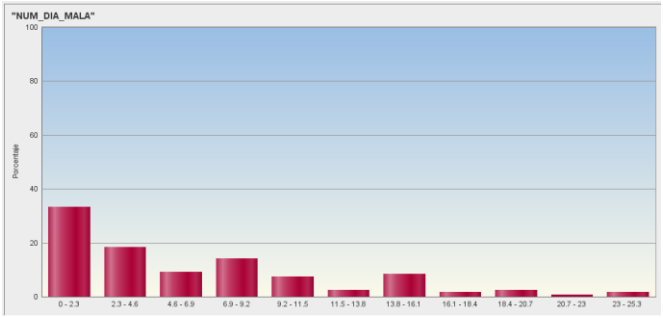


Figura 23. Histogramas con la distribución del número de días con calidad ambiental mala.

- Número de días con una Calidad ambiental Muy Mala

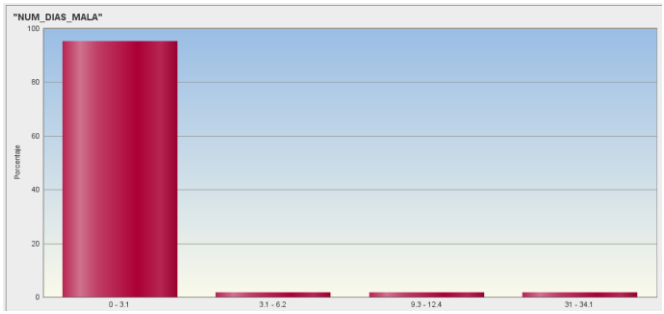


Figura 24. Histogramas con la distribución del número de días con calidad ambiental muy mala.

3.3.2. Detección de Anomalías mediantes algoritmos de Minería de Datos.

Antes de proceder a la agrupación e integración de los datos se lleva a cabo una búsqueda de anomalías en ellos. El objetivo de la detección de anomalías es identificar los casos que no son habituales dentro de los datos que son aparentemente homogéneos.

La detección de anomalías se implementa como una clasificación de clase, ya que sólo una clase está representada en los datos de entrenamiento. Un modelo de detección de anomalías predice si un punto de datos es típico para una distribución dada o no. Un punto de datos atípicos puede ser un valor atípico o un ejemplo de una clase nunca antes vista. El algoritmo utilizado para la detección de anomalías es One Class SVM con kernel Gaussiano. El Modelo SVM de una sola clase, cuando es aplicado, genera una predicción y una probabilidad para cada caso en los datos. Si la predicción es 1, el caso se considera típico. Si la predicción es 0, el caso se considera anómalo.

Al aplicar este proceso sobre los datos de nuestro estudio nos encontramos muchos datos erróneos en el numero de vistas al médico, estos datos anómalos de ser considerados por los modelos afectaría negativamente a la predicción, es por esto que los eliminamos de nuestro estudio. Todas las anomalías se corresponden con días en los que el sistema de Diraya ha registrado muy pocas visitas. Estas anomalías se produjeron principalmente por fallos en los sistemas informáticos de los centros de salud. En la Tabla 11 se muestra un ejemplo donde el 7 de febrero de 2007 miércoles sólo hubo 104 visitas al centro de Salud de San Felipe, sin embargo otros días similares se registraron más de 400 visitas. Todas estas irregularidades son detectadas por el algoritmo y eliminadas del estudio.

Fecha	Día	Mes	Numero de Visita	Predic.	Probabilidad
07/02/2007	Miércoles	Febrero	104	0	0,50184638
14/02/2007	Miércoles	Febrero	432	1	0,50308673
21/02/2007	Miércoles	Febrero	474	1	0,50043688
06/02/2008	Miércoles	Febrero	531	1	0,5014675
27/02/2008	Miércoles	Febrero	515	1	0,50300576
04/02/2009	Miércoles	Febrero	413	1	0,50037512
11/02/2009	Miércoles	Febrero	427	1	0,50106697
18/02/2009	Miércoles	Febrero	484	1	0,50087795
25/02/2009	Miércoles	Febrero	448	1	0,50168938
03/02/2010	Miércoles	Febrero	449	1	0,50133183
10/02/2010	Miércoles	Febrero	446	1	0,50253778
17/02/2010	Miércoles	Febrero	434	1	0,50094404
24/02/2010	Miércoles	Febrero	459	1	0,50377471
02/02/2011	Miércoles	Febrero	489	1	0,50249386
09/02/2011	Miércoles	Febrero	514	1	0,50562715
16/02/2011	Miércoles	Febrero	522	1	0,50391194
23/02/2011	Miércoles	Febrero	477	1	0,50026287

Tabla 11. Ejemplo de la información de salida del algoritmo de detección de anomalías.

3.3.3. Agrupación de la información.

Otra parte fundamental del estudio es establecer el mejor criterio para realizar la agrupación de los datos. Es fundamental agrupar la información con el objetivo de maximizar los datos de entrenamiento para generar el modelo, sin perjudicar la calidad de la predicción. Cuantos más datos de entrenamiento tengamos en un periodo de tiempo mucho mejor será

nuestro modelo. Para explicar mejor en qué consiste la agrupación, lo haremos con un ejemplo: supongamos que tenemos para generar un modelo datos del mes de agosto de cualquier año, si en nuestro modelo no hacemos ninguna agrupación y consideramos el día del mes (1 de agosto, 2 de agosto, etc.), para generar el modelo solo tendríamos 1 dato por día, con lo cual nuestra predicción sería muy pobre. Sin embargo si agrupamos los datos por día de la semana (lunes a domingo), para la predicción de un lunes tendríamos 4 datos (primer lunes de agosto, segundo lunes de agosto, tercer lunes de agosto y cuarto lunes de agosto). Con esta agrupación nuestro sistema seria más preciso ya que tenemos más datos de entrenamiento. Por otro lado también hay que tener en cuenta que al agrupar también podemos estar cometiendo más errores si hay diferencias importantes en las visitas del primer lunes de mes con respecto al 2ª, 3ª o 4ª. Es por ello que hay que buscar una agrupación óptima que garantice el máximo número de datos de entrenamiento de los modelo, sin perjudicar la calidad del dato a predecir.

Para determinar la mejor forma de agrupar la información realizamos un estudio completo de la distribución de la información. Para entender la distribución de los datos, comparamos el número de visitas a los centros de salud durante los años que forman parte del estudio (2007, 2008, 2009 y 2010). En la Figura 25 podemos ver superpuestos las vistas a los centros de salud de Jaén durante los años del estudio.

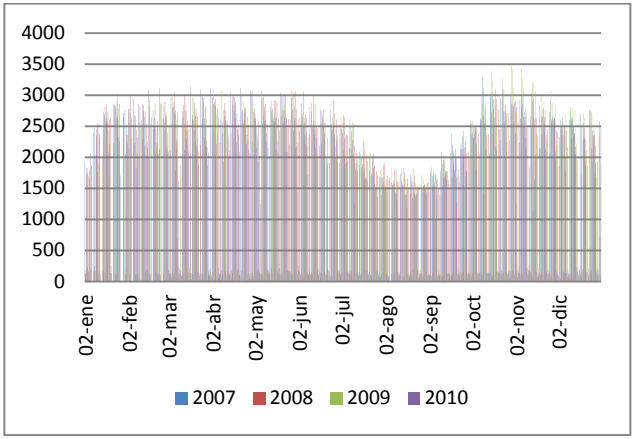


Figura 25. Distribución de los datos de afluencia de pacientes a los centros de salud durante en los años 2007-2010.

En los datos podemos observar que hay mucha estacionalidad, es decir tienen un comportamiento muy similar que se repite cíclicamente cada año en función del día de la semana, mes y el tipo de día. En la Figura 25 mostramos una comparativa del número de visitas a atención primaria agrupadas por meses, como puede verse el comportamiento es muy parecido de unos años a otros.

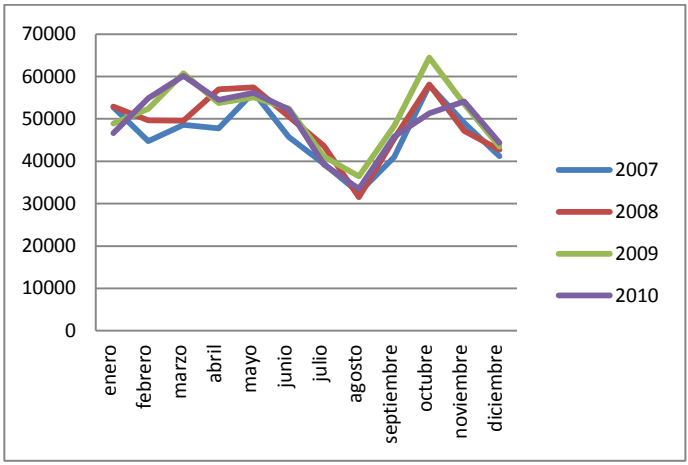


Figura 26. Distribución de la afluencia de pacientes a los centros de salud en el periodo 2007-2010.

Si bajamos el nivel y comparamos los datos de vistas al médico por día de la semana, vemos que el comportamiento también es muy similar y se repite periódicamente todos los días de la semana (Figura 27).

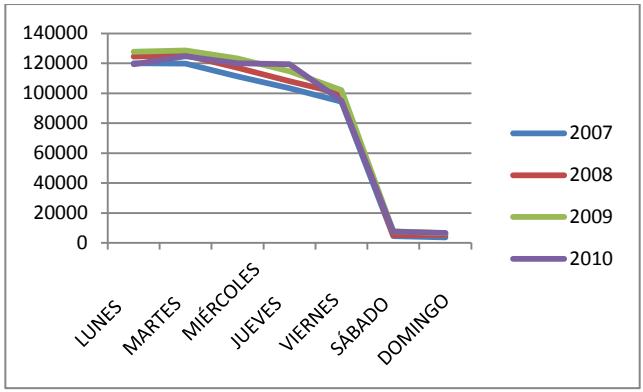


Figura 27. Distribución de la afluencia de pacientes por día de la semana desde el 2007 al 2010.

Otra distribución importante a estudiar es la comparativa del comportamiento de los datos por tipo de día (Festivo o Laborable). Los Festivos y fines de semana sólo se atienden urgencias y evidentemente el número de visitas es mínimo comparado con un día laborable. En la Figura 28 puede verse la diferencia entre los lunes laborables de marzo de 2010 y un lunes Festivo de marzo (01 de Marzo).

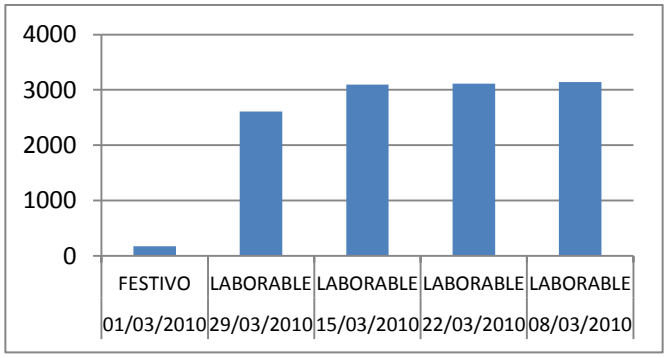


Figura 28. Comparativa del número de visitas de un festivo de marzo con respecto a un día laborable

Finamente analizamos los datos por día de la semana dentro del mismo mes, en este caso también se puede comprobar (Figura 29) que hay una gran estabilidad en días similares, es decir, un martes de Junio se atiende un número muy parecido al segundo, tercer y cuarto martes del mes de Junio. Aquí puede verse una tabla con la comparativa de visitas a los Centros de Salud de Jaén los martes Laborables de Junio.

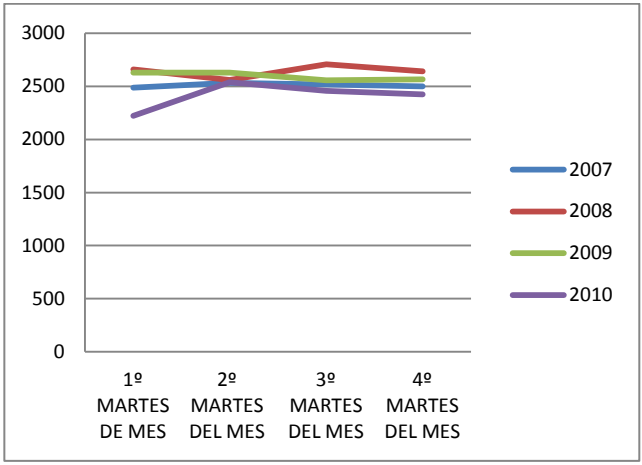


Figura 29. Distribución de visitas a los centros de salud durante 4 martes de mismo mes de los años 2007 - 2010

Con los datos analizados anteriormente, la agrupación más óptima para modelar nuestro problema es agrupar la información por los siguientes campos:

- Día de la semana (lunes, martes, miércoles,…, domingo).
- Tipo de día (Laborable, Festivo)
- Mes (01, 02,… 12).
- Año (2007, 2008, 2009, 2010)

Con esta forma de agrupar la información tenemos dos ventajas, la primera es que para generar nuestro modelo tenemos como mínimo 4 días de la semana por mes (4 lunes en enero, 4 lunes en febrero, etc.), con lo cual nuestro modelo dispondrá de un gran número de valores de entrenamiento. La segunda gran ventaja es que a pesar de que los datos se han analizado en profundidad (tanto con el análisis de los histogramas y la detección de anomalías) y se han eliminado del estudio todos aquellos anómalos, es posible que algunos no se hayan limpiado por estar dentro de un rango permisible de error, en nuestro caso es muy posible que Diraya haya tenido problemas puntuales de 20-30 minutos que ha provocado que se hayan contabilizado menos pacientes de los que realmente han necesitado una asistencia, para estos caso con esta agrupación tiene la ventaja de que ese error mínimo se minimiza aun más, al calcular la media de esos 4 días de la semana, este error se reparte entre los 4 días.

3.3.4. Transformaciones de datos

A continuación se describen todas las transformaciones que se realiza sobre los datos. Este paso es uno de los más importantes dentro del proceso de minería de datos ya que la información tiene que ser transformada y adaptada para que los algoritmos de minería de datos puedan explotar el conocimiento oculto que hay en la información.

3.3.4.1 Datos de Visitas a los centros de salud

En el análisis previo hemos visto cómo el número de visitas a los centros de Salud se modela por el día de la semana, tipo de día, mes y año. Por tanto es necesario realizar varias transformaciones a los datos de vistas para que éstos estén en el formato adecuado.

En la tabla 12 tenemos un ejemplo de los datos de Partida:

FECHA DE CONSULTA	NUMERO TOTAL DE VISITAS	CODIGO DEL CENTRO DE SALUD	TIPO DE ASISTENCIA
12/04/2011	320	22126	clínica
31/05/2011	330	22126	clínica
14/07/2011	184	22126	receta
17/08/2011	189	22126	receta
17/11/2011	369	22126	clínica
21/02/2011	428	22566	clínica
13/06/2011	442	22566	clínica
20/10/2011	35	22566	certificado

Tabla 12. Ejemplo de los datos de visitas a los centros de Salud proporcionados por el distrito sanitario de Jaén.

A estos datos es necesario realizarles las siguientes transformaciones para que cumplan con los criterios que hemos marcado de agrupación:

i. A partir de la fecha obtenemos el día de la semana (Lunes, Martes, etc.).
ii. A partir de la Fecha Obtenemos el mes (01, 02, etc.).
iii. A partir de la Fecha obtenemos el año.
iv. En la primera aproximación del modelo no vamos a distinguir por centro de Salud, es decir vamos a tener los datos totales de Jaén capital. por tanto es necesario sumar el número de visitas de cada centro de Salud.
v. En una primera aproximación el tipo de día se marcara como Laborable todos los días de Lunes a Viernes y festivos los Sábados y Domingos (estos datos son extraídos también de la fecha). En una segunda fase insertaremos los días festivos que ha caído entre semana en los años del estudio.

Como comentamos en el diseño del estudio, en esta parte del estudio no queremos hacer distinción entre los centros de Salud porque entonces el modelo recogería las particularidades de cada centro y como se ha indicado al principio de la memoria, uno de los objetivos es realizar un estudio especifico de cómo afecta variables locales (tipo de población atendida y nivel económico) en la afluencia a los centros de salud.

Con una sentencia SQL insertamos en la Base de datos la información de la asistencia a los centros de salud, desglosada por día de la semana, tipo de día, mes y año. En la tabla 13 podemos ver un ejemplo de esta información:

FECHA	AÑO	MES	DIA DE LA SEMANA	TIPO DIA	NUMERO CASOS
20/10/2011	2011	10	DOMINGO	FESTIVO	125
15/01/2011	2011	01	VIERNES	LABORABLE	289
19/02/2011	2011	02	VIERNES	LABORABLE	229
18/08/2011	2011	08	LUNES	LABORABLE	353
29/07/2011	2011	07	MIÉRCOLES	LABORABLE	359
26/06/2011	2011	06	MIÉRCOLES	LABORABLE	331
06/03/2011	2011	03	DOMINGO	FESTIVO	119
03/03/2011	2011	03	LUNES	LABORABLE	401
15/01/2011	2011	01	VIERNES	LABORABLE	268
19/02/2011	2011	02	LUNES	LABORABLE	325

Tabla 13. Información de visitas a los centros de salud transformada para la generación de los modelos.

3.3.4.2 Datos Meteorológicos.

Los datos climáticos tenemos que transformarlos con una agrupación similar a la de los datos Estacionales. Además, como vimos en su análisis, tenemos muchos nulos en una de las estaciones, por tanto trabajaremos con la media de las dos estaciones.

En la tabla 14 tenemos un ejemplo de la información de partida:

Est	Día	T.Med	T.Máx	T.Mín	Prec.	Hum.R	V.Med
SIVA63	01/01/2007	9.3 ºC	14 ºC	6 ºC	0 mm	66.8 %	n/a
SIVA63	02/01/2007	8.8 ºC	14 ºC	4 ºC	0 mm	79%	n/a
SIVA63	03/01/2007	9.5 ºC	17 ºC	3 ºC	0.2 mm	50.7 %	n/a
SIVA63	04/01/2007	10 ºC	15 ºC	7 ºC	0 mm	47.2 %	n/a
SIVA63	05/01/2007	8.8 ºC	14 ºC	3 ºC	0.2 mm	69.1 %	n/a
SIVA63	06/01/2007	9.5 ºC	14 ºC	4 ºC	0.2 mm	66%	n/a
SIVA63	07/01/2007	8.9 ºC	14 ºC	4 ºC	0.2 mm	77.2 %	n/a
SIVA63	08/01/2007	7.7 ºC	11 ºC	4 ºC	0.2 mm	87.2 %	n/a
SIVA63	09/01/2007	8.1 ºC	15 ºC	3 ºC	0 mm	76.3 %	n/a
SIVA63	10/01/2007	7.9 ºC	n/a	n/a	n/a	73.8 %	n/a
SIVA63	11/01/2007	8.1 ºC	16 ºC	1 ºC	0 mm	56.1 %	n/a
SIVA63	12/01/2007	9.1 ºC	17 ºC	3 ºC	0 mm	42.2 %	n/a

Tabla 14. Ejemplo de la información meteorológica obtenida de REDIAM.

Sobre estos datos realizaremos las siguientes transformaciones:

i. Calculamos la media de la temperatura mínima, máxima, media, precipitaciones y humedad relativa de las dos estaciones agrupando por fecha.
ii. Agrupamos los datos por año, mes, día de la semana y tipo de día.

En la tabla 15 podemos ver un ejemplo del contenido de la tabla una vez realizadas esas transformaciones.

MES	DIA SEMANA	AÑO	TIPO DIA	TEMP MEDIA	TEMP MAXIMA	TEMP MINIMA	HUMEDAD RELATIVA	PRECIP
01	LUNES	2011	LABORABLE	9,38	13	4,2	79,42	0,4
01	MARTES	2011	LABORABLE	13,1	10,4	4,6	41,49	0
01	MIERCOLES	2011	LABORABLE	25,875	31,5	19,5	41,35	7,2
01	JUEVES	2011	LABORABLE	11,6	14,4	8,2	65,78	4,2
01	VIERNES	2011	LABORABLE	18,925	17,5	10,25	58,4	0,2
01	SABADO	2011	LABORABLE	19,32	24,4	13,8	31,6	0

Tabla 15. Ejemplo de la información meteorológica transformada.

3.3.4.3 Datos de Calidad Ambiental.

Los datos de Calidad del aire ya los hemos obtenido de la Web de REDIAM agrupados por meses, año y estación que ha recogido el dato. Al igual que con los datos Climáticos, tenemos que transformarlos calculando la media de días con calidad de aire: Muy Buena, Aceptable, Mala y Muy Mala.

Actualmente tenemos los datos de origen como se muestra en la tabla 16.

MES	AÑO	ESTACION DE RECOGIDA	NUMRO DIAS CALIDAD BUENA	NUM DIAS CALIDAD ADMISIBLE	NUM DIAS CALIDAD MALA	NUM DIAS CALIDAD MUY MALA
01	2007	LAS FUENTEZUELAS	21	10	0	0
01	2007	RONDA DEL VALLE	3	12	5	11
02	2007	LAS FUENTEZUELAS	4	24	0	0
02	2007	RONDA DEL VALLE	2	21	5	0
03	2007	LAS FUENTEZUELAS	0	31	0	0
03	2007	RONDA DEL VALLE	0	28	3	0
04	2007	LAS FUENTEZUELAS	0	26	4	1
04	2007	RONDA DEL VALLE	0	20	10	0
05	2007	LAS FUENTEZUELAS	0	23	8	31
05	2007	RONDA DEL VALLE	0	24	7	31

Tabla 16. Ejemplo de la información obtenida de REDIAM con la calidad del aire.

Tras realizar las transformaciones de los datos de calidad ambiental, la información se almacena como se muestra en la tabla 17.

MES	AÑO	NUMRO DIAS CALIDAD BUENA	NUM DIAS CALIDAD ADMISIBLE	NUM DIAS CALIDAD MALA	NUM DIAS CALIDAD MUY MALA
01	2010	4	26,5	0,5	0
06	2010	0,5	13,5	16	0
08	2007	0	12,5	13,5	5
08	2010	0	12,5	16	1
09	2007	2	18,5	9,5	0
05	2008	0	23	6	0,5
10	2008	4,5	23	2	1,5
04	2009	0	26	9	0
11	2009	1	24	4	1
04	2010	0	22	9,5	0

Tabla 17. Ejemplo de la información transformada de la calidad del aire.

3.3.4.4 Datos económicos

En los datos económicos obtenidos de las zonas de influencia de los centros de salud hemos encontrado importantes diferencias de unas zonas a otras, por ejemplo la zona de virgen de la capilla triplica el precio de los pisos de la zona de la magdalena.

Como hay diferencias muy importantes entre unos precios y otros, para suavizar el impacto de estos datos sobre el modelo predictivo aplicamos una transformación sobre la variable Precio Medio, aplicándole el logaritmo Neperiano. Esta es una técnica muy utilizada en minería de datos para suavizar valores cuyas distancias entre ellos son muy grandes y pueden generar ruido en el modelo. En la tabla 18 podemos ver el resultado y los datos que usaremos para generar nuestro modelo.

Centro de Salud	PRECIO MEDIO	LN
FEDERICO DEL CASTILLO	161000	11,98915964
BELEN	139125	11,84312809
VIRGEN DE LA CAPILLA	244750	12,40799256
FUENTEZUELAS	137000	11,8277362
SAN FELIPE	101500	11,52781408
EL VALLE	109250	11,60139411
LA MAGDALENA	94500	11,45635511

Tabla 18. Calculo de Logaritmo Neperiano del precio de un piso estándar para suavizar su impacto en el modelo.

3.4. Integración de la Información

Una vez limpiados los datos anómalos y agrupada la información extraída de las distintas fuentes, procedemos a integrar y relacionar todos los datos. Esta integración se realiza sobre una tabla, uniendo los datos por mes, día de la semana y año.

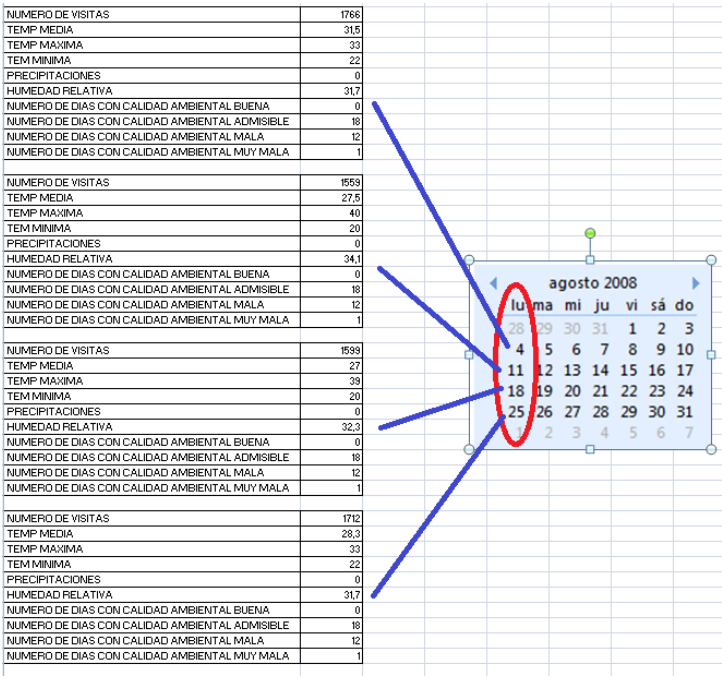
Tras realizar la integración de los datos, en la tabla 19 se muestra un ejemplo de cómo quedaría la información agregada con la que realizaremos el estudio de Data Mining.

MES	AÑO	Numero Visitas	DIA SEMANA	TIPO DIA	TEM MEDIA	TEM MAX	TEM MIN	PRECIP	HUME RELATIVA	NUMERO DE DIAS CON CALIDAD DEL AIRE BUENA	NUMERO DE DIAS CON CALIDAD DEL AIRE ADMISIBLE	NUMERO DE DIAS CON CALIDAD DEL AIRE MALA	NUMERO DE DIAS CON CALIDAD DEL AIRE MUY MALA	Número de Visitas	MEDIA VISITAS
2	2010	10607	L	LA	10,4	13,0	7,8	0,2	66,4	4,0	26,5	0,5	0,0	4	2651
3	2010	12819	M	LA	10,4	7,8	4,3	1,4	58,2	0,5	13,5	16,0	0,0	5	2563
7	2007	8912	S	FE	27,7	33,5	20,3	0,0	27,1	0,0	12,5	13,5	5,0	4	2228
1	2010	8451	M	LA	9,4	13,0	4,2	0,4	79,4	0,0	12,5	16,0	1,0	4	2112
1	2007	11941	M	FE	9,4	13,0	4,2	0,4	79,4	2,0	18,5	9,5	0,0	5	2388
3	2008	516	L	LA	13,1	10,4	4,6	0,0	41,5	0,0	23,0	6,0	0,5	4	129
7	2008	8891	X	LA	25,9	31,5	19,5	7,2	41,4	4,5	23,0	2,0	1,5	3	2963
8	2009	10869	V	LA	27,6	32,8	21,0	0,0	31,8	0,0	26,0	9,0	0,0	4	2717
8	2009	109	V	FE	27,6	32,8	21,0	0,0	31,8	1,0	24,0	4,0	1,0	1	109

Tabla 19. Ejemplo de la información integrada para la generación de los modelos.

Al agrupar la información por mes, año, día de la semana y tipo de día, es fundamental conocer cuántos días del mes cumple con esos criterios, para ello hemos creado el campo "numero_visitas" que almacena este dato. Si nos fijamos en la primera columna, el numero_visitas es igual a 4, esto quiere decir que en el mes 02 de de 2010 tenemos sumadas las visitas de 4 Lunes. En el caso de los datos meteorológicos, tenemos las medias de los 4 lunes. A continuación se muestra la figura 30 que nos ayuda a entender mejor la agrupación e integración que hemos realizado.

Nos centramos en los lunes de Agosto de 2008. La agrupación se realiza sobre los días 4, 11, 18 y 25, que son todos los lunes laborables de 2008. En la Figura 30 podemos ver de forma grafica la integración y agrupación de los datos.



NUMERO DE VISITAS	1768
TEMP MEDIA	31,5
TEMP MAXIMA	33
TEM MINIMA	22
PRECIPITACIONES	0
HUMEDAD RELATIVA	31,7
NUMERO DE DIAS CON CALIDAD AMBIENTAL BUENA	0
NUMERO DE DIAS CON CALIDAD AMBIENTAL ADMISIBLE	18
NUMERO DE DIAS CON CALIDAD AMBIENTAL MALA	12
NUMERO DE DIAS CON CALIDAD AMBIENTAL MUY MALA	1

NUMERO DE VISITAS	1559
TEMP MEDIA	27,5
TEMP MAXIMA	40
TEM MINIMA	20
PRECIPITACIONES	0
HUMEDAD RELATIVA	34,1
NUMERO DE DIAS CON CALIDAD AMBIENTAL BUENA	0
NUMERO DE DIAS CON CALIDAD AMBIENTAL ADMISIBLE	18
NUMERO DE DIAS CON CALIDAD AMBIENTAL MALA	12
NUMERO DE DIAS CON CALIDAD AMBIENTAL MUY MALA	1

NUMERO DE VISITAS	1599
TEMP MEDIA	27
TEMP MAXIMA	39
TEM MINIMA	20
PRECIPITACIONES	0
HUMEDAD RELATIVA	32,3
NUMERO DE DIAS CON CALIDAD AMBIENTAL BUENA	0
NUMERO DE DIAS CON CALIDAD AMBIENTAL ADMISIBLE	18
NUMERO DE DIAS CON CALIDAD AMBIENTAL MALA	12
NUMERO DE DIAS CON CALIDAD AMBIENTAL MUY MALA	1

NUMERO DE VISITAS	1712
TEMP MEDIA	28,3
TEMP MAXIMA	33
TEM MINIMA	22
PRECIPITACIONES	0
HUMEDAD RELATIVA	31,7
NUMERO DE DIAS CON CALIDAD AMBIENTAL BUENA	0
NUMERO DE DIAS CON CALIDAD AMBIENTAL ADMISIBLE	18
NUMERO DE DIAS CON CALIDAD AMBIENTAL MALA	12
NUMERO DE DIAS CON CALIDAD AMBIENTAL MUY MALA	1

Figura 30. Representación grafica del modelado de un día tipo.

En la tabla 20 puede verse los datos completos de los 4 lunes y en la última fila tenemos calculadas las medias de todos los valores. Estos datos son los que usaremos para modelar los lunes laborables de Agosto del año 2008.

Fecha	Numero Visitas	tem media	tem máxima	tem mínima	precip	húmeda relativa	días buenos	admisible	malos	Muy Malos
04/08/2008	1766	31,5	33	22	0	31,7	0	18	12	1
18/08/2008	1559	27,5	40	20	0	34,1	0	18	12	1
25/08/2008	1599	27	39	20	0	32,3	0	18	12	1
11/08/2008	1712	28,3	33	22	0	31,7	0	18	12	1
MEDIAS DE LOS LUNES DE AGOSTO	**1659**	**28,575**	**36,25**	**21**	**0**	**32,45**	**0**	**18**	**12**	**1**

Tabla 20. Ejemplo del cálculo de las medias para modelar un día estándar.

Una vez realizadas todos las agrupaciones en la tabla 21 vemos como quedaría el registro completo agrupado y agregado.

MES	AÑO	Numero total de Visitas	DIA SEMANA	TIPO DIA	TEM MEDIA	TEM MAX	TEM MIN	PRECIP	HUME RELATIVA	NUM D CALIDAD BUENA	DIAS CON CALIDAD DEL AIRE ADMISIBLE	DIAS CON CALIDAD DEL AIRE MALA	DIAS CON CALIDAD DEL AIRE MUY MALA	Número de Visitas	MEDIA VISITAS
8	2008	6636	LUN	LAB	28,57	36,25	21	0	32,45	0	18	12	0	1	1659

Tabla 21. Ejemplo de los datos de un día modelado (un lunes de agosto del año 2008).

Este proceso lo realizamos para todos los datos desde el año 2007 hasta el año 2011.

4. RESULTADOS SIN INFORMACIÓN ESPACIAL

Antes de presentar los resultados del estudio es importante recordar que todos los modelos predictivos se generan con datos de 2007 hasta 2010 y usaremos los datos del año 2011 para validar la eficiencia del modelo. De todos los modelos que se generen en el estudio se realizará un análisis de su eficiencia desde el punto de vista teórico (con datos estadísticos teóricos obtenidos de los modelos) como practico comprobando el error real en la predicción de pacientes de año 2011, comparando el error absoluto de cada predicción y el dato real de visitas de todos los días del año 2011.

Desde el punto de vista teórico se medirán de los modelos las siguientes variables:

- **Porcentaje del nivel de confianza predictiva**.- Oracle Data Mining calcula la confianza de predicción de los modelos de regresión. La confianza predictiva es una medida de la mejora obtenida por el modelo sobre una estimación realizada al azar. La confianza predictiva es el porcentaje de incremento obtenido por el modelo en un modelo ingenuo, es decir que si el modelo indica una confianza predictiva del 43%, quiere decir que el modelo es un 43% mejor que un modelo ingenuo.

- **Error de promedio absoluto.-** El Error medio cuadrático y el error absoluto medio se utilizan comúnmente para evaluar la calidad global de un modelo de regresión.

- **El error absoluto medio (MAE).-** Es el promedio del valor absoluto de los residuos (error). El MAE es muy similar al Error medio cuadrático (RMSE) pero es menos sensible a los errores grandes.

$$\mathbf{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- **Error medio Cuadrático.-** Es la raíz cuadrada de la distancia al cuadrado del promedio de un punto de datos de la línea ajustada.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- **Grafico de valores Residuales**.- Un grafico de valores residuales es un diagrama de dispersión en el que el eje "x" es el valor predicho de "x", y el eje "y" es el residual para "x". El residual es la diferencia entre el valor real de "x" y el valor predicho de "x".

En la comprobación de la eficiencia del modelo con respecto a los datos reales del año 2011, mostraremos una tabla que contendrá los siguientes datos:

- **Mes.-** mes del día a predecir.
- **Tipo de día.-** Se caracteriza el día por Laborable y Festivo
- **Día de la Semana.-** Día de la semana (Lunes a Domingo).
- **Número de Visitas (este es el atributo target).-** Media de Visitas realizadas en uno de los día caracterizado.
- **GLM.-** Valor de la Predicción con el algoritmo GLM.
- **ERROR ABSOLUTO GLM.-** Error absoluto cometido en la predicción del día caracterizado, comparando el valor real con la predicción.
- **SVM LINEAL.-** Valor de la Predicción con el algoritmo SVM con Kernel Lineal.
- **ERROR ABSOLUTO SVM LINEAL.-** Error absoluto cometido en la predicción del día caracterizado.

- **SVM GAUSIANO.-** Valor de la Predicción con el algoritmo SVM con Kernel Gausiano.
- **ERROR ABSOLUTO SVM GAUSIANO.-** Error absoluto cometido en la predicción del día caracterizado.

En la tabla 22 puede verse un ejemplo de la información que contendrá la comparativa del dato real de 2011 y la predicción de los distintos algoritmos.

DATOS REALES DE 2011				PREVISION					
MES	TIPO DIA	DIA SEMANA	VISITAS REALES	GML	ERROR ABSOLUTO GML	SVM GUISIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA
01	LABO	MARTES	1659	104	16	98	21	102	18
01	LABO	VIERNES	86	87	1	75	11	85	0

Tabla 22. Tabla ejemplo de la presentación de resultados.

En la Figura 31 puede observar de forma grafica el volcado de información de origen con respecto a la tabla de resultados. El dato más importante es el número de vistas, que realmente es la Media de Visitas (en rojo) del día caracterizado. En el ejemplo estamos calculando la media de visitas realizadas los martes Laborables del mes de enero de 2011.

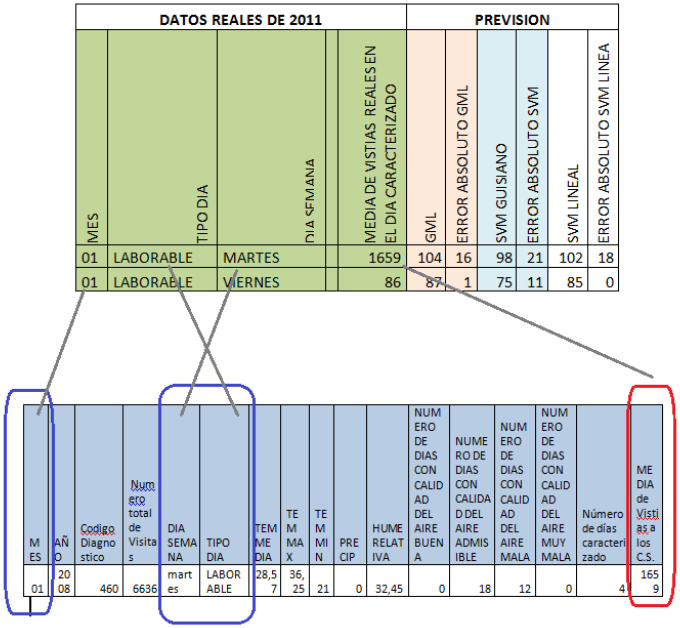


Figura 31. Representación grafica de la obtención de los datos para la generación de los modelos.

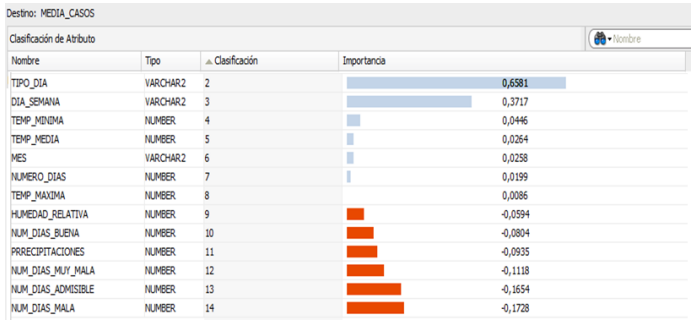
4.1. Modelo global de visitas de pacientes de Jaén

A continuación presentamos los resultados de los modelos generados para predecir de forma global el número de pacientes que necesitaran asistencia médica en atención primaria en Jaén, es importante recordar que en este modelo general no distinguimos entre centro de salud ni entre tipo de demanda (clínica o administrativa).

En los resultados se presentará los datos de importancia de atributos, resultados teóricos y prácticos de los modelos generados y finalmente los coeficientes estandarizados de los modelos.

4.1.1. Importancia de los atributos

Antes de generar nuestro modelo verificamos la importancia de nuestros atributos con el algoritmo: MINIMUN DESCRIPTION LENGTH. En la figura 32 podemos observar los resultados:



Destino: MEDIA_CASOS

Clasificación de Atributo

Nombre	Tipo	Clasificación	Importancia
TIPO_DIA	VARCHAR2	2	0,6581
DIA_SEMANA	VARCHAR2	3	0,3717
TEMP_MINIMA	NUMBER	4	0,0446
TEMP_MEDIA	NUMBER	5	0,0264
MES	VARCHAR2	6	0,0258
NUMERO_DIAS	NUMBER	7	0,0199
TEMP_MAXIMA	NUMBER	8	0,0086
HUMEDAD_RELATIVA	NUMBER	9	-0,0594
NUM_DIAS_BUENA	NUMBER	10	-0,0804
PRRECIPITACIONES	NUMBER	11	-0,0935
NUM_DIAS_MUY_MALA	NUMBER	12	-0,1118
NUM_DIAS_ADMISIBLE	NUMBER	13	-0,1654
NUM_DIAS_MALA	NUMBER	14	-0,1728

Figura 32. Distribución de la importancia de atributos.

En este ranking podemos ver como las variables con más peso son las que esperábamos: Tipo de Día y Día de la semana, temperatura mínima, temperatura máxima, mes y temperatura máxima.

Por otra parte vemos que hay atributos con valor negativo, lo cual quiere decir que no se relacionan con la variable objetivo, para evitar ruido con estas variables son eliminamos del estudio. Esto supone eliminar algunas variables meteorología como: Humedad Relativa y precipitaciones y todas las variables de calidad ambiental.

4.1.2. Comparativa de los Modelos desde un punto de vista teórico

En este punto presentamos los resultados de los principales parámetros teóricos de los modelos probados. Como puede observarse en la tabla 23 desde el punto de vista teórico el algoritmo más eficiente es SVM con Kernel Gaussiano.

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático	Valor Promedio Previsto	Valor promedio Real
SVM (kernel Lineal)	85,38	91,2	160,73	1446	1460
SVM (Kernel Gaussiano)	91,93	65,67	89,79	1461	1460
GLM	86,17	113,34	151,99	1478	1460

Tabla 23. Resultados teóricos de los modelos.

En la Figura 33 presentamos el grafico de residuos, como se puede comprobar tenemos muchos residuos al principio de la grafica que son los puntos de los días festivos, lo que es indicativo de que el modelo no predecirá bien la afluencia de pacientes los días festivos y fines de semana.

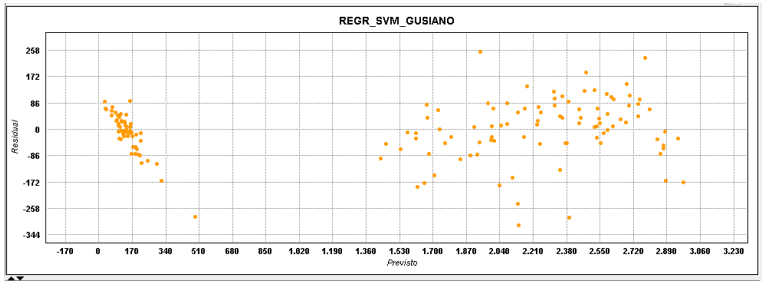


Figura 33. Grafico de los residuos del modelo.

4.1.3. Comparativa del modelo aplicado sobre datos reales de 2011

Como se comenta en el planteamiento del problema, los modelos predictivos se han generados con los datos de 2007, 2008, 2009 y 2010. Usaremos el 2011 para testear los 3 algoritmos. En la tabla 24 tenemos los resultados de la ejecución de los modelos sobre el año 2011. En la tabla se presenta el dato real de afluencia a los centros de Salud, la predicción de cada modelo y el error absoluto cometido por cada modelo.

DATOS REALES DE 2011				PREVISION					
MES	TIPO DIA	DIA SEMANA	DATO REAL	GML	ERROR ABSOLUTO GML	SVM GAUSIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA

1	FESTIVO	SÁBADO	184	270	86	185	1	198	14
1	FESTIVO	DOMINGO	150	-41	191	102	48	-251	401
1	LABORABLE	MIÉRCOLES	2554	2564	10	2445	109	2527	27
1	LABORABLE	VIERNES	2036	2115	79	2067	30	2059	23
1	LABORABLE	LUNES	2657	2772	114	2568	90	2728	71
1	LABORABLE	MARTES	2553	2649	96	2518	35	2579	26
1	LABORABLE	JUEVES	2687	2489	198	2632	54	2531	155
1	FESTIVO	JUEVES	156	224	68	173	17	154	2
2	LABORABLE	LUNES	3054	2623	431	2702	351	2767	287
2	FESTIVO	SÁBADO	173	177	5	201	28	186	14
2	LABORABLE	JUEVES	2731	2553	178	2597	134	2675	56
2	LABORABLE	MIÉRCOLES	2838	2656	182	2554	284	2757	81
2	LABORABLE	MARTES	2950	2796	154	2747	203	2903	47
2	FESTIVO	LUNES	198	161	37	180	18	161	37
2	LABORABLE	VIERNES	2255	2160	95	2156	99	2240	15
2	FESTIVO	DOMINGO	177	197	20	167	10	190	12
3	LABORABLE	VIERNES	2269	2195	73	2179	90	2224	45
3	LABORABLE	LUNES	2855	2742	113	2835	20	2780	75
3	LABORABLE	MIÉRCOLES	2943	2909	34	2748	196	2994	51
3	FESTIVO	SÁBADO	173	198	25	163	9	170	3
3	FESTIVO	DOMINGO	160	224	64	152	8	165	5
3	LABORABLE	MARTES	2888	2941	53	2783	104	2989	101
3	LABORABLE	JUEVES	2621	2625	4	2566	55	2659	38
4	LABORABLE	MARTES	2708	2705	3	2781	73	2715	7
4	LABORABLE	VIERNES	2224	2212	13	2242	18	2224	0
4	FESTIVO	SÁBADO	149	-120	269	135	13	-258	407

4	FESTIVO	DOMINGO	148	232	85	163	15	175	27
4	FESTIVO	VIERNES	103	27	76	116	13	82	21
4	LABORABLE	JUEVES	2780	2528	252	2701	80	2604	177
4	LABORABLE	LUNES	2831	2736	95	2799	32	2774	57
4	LABORABLE	MIÉRCOLES	2605	2577	28	2682	77	2576	29
5	FESTIVO	SÁBADO	110	243	133	183	73	159	49
5	FESTIVO	DOMINGO	123	-24	147	119	4	-255	378
5	LABORABLE	LUNES	3063	2967	96	3008	55	2992	71
5	FESTIVO	LUNES	105	221	116	228	123	137	32
5	LABORABLE	JUEVES	2900	2793	107	2730	169	2828	71
5	LABORABLE	VIERNES	2215	2246	32	2213	2	2219	4
5	LABORABLE	MARTES	2927	3100	174	2867	59	3073	147
5	LABORABLE	MIÉRCOLES	2845	2774	71	2759	86	2778	67
6	LABORABLE	MIÉRCOLES	2448	2469	21	2421	27	2470	22
6	LABORABLE	LUNES	2676	2580	97	2616	61	2623	53
6	FESTIVO	SÁBADO	142	122	19	136	5	129	13
6	LABORABLE	JUEVES	2411	2412	1	2324	87	2433	22
6	LABORABLE	MARTES	2529	2523	7	2549	19	2536	7
6	FESTIVO	DOMINGO	133	175	42	177	44	135	2
6	LABORABLE	VIERNES	2014	1982	32	2001	13	2007	7
7	LABORABLE	MIÉRCOLES	1684	1802	118	1667	17	1745	62
7	LABORABLE	JUEVES	1741	1801	60	1770	29	1795	54
7	FESTIVO	DOMINGO	125	-244	370	63	63	-327	452
7	LABORABLE	LUNES	1917	1972	55	1935	18	1954	37
7	LABORABLE	MARTES	1869	1996	127	1928	59	1955	86
7	FESTIVO	SÁBADO	152	-249	400	107	45	-297	449

7	LABORABLE	VIERNES	1444	1437	7	1525	81	1360	84
8	LABORABLE	MARTES	1600	1728	127	1737	136	1567	33
8	LABORABLE	JUEVES	1426	1563	137	1485	59	1488	62
8	LABORABLE	VIERNES	1284	1385	102	1391	108	1329	45
8	LABORABLE	MIÉRCOLES	1583	1638	56	1634	52	1489	93
8	FESTIVO	DOMINGO	112	36	76	75	37	46	66
8	LABORABLE	LUNES	1725	1817	93	1670	55	1744	19
8	FESTIVO	SÁBADO	117	6	112	92	25	46	71
8	FESTIVO	LUNES	103	63	40	194	91	61	42
9	LABORABLE	MIÉRCOLES	2150	2198	48	2091	58	2150	1
9	LABORABLE	VIERNES	1785	1788	3	1800	15	1717	68
9	LABORABLE	JUEVES	2033	2081	48	2051	18	2022	11
9	LABORABLE	MARTES	2194	2311	117	2170	24	2244	50
9	LABORABLE	LUNES	2326	2342	15	2283	43	2310	16
9	FESTIVO	SÁBADO	129	92	37	126	3	101	28
9	FESTIVO	DOMINGO	148	158	9	136	12	124	24
10	LABORABLE	LUNES	2665	2850	185	2791	125	2744	78
10	FESTIVO	DOMINGO	124	16	108	158	33	-273	398
10	LABORABLE	MIÉRCOLES	2643	2556	87	2787	144	2511	132
10	FESTIVO	MARTES	136	363	227	190	54	196	60
10	LABORABLE	MARTES	2834	2741	93	2920	86	2686	148
10	FESTIVO	SÁBADO	161	-6	166	176	15	-238	399
10	FESTIVO	MIÉRCOLES	130	342	212	329	199	158	28
10	LABORABLE	JUEVES	2741	2731	10	2801	60	2692	49
10	LABORABLE	VIERNES	2224	2301	77	2315	91	2222	3
11	FESTIVO	DOMINGO	145	495	350	149	4	591	446

11	LABORABLE	LUNES	2905	2755	149	2816	89	2830	75
11	LABORABLE	MIÉRCOLES	2645	2678	33	2599	46	2701	57
11	FESTIVO	MARTES	104	274	170	119	15	193	89
11	LABORABLE	JUEVES	2588	2504	84	2567	22	2559	29
11	FESTIVO	SÁBADO	142	485	343	205	64	595	454
11	LABORABLE	VIERNES	1658	1806	148	1888	230	1742	84
11	LABORABLE	MARTES	2816	2745	71	2670	146	2796	21
12	FESTIVO	MARTES	161	161	0	153	8	160	1
12	LABORABLE	JUEVES	2252	2210	42	2203	49	2246	6
12	LABORABLE	MARTES	2493	2371	123	2505	12	2432	61
12	FESTIVO	JUEVES	137	89	48	107	30	109	28
12	FESTIVO	LUNES	191	140	51	133	58	132	59
12	FESTIVO	DOMINGO	163	141	22	138	24	149	13
12	LABORABLE	LUNES	2519	2326	193	2386	133	2407	112
12	FESTIVO	SÁBADO	174	-176	350	23	151	-259	433
12	LABORABLE	VIERNES	1871	1852	19	1673	198	1824	47
12	LABORABLE	MIÉRCOLES	2192	2202	10	2223	31	2197	5

Tabla 24. Aplicación de los modelos y comparativa sobre el año 2011.

Al analizar los datos vemos como en general el modelo que hace la predicción más exacta es el SVM con Kernel Gausiano. El porcentaje absoluto de error es del 4,26%. En la tabla 25 podemos ver un resumen de los resultados.

Algoritmo	% DE ERROR ABOSOLUTO	Error absoluto medio en cada predicción
GLM	6,59446157	102,673684
SVM con Kernel Lineal	5,69527827	88,6736842
SVM con Kernel Gausiano	4,26334577	66,3789474

Tabla 25. Resumen del error real cometido por los modelos en la predicción del año 2011.

A continuación podemos visualizar en la Figura 34 las graficas con la comparativa del dato real de pacientes que necesitaron atención en 2011 frente a la previsión del algoritmo SVM con Kernel Gaussiano.

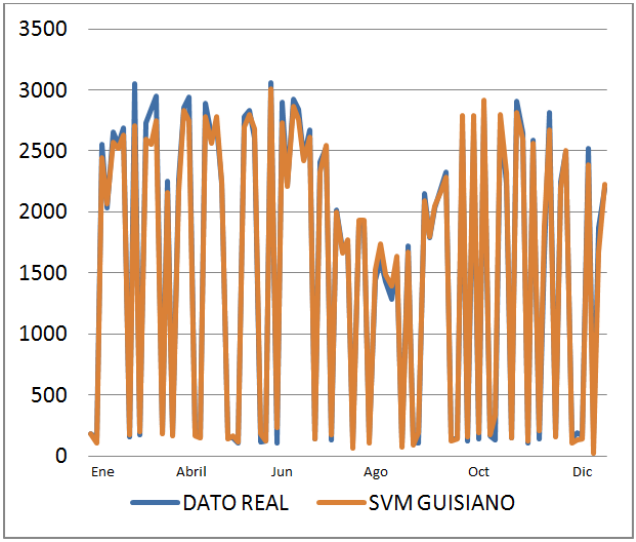


Figura 34. Comparativa del dato real y la predicción del modelo.

Si analizamos en profundidad los errores, estos se producen en mayor medida en los días Festivos, tal y como se podía interpretar del diagrama de residuos. El porcentaje de error absoluto en días festivos es del 27%.

En los días festivos sólo se atiende Urgencias y el número de visitas aparte de ser mucho menor que los días laborables, no tienen tanta estacionalidad en los datos, es decir, no se repiten tanto los datos en función del mes, día, etc. En la figura 35 podemos ver de forma gráfica las visitas de los días festivos durante el año 2011:

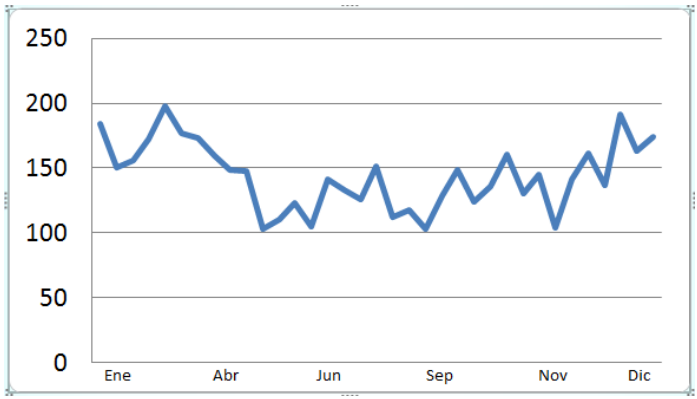


Figura 35. Distribución de afluencia de pacientes durante el año 2011 en días festivos y fines de semana.

Puede observarse cómo los valores varían prácticamente entre 100 y 190 visitas. En la figura 36 comparamos los datos reales de los festivos del 2011 con la predicción del modelo, podemos ver cómo los modelos lineales tienen un comportamiento muy inestable. Sin embargo el modelo generado con SVM Gausiano se comporta mucho mejor.

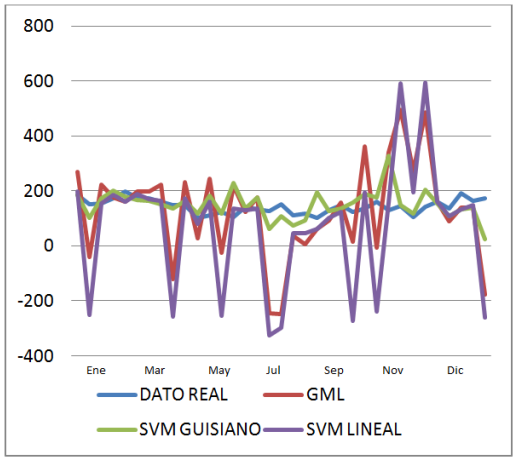


Figura 36. Comparativa grafica de la predicción de los modelos y el número de visitas reales del año 2011.

Sin embargo al comparar el modelo para los días laborables, los tres modelos son muy precisos. En la Figura 37 podemos ver la comparativa.

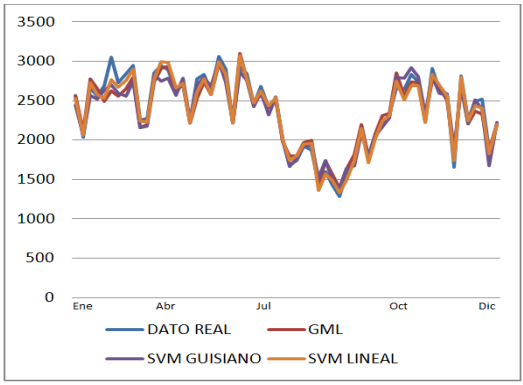


Figura 37. Comparativa grafica de la predicción de los modelos y el número de visitas reales del año 2011.

En el caso de los días Laborables el algoritmo que tiene menor error absoluto es el SVM con Kernel Lineal. El error total es de 3.469 pacientes en total durante todo el año frente a los 4.949 y 5.275 del SVM con Kernel Gausiano y GLM respectivamente.

Por tanto para realizar la mejor predicción sería necesario utilizar el algoritmo SVM con Kernel Gausiano para la predicción de los días Festivos y el Algoritmo SVM con kernel Lineal para los días Laborables.

El error cometido para los días festivos es muy alto. Si analizamos los datos con mayor profundidad y graficamos las visitas desde el año 2007 al 2010 agrupadas por meses (Figura 38) vemos claramente que hay dos conjunto de datos independientes: uno con los días Festivos (donde sólo se atienden urgencias) y otro con los días Laborables donde se atienden demanda clínica y administrativa (emisión de recetas, emisión de partes de altas/bajas, revisiones periódicas, etc.).

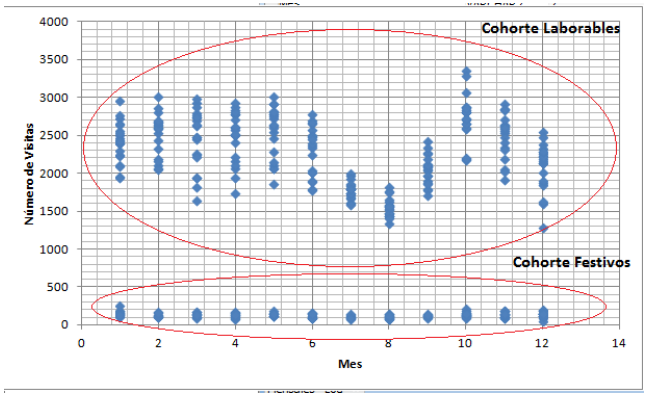


Figura 38. Representación grafica de las visitas de pacientes por meses, tanto en días festivos como laborables.

En los días Laborables puede observarse cómo hay mucha estacionalidad. En cuanto a los días Festivos puede verse cómo el comportamiento es muy lineal durante todo el año.

Una mejora del sistema sería crear dos cohortes independientes de datos y realizar el estudio por separado.

A partir de ahora haremos referencia a tres cohortes:

- **Cohorte Completa.-** Es el conjunto de datos que hemos estudiado anteriormente y que contienen toda la información: Festivos y Laborables.

- **Cohorte Festivos.-** Es el conjunto de datos que sólo contiene la información de los días Festivos. Esta cohorte contiene datos de Sábados, Domingos y Festivos. Estos días solo se atienden Urgencias en los Centros de Salud.

- **Cohorte Laborables.-** Es el conjunto de datos que sólo contiene la información de los días Laborables.

4.1.4. Mejora del Modelo para los días Festivos

Una vez creada las dos cohortes nuevas independientes para generar los nuevos modelos, volvemos a aplicar las técnicas de minería de datos siguiendo la misma metodología de trabajo.

4.1.4.1 Importancia de Atributos para los días festivos

Al cambiar las cohortes, volvemos a calcular nuevamente el peso de los atributos con el algoritmo: MINIMUN DESCRIPTION LENGTH. En la Figura 39 podemos ver los resultados.

Destino: MEDIA_CASOS

Clasificación de Atributo

Nombre	Tipo	Clasificación	Importancia
MES	VARCHAR2	2	0,3183
TEMP_MINIMA	NUMBER	3	0,2842
TEMP_MEDIA	NUMBER	4	0,2546
TEMP_MAXIMA	NUMBER	5	0,2343
HUMEDAD_RELATIVA	NUMBER	6	0,0945
NUM_DIAS_MALA	NUMBER	7	0,0304
TIPO_DIA	VARCHAR2	8	0
DIA_SEMANA	VARCHAR2	9	-0,0296
PRRECIPITACIONES	NUMBER	10	-0,0448
NUM_DIAS_ADMISIBLE	NUMBER	11	-0,0581
NUM_DIAS_BUENA	NUMBER	12	-0,0757
NUMERO_DIAS	NUMBER	13	-0,0891
NUM_DIAS_MUY_MALA	NUMBER	14	-0,1014

Figura 39. Representación grafica de la importancia de atributos.

En este ranking podemos ver como las variables con más peso ahora son: mes, temperatura mínima, temperatura media, temperatura máxima, Humedad relativa y número de días con mala calidad ambiental, aunque el peso de importancia de este último atributo es muy bajo, sólo un 0,03.

Hay que destacar que al separar las cohortes en dos independientes, aparecen nuevos atributos que están relacionados con el número de vistas al médico, como son Humedad relativa y la calidad del aire. En teoría estos dos atributos deberían tener una relación directa ya que como hemos visto en el estado del arte hay muchas patologías que tienen una relación con dichas variables y en teoría deben desencadenar en un amento o disminución de visitas a los centros de Salud.

4.1.4.2 Comparativa de los Modelos desde un punto de vista teórico (Cohorte Festivos)

Desde un punto de vista teórico, tras estudiar los parámetros de los modelos, el más eficiente es SVM con Kernel Lineal (tabla 26).

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático	Valor Promedio Previsto	Valor promedio Real
SVM (kernel Lineal)	39,04	12,32	21,81	132	131
SVM (Kernel Gaussiano)	2,91	26,27	34,75	132	132
GLM	35,53	15,75	23	131	132

Tabla 26. Resultados teóricos de la eficacia de los modelos.

En el grafico de valores residuales (Figura 40) observamos que existen muy pocos residuos y casi todos los valores convergen en el centro, lo cual quiere decir que el modelo se ajusta bien a la realidad.

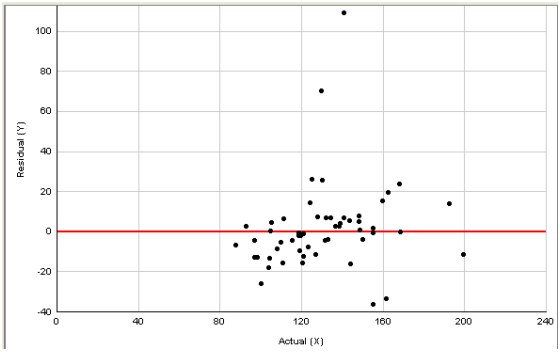


Figura 40. Representación grafica de los resididos del modelo.

4.1.4.3 Comparativa del Modelo con datos reales de 2011 (Cohorte Festivos)

En este caso el mejor algoritmo es el SVM con Kernel Lineal. Esto se confirma al comprar la predicción con el dato real de visitas del año 2011. En la tabla 27 podemos ver los resultados.

DATOS REALES DE 2011				PREVISION					
MES	TIPO DIA	DIA SEMANA	DATO REAL	GML	ERROR ABSOLUTO GML	SVM GAUSIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA

1	FESTI	SÁBADO	184	196	11	155	29	186	2
1	FEST	DOMIN	150	168	18	130	20	164	14
1	FEST	JUEVES	156	136	20	141	15	133	23
2	FEST	SÁBADO	173	182	9	149	23	169	3
2	FEST	LUNES	198	170	28	150	48	163	35
2	FEST	DOMIN	177	193	16	133	45	170	8
3	FEST	SÁBADO	173	174	1	161	12	162	10
3	FEST	DOMIN	160	163	3	121	39	150	10
4	FEST	SÁBADO	149	157	8	159	10	155	7
4	FEST	DOMIN	148	154	6	131	16	146	2
4	FEST	VIERNES	103	121	18	115	12	116	13
5	FEST	SÁBADO	110	89	21	138	28	102	8
5	FEST	DOMIN	123	111	12	134	11	122	1
5	FEST	LUNES	105	123	18	146	41	132	27
6	FEST	SÁBADO	142	141	1	142	1	144	2
6	FEST	DOMIN	133	133	1	135	2	135	3
7	FEST	DOMIN	125	141	16	115	11	127	2
7	FEST	SÁBADO	152	177	26	141	10	157	5
8	FEST	DOMIN	112	118	6	111	1	112	1
8	FEST	SÁBADO	117	120	2	110	7	119	2
8	FEST	LUNES	103	140	37	138	35	140	37
9	FEST	SÁBADO	129	131	2	127	2	129	0
9	FEST	DOMIN	148	156	8	128	21	143	5
10	FEST	DOMIN	124	112	12	133	8	125	1
10	FEST	MARTES	136	122	14	113	23	115	21
10	FEST	SÁBADO	161	167	6	163	2	165	4
10	FEST	MIÉRCO	130	100	30	106	24	111	19
11	FEST	DOMIN	145	148	3	115	30	137	8
11	FEST	MARTES	104	123	19	130	26	115	11

11	FEST	SÁBADO	142	139	2	163	21	138	3
12	FEST	MARTES	161	137	24	140	21	124	37
12	FEST	JUEVES	137	120	17	128	9	121	16
12	FEST	LUNES	191	157	34	143	48	162	29
12	FEST	DOMIN	163	170	8	139	24	159	4
12	FEST	SÁBADO	174	192	18	164	10	184	10

Tabla 27. Aplicación de los modelos y comparativa sobre el año 2011.

En este caso el mejor algoritmo es SVM con Kernel Lineal, este algoritmo comete un error absoluto del 7,6%. En la tabla 28 podemos ver un resumen de los resultados de los tres algoritmos.

Algoritmo	% DE ERROR ABOSOLUTO	Error absoluto medio en cada predicción
GLM	9,42834458	13,5714286
SVM con Kernel Lineal	7,6022231	10,9428571
SVM con Kernel Gausiano	13,5966653	19,5714286

Tabla 28. Resumen del error real cometido por los modelos en la predicción del año 2011.

Si observamos la figura 41 que contiene una grafica con la comparativa del dato real del año 2011 y la predicción de los modelos, observaremos como al generar un modelo especifico con la cohorte de datos de los días Festivos, los algoritmos modelan mucho mejor la realidad de las visitas en días festivos. El error para los días festivos baja de 1.360 pacientes en total para todos los días festivos y fines de semana del año a 385 pacientes.

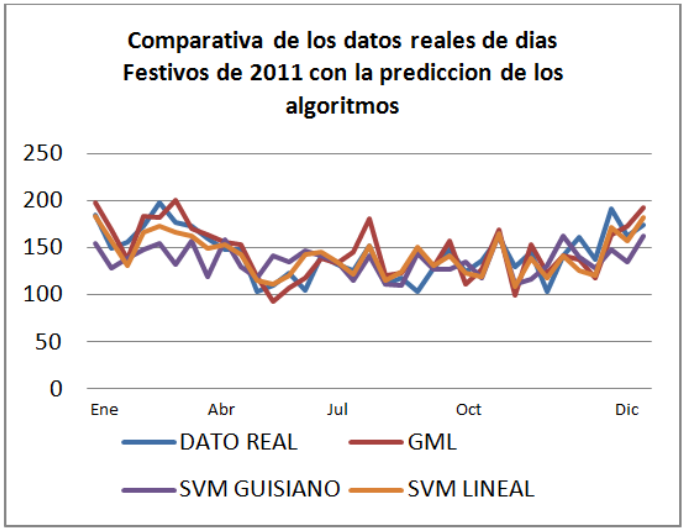


Figura 41. Comparativa grafica de la predicción de los modelos y el número de visitas reales del año 2011 en días Festivos.

La división de los datos en dos cohortes ha mejorado sustancialmente la predicción de los días festivos. En la figura 42 podemos ver de forma grafica la diferencia de la predicción antigua realizada con el algoritmo SVM con Kernel Gaussiano, frente a la actual realizada con el SVM con Kernel Lineal.

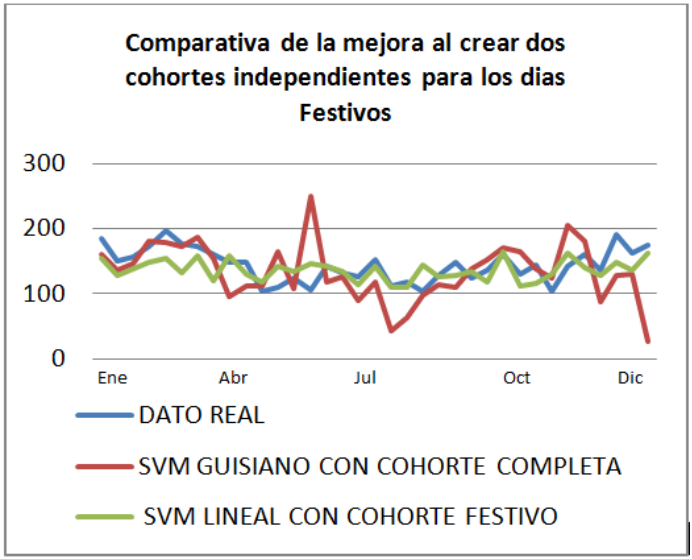


Figura 42. Comparativa grafica de la predicción del modelo con la cohorte completa, la cohorte de días festivos y el número de visitas reales del año 2011 en días festivos.

4.1.5. Mejora del Modelo para los días Laborables

De la misma manera que en el punto anterior ahora generamos nuevamente el modelo para los días laborables a partir de la cohorte de datos de los días laborables.

4.1.5.1 Importancia de Atributos para los días festivos

En cuanto a la importancia de atributos es prácticamente igual que la de los días laborables. En el ranking las variables con más peso son: mes, temperatura mínima, temperatura media, temperatura máxima, Humedad relativa y número de días con mala calidad ambiental.

4.1.5.2 Comparativa de los Modelos desde un punto de vista teórico (Cohorte laborables)

Desde un punto de vista teórico, tras estudiar los parámetros de los modelos, el más eficiente es GLM con una confianza predictiva del 85,17% y el error absoluto es de 49,09. Estos datos indican que sin duda es el mejor algoritmo para modelar los días laborables. En la tabla 29 podemos ver un resumen de los resultados de los tres algoritmos.

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático	Valor Promedio Previsto	Valor promedio Real
SVM (kernel Lineal)	83,2	52	149,3	2312	2306
SVM (Kernel Gaussiano)	68,75	102	140,67	2312	2303
GLM	85,45	49,09	66,17	2312	2308

Tabla 29. Resultados teóricos de la eficacia de los modelos.

Al analizar el grafico de residuos (figura 43), también podemos observar la eficacia del modelo GLM, confirmando que la distancia de los residuos es pequeña y convergen en el centro.

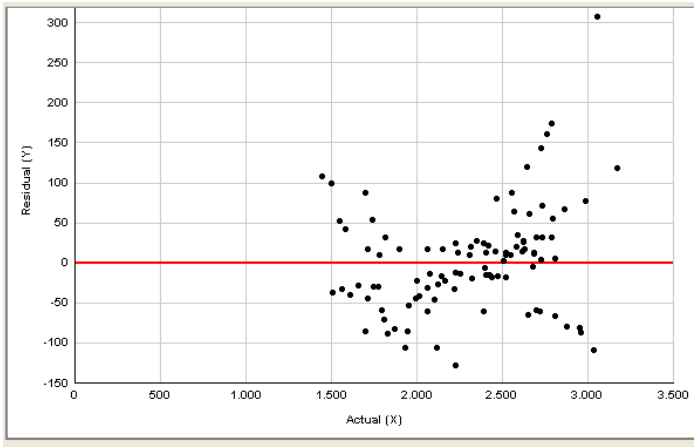


Figura 43. Representación grafica de los resididos del modelo.

4.1.5.3 Comparativa del Modelo con datos reales de 2011 (Cohorte Laborable)

En la comparación de los datos reales de visitas al médico en el año 2011 con respecto a la predicción realizada por el algoritmo GLM (generado con la ***cohorte Laborables)*** se observa también una importante mejora. Con el algoritmo GLM la suma total de errores en la predicción de 2011 baja de 3.469 generado por el algoritmo SVM con Kernel lineal sobre la ***cohorte completa*** a 3018. Esto supone un error de absoluto en la predicción del 2011 del 2,11% y una error medio absoluto en la predicción en un día de 50 pacientes de un total un total de 2.381.

En la tabla 30 podemos ver los datos reales del año 2011 y las predicciones de los modelos.

DATOS REALES DE 2011				PREVISION					
MES	TIPO DIA	DIA SEMANA	DATO REAL	GML	ERROR ABSOLUTO GML	SVM GAUSIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA
1	LABO	MIÉRCOLES	2554	2534	20	2509	45	2544	10
1	LABO	VIERNES	2036	2050	14	2064	27	2068	32
1	LABO	LUNES	2657	2732	75	2581	76	2738	80
1	LABO	MARTES	2553	2541	12	2607	53	2562	9
1	LABO	JUEVES	2687	2598	88	2378	308	2569	118
2	LABO	LUNES	3054	2889	164	2719	335	2849	205
2	LABO	JUEVES	2731	2735	4	2600	131	2735	4
2	LABO	MIÉRCOLES	2838	2816	23	2721	117	2810	29
2	LABO	MARTES	2950	2915	35	2864	86	2915	35
2	LABO	VIERNES	2255	2282	27	2187	68	2284	29
3	LABO	VIERNES	2269	2253	15	2254	15	2254	14
3	LABO	LUNES	2855	2819	36	2848	7	2829	26
3	LABO	MIÉRCOLES	2943	3028	85	2600	343	2978	34
3	LABO	MARTES	2888	2973	85	2630	258	2953	65
3	LABO	JUEVES	2621	2677	56	2344	277	2661	40
4	LABO	MARTES	2708	2672	36	2692	16	2676	32
4	LABO	VIERNES	2224	2211	13	2179	45	2215	9
4	LABO	JUEVES	2780	2636	144	2410	370	2601	180

4	LABO	LUNES	2831	2782	49	2774	57	2794	37
4	LABO	MIÉRCOLES	2605	2568	37	2531	74	2568	37
5	LABO	LUNES	3063	3001	62	2990	73	3022	41
5	LABO	JUEVES	2900	2842	57	2753	147	2855	45
5	LABO	VIERNES	2215	2205	9	2302	87	2225	11
5	LABO	MARTES	2927	3009	83	2642	284	3006	80
5	LABO	MIÉRCOLES	2845	2762	83	2814	31	2780	65
6	LABO	MIÉRCOLES	2448	2441	7	2452	4	2425	23
6	LABO	LUNES	2676	2614	62	2626	51	2628	48
6	LABO	JUEVES	2411	2404	7	2371	40	2397	14
6	LABO	MARTES	2529	2477	52	2517	12	2487	43
6	LABO	VIERNES	2014	1999	14	2075	61	2005	9
7	LABO	MIÉRCOLES	1684	1724	40	1800	116	1743	60
7	LABO	JUEVES	1741	1782	41	1793	52	1804	63
7	LABO	LUNES	1917	1952	36	1981	65	1983	66
7	LABO	MARTES	1869	1894	25	1886	17	1919	49
7	LABO	VIERNES	1444	1322	121	1500	56	1327	117
8	LABO	MARTES	1600	1474	126	1620	20	1450	150
8	LABO	JUEVES	1426	1462	36	1490	64	1450	24
8	LABO	VIERNES	1284	1306	23	1439	155	1393	110
8	LABO	MIÉRCOLES	1583	1448	135	1636	53	1415	168
8	LABO	LUNES	1725	1730	5	1733	9	1722	2
9	LABO	MIÉRCOLES	2150	2135	15	2021	129	2135	14
9	LABO	VIERNES	1785	1697	88	1759	26	1680	105
9	LABO	JUEVES	2033	1996	37	2011	22	1984	49
9	LABO	MARTES	2194	2179	15	2088	106	2188	6

9	LABO	LUNES	2326	2311	15	2269	57	2323	3
10	LABO	LUNES	2665	2718	53	2827	161	2722	56
10	LABO	MIÉRCOLES	2643	2536	107	2564	79	2511	132
10	LABO	MARTES	2834	2689	145	2623	211	2655	179
10	LABO	JUEVES	2741	2690	52	2690	52	2696	45
10	LABO	VIERNES	2224	2220	4	2238	14	2223	1
11	LABO	LUNES	2905	2854	50	2822	83	2861	44
11	LABO	MIÉRCOLES	2645	2672	28	2586	59	2660	16
11	LABO	JUEVES	2588	2545	43	2608	20	2556	32
11	LABO	VIERNES	1658	1733	75	1927	270	1746	88
11	LABO	MARTES	2816	2771	45	2747	70	2772	44
12	LABO	JUEVES	2252	2238	14	2188	64	2243	9
12	LABO	MARTES	2493	2434	60	2315	178	2397	97
12	LABO	LUNES	2519	2469	50	2394	125	2435	84
12	LABO	VIERNES	1871	1797	74	1693	178	1783	88
12	LABO	MIÉRCOLES	2192	2190	3	2218	26	2192	0

Tabla 30. Aplicación de los modelos y comparativa sobre el año 2011.

En la tabla 31 podemos ver un resumen de los errores totales que comenten los modelos, confirmándose que GLM es el más eficiente, aunque hay que destacar que los tres algoritmos tiene un comportamiento muy bueno.

Algoritmo	% DE ERROR ABSOLUTO	Error absoluto medio en cada predicción
GLM	2,11025099	50,25
SVM con Kernel Lineal	2,31322704	55,0833333
SVM con Kernel Gausiano	4,22400157	100,583333

Tabla 31. Resumen del error real cometido por los modelos en la predicción del año 2011.

En la Figura 44 podemos ver de forma grafica la comparativa de las previsiones y los datos reales del año 2011.

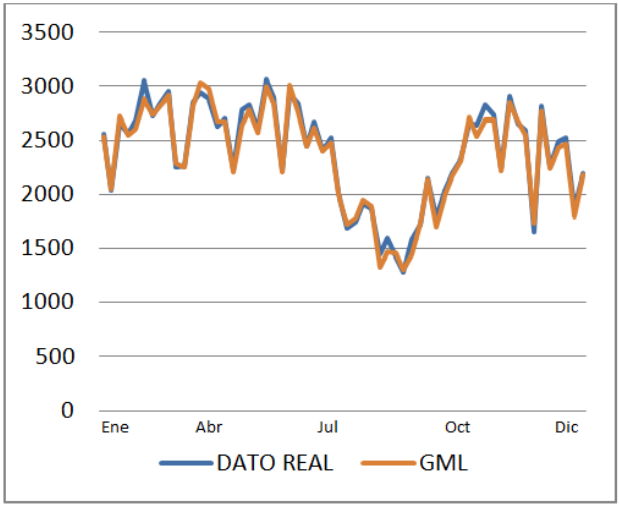


Figura 44. Comparativa grafica de la predicción del modelo GLM y el número de visitas reales del año 2011.

4.1.6. Optimización del Modelo global para la ciudad de Jaén

Una vez demostrado que es más eficiente generar los modelos separando los datos en dos cohortes independientes. Procedemos a realizar predicción del año 2011 utilizando para los días laborables el algoritmo GLM, mientras que para los días Festivos es el algoritmo SVM con kernel Lineal. En la tabla 32 mostramos la comparativa de la predicción de los dos modelos con respecto a los datos reales del año 2011.

DATOS REALES DE 2011				PREVISION
MES	TIPO DIA	DIA SEMANA	DATO REAL	PREDICCIÓN COMBINADA CON SVM LINEAL PARA FESTIVOS Y GLM PARA LABORABLES
1	FESTIVO	DOMINGO	150	157
1	FESTIVO	JUEVES	156	131
1	LABORABLE	JUEVES	2687	2598
1	LABORABLE	LUNES	2657	2732
1	LABORABLE	MARTES	2553	2541
1	LABORABLE	MIÉRCOLES	2554	2534
1	FESTIVO	SÁBADO	184	183
1	LABORABLE	VIERNES	2036	2050
2	FESTIVO	DOMINGO	177	166
2	LABORABLE	JUEVES	2731	2735
2	FESTIVO	LUNES	198	173
2	LABORABLE	LUNES	3054	2889
2	LABORABLE	MARTES	2950	2915

2	LABORABLE	MIÉRCOLES	2838	2816
2	FESTIVO	SÁBADO	173	167
2	LABORABLE	VIERNES	2255	2282
3	FESTIVO	DOMINGO	160	149
3	LABORABLE	JUEVES	2621	2677
3	LABORABLE	LUNES	2855	2819
3	LABORABLE	MARTES	2888	2973
3	LABORABLE	MIÉRCOLES	2943	3028
3	FESTIVO	SÁBADO	173	162
3	LABORABLE	VIERNES	2269	2253
4	FESTIVO	DOMINGO	148	143
4	LABORABLE	JUEVES	2780	2636
4	LABORABLE	LUNES	2831	2782
4	LABORABLE	MARTES	2708	2672
4	LABORABLE	MIÉRCOLES	2605	2568
4	FESTIVO	SÁBADO	149	154
4	FESTIVO	VIERNES	103	116
4	LABORABLE	VIERNES	2224	2211
5	FESTIVO	DOMINGO	123	121
5	LABORABLE	JUEVES	2900	2842
5	FESTIVO	LUNES	105	142
5	LABORABLE	LUNES	3063	3001
5	LABORABLE	MARTES	2927	3009
5	LABORABLE	MIÉRCOLES	2845	2762
5	FESTIVO	SÁBADO	110	112
5	LABORABLE	VIERNES	2215	2205
6	FESTIVO	DOMINGO	133	135
6	LABORABLE	JUEVES	2411	2404
6	LABORABLE	LUNES	2676	2614

6	LABORABLE	MARTES	2529	2477
6	LABORABLE	MIÉRCOLES	2448	2441
6	FESTIVO	SÁBADO	142	146
6	LABORABLE	VIERNES	2014	1999
7	FESTIVO	DOMINGO	125	121
7	LABORABLE	JUEVES	1741	1782
7	LABORABLE	LUNES	1917	1952
7	LABORABLE	MARTES	1869	1894
7	LABORABLE	MIÉRCOLES	1684	1724
7	FESTIVO	SÁBADO	152	152
7	LABORABLE	VIERNES	1444	1322
8	FESTIVO	DOMINGO	112	115
8	LABORABLE	JUEVES	1426	1462
8	FESTIVO	LUNES	103	151
8	LABORABLE	LUNES	1725	1730
8	LABORABLE	MARTES	1600	1474
8	LABORABLE	MIÉRCOLES	1583	1448
8	FESTIVO	SÁBADO	117	124
8	LABORABLE	VIERNES	1284	1306
9	FESTIVO	DOMINGO	148	142
9	LABORABLE	JUEVES	2033	1996
9	LABORABLE	LUNES	2326	2311
9	LABORABLE	MARTES	2194	2179
9	LABORABLE	MIÉRCOLES	2150	2135
9	FESTIVO	SÁBADO	129	131
9	LABORABLE	VIERNES	1785	1697
10	FESTIVO	DOMINGO	124	123
10	LABORABLE	JUEVES	2741	2690
10	LABORABLE	LUNES	2665	2718

10	FESTIVO	MARTES	136	119
10	LABORABLE	MARTES	2834	2689
10	FESTIVO	MIÉRCOLES	130	109
10	LABORABLE	MIÉRCOLES	2643	2536
10	FESTIVO	SÁBADO	161	165
10	LABORABLE	VIERNES	2224	2220
11	FESTIVO	DOMINGO	145	139
11	LABORABLE	JUEVES	2588	2545
11	LABORABLE	LUNES	2905	2854
11	FESTIVO	MARTES	104	117
11	LABORABLE	MARTES	2816	2771
11	LABORABLE	MIÉRCOLES	2645	2672
11	FESTIVO	SÁBADO	142	142
11	LABORABLE	VIERNES	1658	1733
12	FESTIVO	DOMINGO	163	157
12	FESTIVO	JUEVES	137	121
12	LABORABLE	JUEVES	2252	2238
12	FESTIVO	LUNES	191	171
12	LABORABLE	LUNES	2519	2469
12	FESTIVO	MARTES	161	125
12	LABORABLE	MARTES	2493	2434
12	LABORABLE	MIÉRCOLES	2192	2190
12	FESTIVO	SÁBADO	174	182
12	LABORABLE	VIERNES	1871	1797

Tabla 32. Aplicación de la combinación de los modelos más precisos (festivos y laborables) sobre el número de visitas reales del año 2011.

Comparando los datos reales del año 2011 de forma global, el error absoluto es de sólo un 2,29%. En la Figura 45 podemos ver la

comparativa grafica de los datos reales de 2011 y la predicción realizada con los modelos generados con las cohortes independientes.

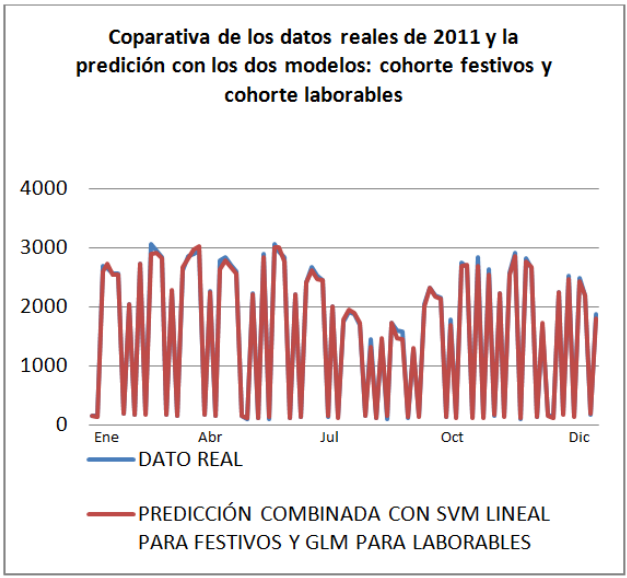


Figura 45. Comparativa grafica de la predicción de la combinación de modelos (festivos y laborables) y el número de visitas reales durante el año 2011.

4.1.6.1 Coeficientes de regresión estandarizados para los días Laborables

Cuando se construye un modelo de regresión lineal multivariable, el algoritmo calcula un coeficiente para cada uno de los predictores utilizados por el modelo. El coeficiente es una medida del impacto del predictor "x" en el objetivo "y". Los coeficientes de regresión nos permiten analizar el comportamiento del modelo y conocer como cada atributo influyen en el, esto nos permite entender la lógica del modelo y

por extensión la lógica de nuestro negocio. En la tabla 33 podemos observar los coeficientes obtenidos:

ATRIBUTO	VALOR	COEFICIENTE ESTANDAR
Intercept		**0**
día semana	VIERNES	-0,432
día semana	JUEVES	-0,173
día semana	LUNES	0,108
día semana	MIERCOLES	-0,093
Humedad Relativa		0,115
Mes	8	-0,61
Mes	7	-0,368
Mes	2	0,356
Mes	10	0,314
Mes	11	0,262
Mes	4	0,242
Mes	5	0,213
Mes	1	0,213
Mes	3	0,185
Mes	9	-0,171
Mes	12	0,096
Calidad del aire	1	0,068
temperatura máxima		-0,168
temperatura media		0,752
temperatura mínima		-0,175

Tabla 33. Coeficientes estandarizados de regresión.

Con un análisis de los datos obtenidos, podemos deducir que a nivel de los días de la semana los miércoles, jueves y sobre todo viernes resta

pacientes a los centros de salud, mientras que el lunes añade un número significativo de pacientes. Esto puede ser debido porque durante el fin de semana sólo se proporciona servicio de urgencias y si un paciente enferma durante el fin de semana y no es algo realmente urgente, este ira al centro de salud el lunes.

Si analizamos el atributo del mes, podemos detectar que, durante los meses de julio, agosto y septiembre, el número de pacientes disminuye. Lógicamente, esto se debe a que coinciden con el período de vacaciones y las altas temperaturas. Por otro lado febrero y octubre son los meses con mayor asistencia de los pacientes a los centros de salud.

A nivel de datos ambiental, podemos ver el peso importante que el modelo tiene la temperatura máxima, mínima y media, con coeficientes estándar estimados -0,175 y 0,752 a 0,168, respectivamente. En términos de humedad relativa se ha demostrado ser el atributo meteorológico menos significativo, pero este tiene su peso en el modelo.

Finalmente, en el nivel de calidad del aire se ha detectado que es el elemento del modelo con menos influencia, con un coeficiente estimado de 0,068. Sin embargo, durante los días con mala calidad del aire se convierte en un atributo importante, con un valor de coeficiente estimado de 102.32.

4.1.6.2 Coeficientes de regresión para los días Festivos

En la tabla 34 podemos ver los coeficientes que usa el algoritmo SVM en la predicción de pacientes para días festivos.

Atributo	valor	Coeficiente Estándar
Intercept		0,49863636

día de la semana	LUNES	0,18892462
día de la semana	MARTES	-0,15558891
día de la semana	SABADO	0,10970659
día de la semana	MIERCOLES	-0,09673843
día de la semana	VIERNES	-0,07478812
día de la semana	JUEVES	0,02517979
día de la semana	DOMINGO	0,00330445
Humedad Relativa		-0,01722341
Mes	8	-0,28251676
Mes	7	-0,24222796
Mes	2	0,18287157
Mes	1	0,14417783
Mes	3	0,12638826
Mes	9	-0,10298506
Mes	12	0,07631651
Mes	11	0,04979944
Mes	10	0,03127474
Mes	5	0,01690143
Mes	4	0
Mes	6	0
Calidad del aire	1	0,03298887
Calidad del aire	0	-0,03298887
temperatura máxima		0,06454707
temperatura media		0,01384894
temperatura mínima		0,04531514

Tabla 34. Coeficientes estandarizados de regresión.

Al analizar los datos de los días de la semana podemos sacar varias conclusiones. El día que más pacientes van a urgencias son aquellos cuyos días festivos coinciden en lunes, con un coeficiente de 0.188, esto es lógico ya que el Sábado y el Domingo no se presta servicio de atención primaria y muchos pacientes el lunes se ve obligada a ir a Urgencias. En cuanto a los días del fin de semana (sábados y Domingo) observamos que el sábado tiene un coeficiente mayor que el domingo, esto también se puede explicar ya que en domingo muchos usuarios enfermos prefieren pedir cita y esperar al lunes para ir a su médico.

En cuanto al análisis del modelo con respecto a los meses, podemos observar que al igual que en días laborales los meses de Julio, Agosto y Septiembre tienen un coeficiente negativo, mientras que los meses de Febrero y Enero son los que tienen mayor coeficiente positivo, esto se explica porque son los meses más fríos en Jaén y suelen coincidir con la aparición enfermedades comunes como resfriados y gripe.

La calidad de aire se puede observar como tiene un peso positivo de 0.03 para los días con mala calidad de aire y -0,03 para los días con buena calidad de aire, lo que quiere decir que en cierta manera cuando hay mala calidad de aire se suman pacientes a los centros de salud.

Finalmente los atributos meteorológicos también tienen su peso en el modelo, teniendo el máximo peso la temperatura Máxima, mínima, humedad relativa y temperatura media respectivamente.

4.2. Uso de los modelos para optimizar el sistema de agendas

Dentro de los servicios que se prestan a los pacientes en los centros de salud hay una serie de tareas administrativas que llevan a cabo los profesionales y que consumen un importante tiempo de su jornada laboral. Durante el año 2011 en Jaén prácticamente el 20% de las visitas se realizaron para obtener la emisión de una receta repetitiva o la emisión de un certificado médico. En la Figura 46 podemos ver una representación grafica de los distintos tipos de demandas por parte de los pacientes:

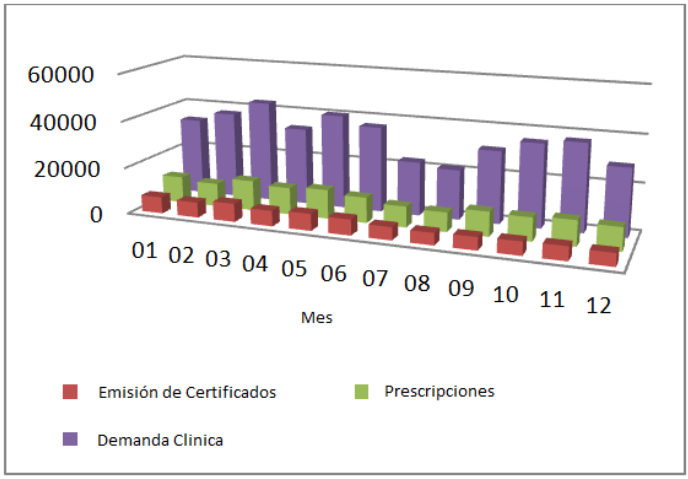


Figura 46. Comparativa del número de demandas clínicas, prescripciones repetitivas y emisión de certificados médicos.

Como se puede ver en la imagen las tareas administrativas representa un importante número de visitas. Actualmente en Jaén las agendas de los profesionales están divididas en huecos regulares de 5 a 10 minutos (normalmente 5 minutos). Tras hablar con varios profesionales nos indican que para estas tareas administrativas (emisión repetida de recetas y emisión de certificados médicos), es suficiente dedicar 1 minuto para una receta repetitiva y 3 minutos para un certificado. El objetivo en esta parte de nuestro trabajo es generar unos modelos de minería de datos para conocer con precisión el número de pacientes que solicitarán estas tareas administrativas y permitir así separar las agendas en tareas administrativas (con distintas programaciones de hueco) y clínica con otra programación, con esto se puede conseguir una importante optimización de los recursos. En la Figura 47 mostramos los periodos de tiempo que son susceptibles de ser optimizados.

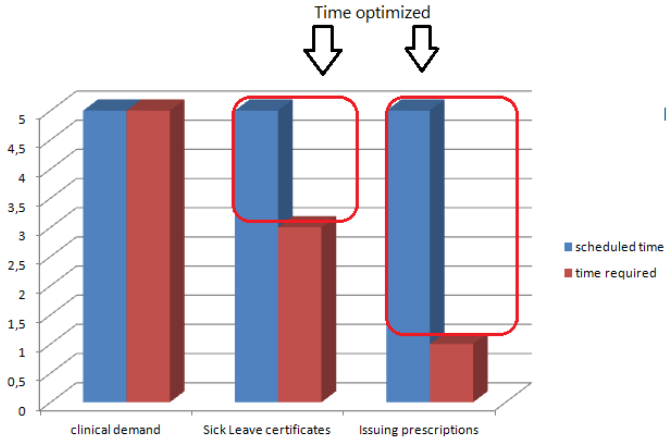


Figura 47. Periodo de tiempo optimizable en las agendas de atención primaria.

Para poder realizar esta separación de las agendas en tareas Administrativa, y Clínicas es necesario disponer de un sistema fiable que sea capaz de realizar una predicción exacta del número de personas que necesitan dichas demanda. Esto es necesario ya que al revisar los datos del 2011 se observa que hay una gran variación de esta demanda de unos días a otros, por ejemplo el 29 de Julio de 2011 la demanda administrativa fue de 341 personas, mientras que el 22 de Noviembre de 2011 esta demanda subió a 1190 pacientes, esto supone una variación de casi un 350%. Un mal dimensionamiento de este modelo podría provocar que quedaran huecos sin ocupar en la demanda administrativa y faltaran huecos en la demanda clínica o a la inversa. Es por esto que necesitamos un modelo fiable que sea capaz de predecir con exactitud la demanda de cada servicio que se ofrece en el centro de salud.

4.2.1. Modelo para la predicción de las receta repetitiva

Nos centramos en la emisión de una renovación de una receta. Este servicio es usado principalmente por usuarios crónicos que necesitan el uso de un medicamento regularmente a lo largo del tiempo.

4.2.1.1 Importancia de atributos

Comenzamos explorando la importancia de atributos con el algoritmo MDL. Este estudio es fundamental ya que nos ayudara a determinar los atributos que usaremos en nuestro estudio.

En la tabla 35 podemos ver el ranking y el peso asignado a cada atributo.

DESTINO: MEDIA_VISITAS		
NOMBRE DE ATRIBUTO	**Ranking**	**IMPORTANCIA**
MES	1	0,25
TEMPERATURA MAXIMA	2	0,15
TEMPERATURA MINIMA	3	0,11
TEMPERATURA MEDIA	4	0,08
DIA DE LA SEMANA	5	0,07
HUMEDAD RELATIVA	6	-0,07
NUMERO DE DIAS CON CALIDAD AMBIENTAL MALA	7	-0,12
PRECIPITACIONES	8	-0,14
NUMERO DE DIAS CON CALIDAD AMBIENTAL ADMISIBLE	9	-0,23
NUMERO DE DIAS CON CALIDAD AMBIENTAL BUENA	10	-0,38
NUMERO DE DIAS CON CALIDAD AMBIENTAL MUY MALA	11	-0,4

Tabla 35. Importancia de atributos.

Como ya habíamos indicado MDL valora los atributos entre 0 y 1. Siendo 1 que tiene la máxima relación con el atributo target, 0 que no tienen ninguna relación con el atributo target. Valores negativos indican que no están relacionados con el atributo target y que pueden generar distorsión en el modelo. En este ranking podemos ver cómo los atributos con más peso son: Mes, temperatura máxima, mínima y media y día de la semana. En la Figura 48 podemos ver también la relación inversa entre las visitas para demanda administrativa y la temperatura.

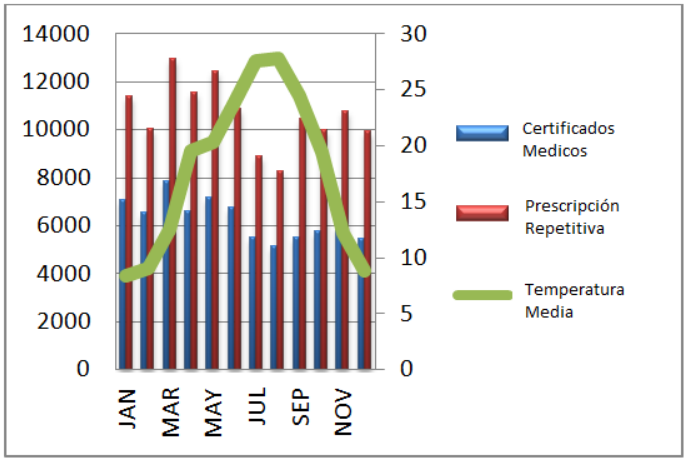


Figura 48. Representación grafica de la relación entre la temperatura media y las tareas administrativas.

En nuestro estudio vamos a eliminar todas las variables cuyo coeficientes son negativo, es decir, número de días con calidad de aire muy Mala, Numero de días con Calidad buena, número de días con calidad admisible, precipitaciones y humedad relativa.

4.2.1.2 Comparativa de los Modelos desde un punto de vista teórico

Tras eliminar los atributos que no tienen peso significativo sobre nuestro atributo target, generamos los modelos de regresión. En la tabla 36 puede verse los resultados teóricos de los tres modelos.

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático
SVM (kernel Gaussiano)	42,38%	166,72	276,68
SVM (Kernel Lineal)	40,27%	164,25	273,56
GLM	36,19%	187,09	279,33

Tabla 36. Resultados teóricos de la eficacia de los modelos.

Desde un punto de vista teórico, tras estudiar los parámetros de los modelos, el más eficiente es SVM con Kernel Gaussiano con una confianza predictiva del 42,38%

4.2.2. Resultados del Modelo para la predicción de demanda de Receta repetitiva

El siguiente paso comparamos los datos reales de pacientes que acudieron a por una receta repetitiva durante el año 2011 y la predicción de los tres modelos. En la tabla 37 podemos ver los resultados.

MES	TIPO DIA	DIA SEMANA	Visitas Reales de Recetas	GML	ERROR ABSOLUTO GML	SVM GUISIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA
1	labo	MIÉRCOLES	567	462	105	710	143	519	48
1	labo	VIERNES	476	368	108	607	132	431	45
1	labo	JUEVES	628	578	51	823	194	636	8
1	labo	MARTES	802	711	91	1054	252	778	24
1	labo	LUNES	583	449	134	757	175	519	63
2	labo	JUEVES	677	697	20	795	118	721	44
2	labo	MIÉRCOLES	560	516	44	670	110	552	9
2	labo	MARTES	547	503	43	710	164	542	5
2	labo	VIERNES	440	397	43	556	116	431	9
2	labo	LUNES	609	579	30	829	220	642	33
3	labo	MIÉRCOLES	617	630	13	740	123	643	26
3	labo	MARTES	617	621	4	739	122	645	28
3	labo	VIERNES	447	407	40	568	121	454	7

3	labo	LUNES	582	529	53	730	148	575	7
3	labo	JUEVES	527	488	39	644	117	520	7
4	labo	MIÉRCOLES	633	581	52	707	74	614	19
4	labo	VIERNES	493	436	56	542	50	468	24
4	labo	JUEVES	623	606	17	718	96	649	26
4	labo	LUNES	840	782	58	868	28	819	22
4	labo	MARTES	884	845	39	971	87	874	10
5	labo	MARTES	643	631	11	765	122	661	18
5	labo	VIERNES	481	453	27	579	99	477	3
5	labo	LUNES	493	389	105	675	181	429	64
5	labo	JUEVES	755	762	7	845	90	785	29
5	labo	MIÉRCOLES	844	835	9	940	96	858	13
6	labo	LUNES	720	674	46	738	18	715	5
6	labo	JUEVES	620	608	11	591	28	611	9
6	labo	MIÉRCOLES	615	586	29	622	8	598	17
6	labo	VIERNES	569	594	25	603	33	616	46
6	labo	MARTES	691	669	22	782	90	708	17
7	labo	JUEVES	443	429	14	500	57	440	3
7	labo	LUNES	449	389	60	530	81	419	30
7	labo	MIÉRCOLES	436	391	45	513	77	421	15
7	labo	MARTES	471	441	30	521	50	464	7
7	labo	VIERNES	341	231	110	424	83	258	83
8	labo	JUEVES	372	326	46	435	64	355	17
8	labo	VIERNES	339	292	47	399	60	328	11
8	labo	MIÉRCOLES	490	455	34	500	10	480	9
8	labo	MARTES	489	453	36	512	23	483	6

8	labo	LUNES	377	275	103	478	100	336	41
9	labo	MIÉRCOLES	544	491	53	555	11	535	9
9	labo	VIERNES	395	269	125	443	49	327	68
9	labo	JUEVES	449	355	94	516	67	398	51
9	labo	LUNES	649	607	42	705	56	672	23
9	labo	MARTES	531	468	63	573	42	521	9
10	labo	JUEVES	534	443	90	601	67	490	44
10	labo	MIÉRCOLES	528	486	42	725	197	538	11
10	labo	MARTES	604	544	60	796	192	608	4
10	labo	LUNES	869	771	98	793	76	824	44
10	labo	VIERNES	464	368	96	536	73	418	46
11	labo	MARTES	1190	1192	2	1235	46	1181	9
11	labo	VIERNES	425	409	16	621	196	481	56
11	labo	LUNES	570	464	106	707	137	523	47
11	labo	MIÉRCOLES	539	437	103	697	158	485	54
11	labo	JUEVES	539	468	71	649	110	509	29
12	labo	JUEVES	497	431	66	555	59	459	37
12	labo	VIERNES	522	477	45	535	14	499	23
12	labo	MARTES	593	568	25	816	223	610	18
12	labo	MIÉRCOLES	767	735	32	805	38	757	10
12	labo	LUNES	599	549	50	734	136	600	1

Tabla 37. Aplicación de los modelos y comparativa sobre el año 2011.

En este caso a pesar de que el algoritmo que tiene un mejor comportamiento desde el punto de visto teórico es el SVM con Kernel Gaussiano, sin embargo al aplicarlo sobre una predicción real, el error absoluto que comete es del 17% frente al 4,33% del SVM con Kernel Lineal. Con lo cual en este caso el algoritmo más eficiente es este último. En la tabla 38 podemos ver el error real cometido por cada algoritmo.

Algoritmo	% DE ERROR ABSOLUTO	Error absoluto medio en cada predicción
GLM	9,05676697	52,2138717
SVM con Kernel Lineal	4,33308465	24,9810033
SVM con Kernel Gausiano	17,0733112	98,4306745

Tabla 38. Resumen del error real cometido por los modelos en la predicción del año 2011.

En la Figura 49 mostramos la comparativa de la predicción del modelo SVM con kernel Lineal y los datos reales de vistas de pacientes que acudieron a por una receta repetitiva en el año 2011.

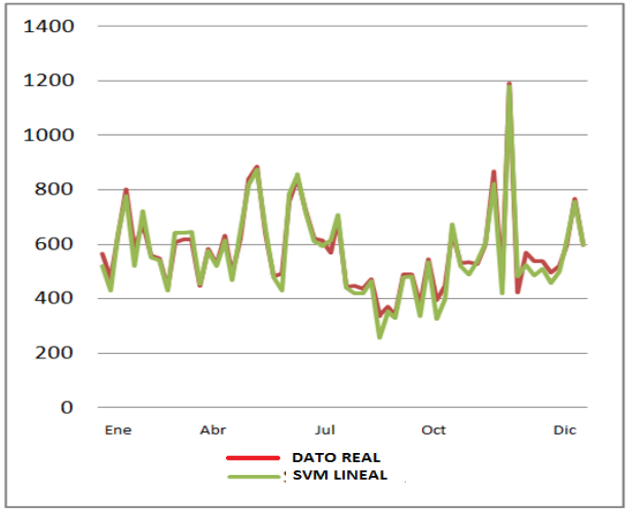


Figura 49. Comparativa grafica de la predicción del modelo SVM con kernel lineal y el número de visitas reales del año 2011.

4.2.3. Modelo para la predicción de la emisión de certificados médicos

Otra de las tareas más importantes realizadas por los facultativos en los centros de salud es la emisión de partes médicos. Según la legislación Andaluza todo usuario que este de baja por enfermedad en sus empresas, necesita un parte de baja que tiene que ser emitido por su médico de cabecera y un parte de renovación semanal de la baja, por tanto, todos los usuarios en esta situación, necesitan ir al centro de salud para que le emitan dicho certificado, como mínimo una vez a la semana.

4.2.3.1 Importancia de atributos

En cuanto a la importancia de atributos es muy similar a de la otra tarea administrativa, en el ranking aparecen con más peso los atributos: Mes, temperatura máxima, mínima y media y día de la semana.

4.2.3.2 Comparativa de los Modelos desde un punto de vista teórico

Al generar el modelo para la otra gran tarea administrativa que se realizan en los centros de Salud, como es la emisión de certificados médicos, el algoritmo más eficiente desde el punto de vista teórico es el modelo GLM. En la tabla 39 podemos ver los parámetros de los tres modelos.

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático
SVM (kernel Lineal)	**46,32%**	**1,24**	**2,87**
SVM (Kernel Gaussiano)	**61.24%**	3,01	3,99
GLM	**61.44%**	1,29	2,87

Tabla 39. Resultados teóricos de la eficacia de los modelos.

4.2.4. Resultados del Modelo para la predicción de demanda de certificados médico

A continuación aplicamos los tres modelos para predecir las visitas de 2011 y compararemos los resultados con los datos reales. En la tabla 40 se muestra los resultados obtenidos.

DATOS REALES DE 2011				PREVISION					
MES	TIPO DIA	DIA SEMANA	MEDIA DE DIAGNOSTICOS REALES EN EL DIA CARACTERIZADO	GML	ERROR ABSOLUTO GML	SVM GUISIANO	ERROR ABSOLUTO SVM	SVM LINEAL	ERROR ABSOLUTO SVM LINEA
1	labo	LUNES	27	31	3	31	4	30	3
1	labo	JUEVES	21	20	0	20	1	20	1
1	labo	MARTES	19	19	0	18	0	18	1

1	labo	VIERNES	19	19	1	20	1	19	1
1	labo	MIÉRCOLES	16	16	0	16	0	16	0
2	labo	MARTES	18	18	0	18	0	18	1
2	labo	MIÉRCOLES	14	14	0	15	0	15	1
2	labo	LUNES	28	27	1	26	2	26	2
2	labo	VIERNES	25	25	0	22	3	25	0
2	labo	JUEVES	16	17	1	18	2	17	1
3	labo	VIERNES	16	16	0	19	3	17	1
3	labo	JUEVES	20	20	0	18	3	21	1
3	labo	MARTES	16	16	1	15	1	16	0
3	labo	LUNES	28	29	1	28	0	28	1
3	labo	MIÉRCOLES	16	14	1	13	3	15	0
4	labo	JUEVES	21	20	1	17	4	19	2
4	labo	LUNES	29	27	2	25	4	26	3
4	labo	MIÉRCOLES	20	19	1	16	3	19	0
4	labo	MARTES	16	16	0	13	2	16	0
4	labo	VIERNES	14	14	0	18	4	14	0
5	labo	MARTES	23	24	0	20	4	24	1
5	labo	MIÉRCOLES	22	20	2	15	7	20	2
5	labo	VIERNES	24	23	0	19	4	23	0
5	labo	JUEVES	24	22	2	20	4	21	2
5	labo	LUNES	37	37	0	30	7	36	0
6	labo	MIÉRCOLES	11	10	1	15	4	11	0
6	labo	LUNES	24	23	1	27	3	23	1
6	labo	JUEVES	20	19	1	23	2	20	0
6	labo	MARTES	19	18	1	18	1	18	1

6	labo	VIERNES	20	19	1	21	1	19	1
7	labo	MIÉRCOLES	7	7	0	12	5	7	1
7	labo	MARTES	7	8	1	16	9	8	1
7	labo	LUNES	12	14	2	22	10	14	2
7	labo	JUEVES	10	10	1	17	7	11	1
7	labo	VIERNES	13	12	1	19	5	13	0
8	labo	JUEVES	13	13	0	17	4	13	0
8	labo	VIERNES	20	20	0	21	1	20	0
8	labo	LUNES	23	24	1	26	3	24	1
8	labo	MIÉRCOLES	14	14	0	14	0	14	0
8	labo	MARTES	16	16	0	17	0	16	0
9	labo	VIERNES	13	11	2	19	6	12	1
9	labo	MARTES	12	11	0	15	3	12	0
9	labo	LUNES	25	24	1	27	2	23	2
9	labo	JUEVES	18	17	1	18	0	18	0
9	labo	MIÉRCOLES	15	15	1	15	0	15	0
10	labo	JUEVES	18	18	0	20	2	18	0
10	labo	VIERNES	17	18	0	21	4	18	0
10	labo	MARTES	16	16	0	15	1	16	0
10	labo	MIÉRCOLES	16	16	0	12	4	16	0
10	labo	LUNES	18	20	1	25	6	19	1
11	labo	MIÉRCOLES	14	12	2	16	2	13	1
11	labo	LUNES	24	24	1	31	8	25	1
11	labo	VIERNES	21	20	1	19	2	20	1
11	labo	MARTES	10	13	4	18	9	12	3
11	labo	JUEVES	21	20	0	23	3	21	0

12	labo	LUNES	30	27	3	21	9	27	4
12	labo	VIERNES	20	19	1	19	1	19	1
12	labo	MARTES	27	23	4	14	13	23	4
12	labo	MIÉRCOLES	15	14	1	8	7	14	1
12	labo	JUEVES	22	20	1	20	2	21	1

Tabla 40. Aplicación de los modelos y comparativa sobre el año 2011.

En este caso se confirma también que GLM es mejor algoritmos para esta tarea administrativa con un error absoluto del 4,60%. En la tabla 41 mostramos los errores reales de los 3 modelos.

Algoritmo	% DE ERROR MEDIO	Error absoluto medio en cada predicción
GLM	4,60176991	0,86666667
SVM con Kernel Lineal	4,77876106	0,9
SVM con Kernel Gausiano	18,1415929	3,41666667

Tabla 41. Resumen del error real cometido por los modelos en la predicción del año 2011.

Para ver de forma grafica la eficiencia del modelo, comparamos los datos reales de 2011 de vistas y la predicción del algoritmo GLM (Figura 50).

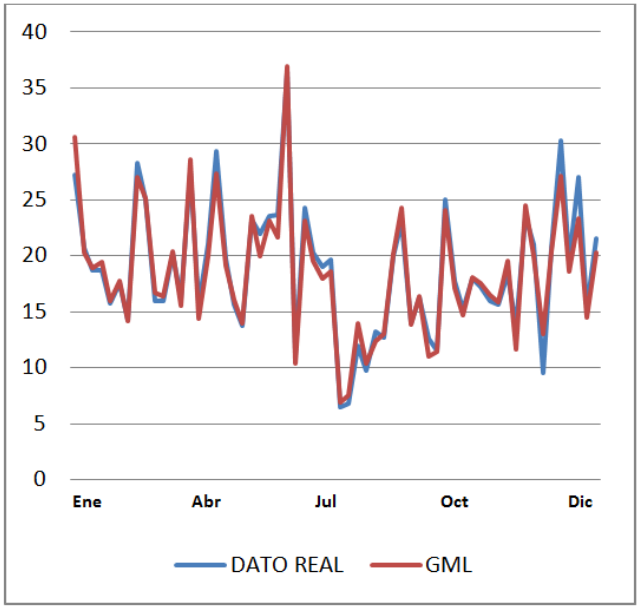


Figura 50. Comparativa grafica de la predicción del modelo GLM y el número de visitas reales del año 2011.

5. RESULTADOS CON INFORMACIÓN ESPACIAL

En la introducción indicábamos que el porcentaje de visitas que tienen los centros de salud no se corresponde con el porcentaje de pacientes adscritos. En la tabla 42 podemos comprobar los datos.

CENTRO DE SALUD	% DE ADSCRIPCIONES	% de visitas totales
Virgen de la Capilla	18,55%	16,40%
Belén	8,17%	8,16%
San Felipe	17,49%	17,84%
La Magdalena	11,11%	11,33%
Fuentezuelas	7,30%	7,32%
El Valle	14,68%	15,55%
Federico Castillo	22,69%	23,41%

Tabla 42. Porcentaje de adscripciones de pacientes y visitas reales a los centros de salud.

Esto es indicativo de que hay ciertos factores locales a cada Centro de Salud o Farmacia que influyen de forma muy notable en que los pacientes asistan con más o menos frecuencia a dichos centros de Salud o a solicitar ciertos medicamentos.

En esta parte del estudio vamos a generar una serie de modelos añadiendo variables espaciales al estudio.

5.1. Generación de los modelos predictivos con las variables espaciales

Para abordar esta parte del estudio, partimos de los modelos generados en el punto anterior que generaliza todas las visitas que se producen en Jaén Capital con un error del 2,11 %. El problema surge cuando queremos extrapolar estos datos a cada Centro de Salud. En la actualidad el 100% de la actividad en Jaén está repartida entre 7 Centros de Salud. En la Figura 51 podemos ver de forma grafica el porcentaje de población que es atendida por cada centro de salud.

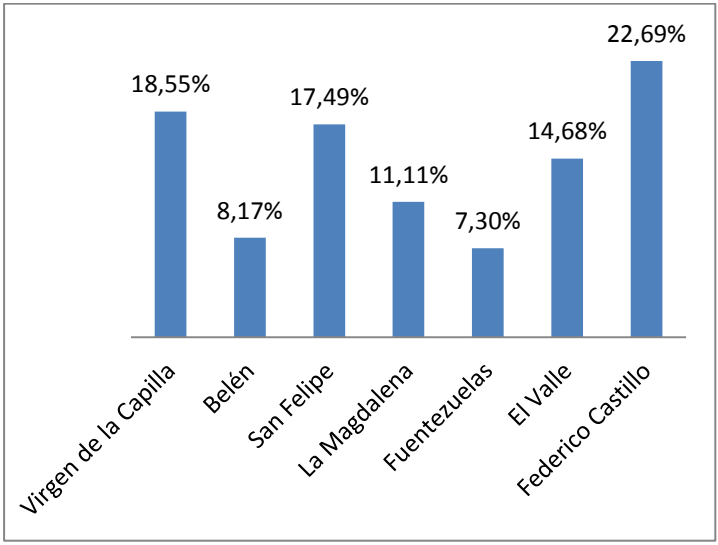


Figura 51. Porcentaje de pacientes adscritos a los centros de salud de Jaén.

Al realizar la extrapolación de nuestro modelo a cada centro de Salud, el porcentaje de error se dispara muy por encima del 2,11%, que era el error para toda la población de Jaén, esto quiere decir que el modelo generado que calcula todas las citas de atención primaria de Jaén modela

correctamente el número de visitas totales, pero al aplicarlo para un centro de salud concreto su porcentaje de error absoluto se dispara. En la tabla 43 observamos los errores que se comente para cada centro de salud. Para algunos centros de Salud el error es de solo el 2,54%, sin embargo para otros es del 11,77%.

CENTRO DE SALUD	% ERRORES ABSOLUTOS
FUENTEZUELAS	11,75
VIRGEN DE LA CAPILLA	11,77
SAN FELIPE	5,32
FEDERICO DEL CASTILLO	2,54
EL VALLE	7,16
LA MAGDALENA	6,67
BELEN	6,6

Tabla 43. Porcentaje de error absoluto cometido por el modelo al aplicarlo a los centros de salud.

La diferencia radica sobre todo en el tipo de población atendida, en teoría no es lo mismo atender un alto porcentaje de población pediátrica o geriátrica, ya que este sector de la población necesita mayor atención sanitaria. Otros factores a tener en cuenta es nivel económico de la zona, a mayor nivel económico es posible que haya un mayor número de seguros privados y que esto reste lógicamente pacientes. En definitiva si buscamos un modelo general que sea optimo para un centro de Salud concreto, tenemos que añadir variables locales de la población que es atendida por este centro de Salud.

5.1.1. Importancia de atributos

Al igual que con los anteriores modelos calculamos la importancia de atributos con el Algoritmo MDL para determinar el peso de los mismos en la predicción. En la tabla 44 podemos ver los valores.

Nombre	Clasificación	Importancia
E0_14	1	0,731
E15_24	1	0,731
E25_34	1	0,731
E35_44	1	0,731
E45_54	1	0,731
E55_64	1	0,731
E65_74	1	0,731
E75_84	1	0,731
MAS_85	1	0,731
NIVEL_ECONOMICO	1	0,731
USUARIOS_ADCRITOS	1	0,731
MES	3	0,064
DIA_SEMANA	3	0,064
TEMP_MINIMA	4	0,052
TEMP_MAXIMA	5	0,046
TEMP_MEDIA	6	0,04
NUM_DIAS_MALA	7	0,007
HUMEDAD_RELATIVA	9	-0,008

Tabla 44. Coeficientes de regresión estandarizados.

Como se puede comprobar los atributos con mayor peso son ahora, el rango etario (edad de los pacientes), nivel económico, número de pacientes adscritos al centro de salud. Es decir las variables espaciales añadidas al modelo.

Por otro lado podemos observar como las variables del mes, día de la semana, temperatura y calidad del aire son las siguientes más importantes.

5.1.2. Comparativa del Modelo desde un punto de vista teórico

Desde un punto de vista teórico, tras estudiar los parámetros de los modelos, el más eficiente es GLM con una confianza predictiva del 89,25%. Hay que destacar que los 3 algoritmos tienen un buen comportamiento y su confianza predictiva se sitúa por encima del 85%, lo cual quiere decir que los tres modelos tienen un comportamiento muy bueno. En la tabla 45 podemos ver un resumen del comportamiento teórico de los algoritmos.

Modelo	% de Confianza predictiva	erro de promedio absoluto	error cuadrático	Valor Promedio Previsto	Valor promedio Real
SVM (kernel Lineal)	87,33	10,87	16,62	2312	2298
SVM (Kernel Gaussiano)	87,73	11,79	22,32	2312	2300
GLM	89,25	9,25	15,25	2312	2308

Tabla 45. Resultados teóricos de la eficacia de los modelos.

5.1.3. Resultados del Modelo Óptimo para la predicción de usuarios que acuden a cada centros de Salud

En las tablas 46, 47, 48, 49, 50, 51 y 52 podemos ver los resultados de las visitas reales a los centros de salud durante el año 2011 y la predicción del nuevo modelo generado que tiene en cuenta las variables espaciales.

- VIRGEN DE LA CAPILLA

mes	día	año	DATO REAL	GLM	error absoluto
1	MIÉRCOLES	2011	423	420	2
1	VIERNES	2011	354	354	0
1	LUNES	2011	435	454	19
1	MARTES	2011	410	414	4
1	JUEVES	2011	472	439	34
2	LUNES	2011	451	435	16
2	JUEVES	2011	417	417	1
2	MIÉRCOLES	2011	482	468	14
2	MARTES	2011	462	453	9
2	VIERNES	2011	360	362	2
3	VIERNES	2011	361	356	5
3	LUNES	2011	448	443	5
3	MIÉRCOLES	2011	497	506	10
3	MARTES	2011	442	456	14
4	MARTES	2011	433	437	4
4	VIERNES	2011	397	394	3
4	LUNES	2011	463	460	2
4	MIÉRCOLES	2011	466	461	6

5	LUNES	2011	493	480	13
5	JUEVES	2011	439	432	7
5	VIERNES	2011	400	388	12
5	MIÉRCOLES	2011	496	472	24
6	MIÉRCOLES	2011	421	427	6
6	MARTES	2011	387	387	0
7	MIÉRCOLES	2011	295	297	2
7	JUEVES	2011	301	300	1
7	MARTES	2011	326	323	3
8	MARTES	2011	255	245	9
8	JUEVES	2011	259	260	1
8	MIÉRCOLES	2011	255	244	11
8	LUNES	2011	284	285	1
9	MIÉRCOLES	2011	364	360	4
9	VIERNES	2011	287	279	8
9	MARTES	2011	388	381	7
9	LUNES	2011	378	376	2
10	MIÉRCOLES	2011	437	418	19
10	JUEVES	2011	418	420	2
11	LUNES	2011	451	448	3
11	JUEVES	2011	394	395	2
11	VIERNES	2011	259	284	25
11	MARTES	2011	447	443	4
12	JUEVES	2011	386	374	11
12	MARTES	2011	452	413	38
12	LUNES	2011	469	426	43
12	VIERNES	2011	326	315	10
12	MIÉRCOLES	2011	381	373	9

3	JUEVES	2011	420	431	12
4	JUEVES	2011	473	445	28
5	MARTES	2011	484	500	16
6	LUNES	2011	424	418	6
6	JUEVES	2011	401	406	6
6	VIERNES	2011	313	315	1
7	LUNES	2011	323	323	0
7	VIERNES	2011	249	233	16
8	VIERNES	2011	206	208	2
9	JUEVES	2011	331	332	2
10	LUNES	2011	436	457	20
10	MARTES	2011	410	411	0
10	VIERNES	2011	364	367	4
11	MIÉRCOLES	2011	450	461	12

Tabla 46. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud Virgen de la Capilla.

El error absoluto total durante en la predicción del año 2011 es del 2,34%

- FUENTEZUELAS

mes	día semana	año	DATO REAL	GLM PREDICCION	ERROR ABSOLUTO
1	LUNES	2011	204	200	4
1	MARTES	2011	194	200	6
1	MIÉRCOLES	2011	224	225	2
1	VIERNES	2011	176	176	0

1	JUEVES	2011	182	187	5
2	VIERNES	2011	221	222	1
2	JUEVES	2011	193	200	8
2	MIÉRCOLES	2011	248	251	4
2	MARTES	2011	208	215	8
2	LUNES	2011	228	234	6
3	JUEVES	2011	189	180	9
3	VIERNES	2011	242	240	2
3	LUNES	2011	222	226	4
3	MARTES	2011	211	211	0
4	MIÉRCOLES	2011	245	234	11
4	LUNES	2011	215	222	7
4	VIERNES	2011	183	185	3
4	MARTES	2011	196	205	9
5	MIÉRCOLES	2011	217	221	5
5	JUEVES	2011	163	170	7
5	VIERNES	2011	167	168	2
5	JUEVES	2011	184	190	6
6	MIÉRCOLES	2011	204	206	2
6	LUNES	2011	223	228	5
7	MARTES	2011	183	186	3
7	VIERNES	2011	161	163	2
7	JUEVES	2011	133	98	35
8	LUNES	2011	182	187	5
8	MARTES	2011	164	171	7
8	MIÉRCOLES	2011	183	180	3
8	JUEVES	2011	152	153	2
9	MIÉRCOLES	2011	163	164	1

9	VIERNES	2011	120	119	1
9	MARTES	2011	163	166	3
9	LUNES	2011	176	179	3
10	JUEVES	2011	84	86	2
10	LUNES	2011	122	124	2
11	MARTES	2011	113	115	2
11	VIERNES	2011	96	94	3
11	MIÉRCOLES	2011	141	145	4
11	MIÉRCOLES	2011	187	185	2
12	LUNES	2011	153	160	7
12	VIERNES	2011	155	156	1
12	MARTES	2011	138	141	4
12	JUEVES	2011	138	120	18
12	MARTES	2011	183	180	3
3	MIÉRCOLES	2011	173	178	5
4	JUEVES	2011	176	182	7
5	VIERNES	2011	164	169	5
6	LUNES	2011	161	152	9
6	VIERNES	2011	171	179	8
6	JUEVES	2011	146	153	7
7	MIÉRCOLES	2011	191	166	25
7	MARTES	2011	181	191	10
8	MARTES	2011	157	159	2
9	MIÉRCOLES	2011	140	146	6
10	JUEVES	2011	137	141	4
10	LUNES	2011	150	156	6
10	VIERNES	2011	135	138	3
11	LUNES	2011	184	193	9

Tabla 47. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud Las Fuentezuelas.

El error absoluto cometido por el modelo en la predicción de todo el año 2011 es del 3,1%.

- BELEN

mes	Día de la semana	año	DATO REAL	GLM PREDICIC.	ERROR
1	VIERNES	2011	160	168	8
1	JUEVES	2011	226	232	7
1	MIÉRCOLES	2011	209	215	6
1	LUNES	2011	228	233	6
1	MARTES	2011	213	221	8
2	VIERNES	2011	168	169	1
2	MARTES	2011	251	258	7
2	MIÉRCOLES	2011	234	242	9
2	LUNES	2011	233	240	7
2	JUEVES	2011	225	230	5
3	MIÉRCOLES	2011	220	224	4
3	VIERNES	2011	176	189	14
3	JUEVES	2011	190	192	2
3	LUNES	2011	227	236	9
4	MARTES	2011	250	248	2
4	MARTES	2011	225	235	9
4	MIÉRCOLES	2011	192	190	2
4	VIERNES	2011	150	152	2

5	JUEVES	2011	268	269	1
5	LUNES	2011	227	235	8
5	MIÉRCOLES	2011	223	230	6
5	JUEVES	2011	219	218	1
6	LUNES	2011	209	209	0
6	VIERNES	2011	166	167	1
7	MARTES	2011	263	269	6
7	VIERNES	2011	172	173	1
7	MIÉRCOLES	2011	211	218	7
8	MARTES	2011	226	224	2
8	LUNES	2011	231	227	4
8	JUEVES	2011	213	212	1
8	JUEVES	2011	136	142	5
9	MIÉRCOLES	2011	135	140	5
9	VIERNES	2011	129	135	5
9	MARTES	2011	145	147	2
9	LUNES	2011	171	168	3
10	JUEVES	2011	115	114	1
10	LUNES	2011	155	155	0
11	MARTES	2011	152	151	1
11	VIERNES	2011	105	106	1
11	MIÉRCOLES	2011	137	141	4
11	MIÉRCOLES	2011	191	185	6
12	LUNES	2011	198	198	1
12	VIERNES	2011	140	143	3
12	MARTES	2011	182	183	1
12	JUEVES	2011	174	174	1
12	MARTES	2011	228	225	3

3	MIÉRCOLES	2011	219	223	4
4	JUEVES	2011	227	226	1
5	VIERNES	2011	153	151	2
6	LUNES	2011	226	227	2
6	VIERNES	2011	180	183	3
6	JUEVES	2011	238	233	4
7	MIÉRCOLES	2011	236	234	2
7	MARTES	2011	219	221	2
8	MARTES	2011	183	180	3
9	MIÉRCOLES	2011	182	187	6
10	JUEVES	2011	183	181	2
10	LUNES	2011	202	204	2
10	VIERNES	2011	143	140	3
11	LUNES	2011	216	224	8

Tabla 48. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud de Belén.

El erro cometido por el modelo en la predicción del año 2011 es del 1,99%.

- **FEDERICO DEL CASTILLO.**

mes	**Día semana**	**año**	**DATO REAL**	**GLM**	**ERROR ABSOLUTO**
1	LUNES	2011	622	620	1,6
1	MARTES	2011	596	591,1364	4,6136
1	MIÉRCOLES	2011	628	618,8591	9,3909

1	VIERNES	2011	501	497,1709	3,8291
1	JUEVES	2011	633	630	3,3333
2	VIERNES	2011	506	506,5613	0,5613
2	JUEVES	2011	662	653,2282	9,0218
2	MIÉRCOLES	2011	643	636,7259	6,5241
2	MARTES	2011	728	714,0407	13,9593
2	LUNES	2011	745	740	5
3	JUEVES	2011	608	600	7,8
3	VIERNES	2011	497	494,4817	2,2683
3	LUNES	2011	658	647,8869	9,8631
3	MARTES	2011	659	659	0,4
3	MIÉRCOLES	2011	668	668	0,2
4	LUNES	2011	673	663,7942	8,9558
4	VIERNES	2011	525	522,3087	2,6913
4	MARTES	2011	644	637,9011	6,0989
4	MIÉRCOLES	2011	602	597,7662	4,4838
4	JUEVES	2011	660	651	8,6667
5	VIERNES	2011	507	503,5165	3,2335
5	JUEVES	2011	651	640,2144	10,2856
5	MIÉRCOLES	2011	668	653,6553	14,3447
5	LUNES	2011	724	708,3132	15,1868
5	MARTES	2011	692	763,477	71,677
6	VIERNES	2011	470	469,9191	0,0809
6	JUEVES	2011	551	599,324	48,124
6	LUNES	2011	654	641,9617	12,0383
6	MARTES	2011	608	600,4872	7,7628

6	MIÉRCOLES	2011	553	550	3
7	JUEVES	2011	376	382,8969	7,1469
7	MIÉRCOLES	2011	378	385,3652	7,6152
7	VIERNES	2011	299	304,5336	5,7336
7	MARTES	2011	431	435,6694	4,4194
7	LUNES	2011	426	431,724	6,224
8	JUEVES	2011	324	330,722	6,472
8	LUNES	2011	405	406,6605	1,9105
8	MARTES	2011	396	415,536	19,736
8	VIERNES	2011	288	294,039	5,789
8	MIÉRCOLES	2011	362	375,9503	14,3503
9	MIÉRCOLES	2011	469	467,557	1,693
9	LUNES	2011	506	503,7732	2,4768
9	VIERNES	2011	395	413,4767	18,6767
9	MARTES	2011	503	498,7208	3,7792
9	JUEVES	2011	491	490	0,8
10	MARTES	2011	726	720	5,6667
10	MIÉRCOLES	2011	596	590	5,6667
10	VIERNES	2011	559	553,8754	5,1246
10	LUNES	2011	664	731,844	67,844
11	LUNES	2011	729	714,329	14,421
11	VIERNES	2011	506	508	2,3333
11	JUEVES	2011	616	605,7663	9,7337
11	MIÉRCOLES	2011	575	628,7498	53,3498
11	MARTES	2011	722	708,5418	13,4582
12	MARTES	2011	587	581	5,6667

12	MIÉRCOLES	2011	512	510,2758	1,9742
12	LUNES	2011	566	567	0,6667
12	VIERNES	2011	421	445,6874	24,2874
10	JUEVES	2011	678	665,2219	13,0281
12	JUEVES	2011	530	524,8289	5,4211

Tabla 499. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud de Federico del Castillo.

El erro cometido por el modelo en la predicción del año 2011 es del 1,87%.

- **SAN FELIPE**

mes	Día semana	año	DATO REAL	GLM	ERROR ABSOLUTO
1	LUNES	2011	462	488	26
1	MARTES	2011	440	444	5
1	MIÉRCOLES	2011	442	443	1
1	VIERNES	2011	341	351	11
1	JUEVES	2011	428	421	7
2	VIERNES	2011	403	403	0
2	JUEVES	2011	488	480	8
2	MIÉRCOLES	2011	501	490	10
2	MARTES	2011	550	529	21
2	LUNES	2011	542	496	46
3	JUEVES	2011	434	453	19
3	VIERNES	2011	410	402	8

3	LUNES	2011	512	500	12
3	MARTES	2011	511	531	19
3	MIÉRCOLES	2011	528	544	17
4	LUNES	2011	473	476	3
4	VIERNES	2011	405	407	3
4	MARTES	2011	493	491	2
4	MIÉRCOLES	2011	463	466	3
4	JUEVES	2011	463	447	16
5	VIERNES	2011	422	412	10
5	JUEVES	2011	595	560	34
5	MIÉRCOLES	2011	512	492	20
5	LUNES	2011	586	560	26
5	MARTES	2011	555	577	22
6	VIERNES	2011	355	355	0
6	JUEVES	2011	410	423	13
6	LUNES	2011	501	485	16
6	MARTES	2011	460	451	9
6	MIÉRCOLES	2011	445	458	13
7	JUEVES	2011	303	309	6
7	MIÉRCOLES	2011	291	302	11
7	VIERNES	2011	259	251	9
7	MARTES	2011	310	318	9
7	LUNES	2011	323	331	7
8	JUEVES	2011	262	271	9
8	LUNES	2011	282	291	9
8	MARTES	2011	270	268	2
8	VIERNES	2011	235	238	3
8	MIÉRCOLES	2011	270	266	4

9	MIÉRCOLES	2011	361	366	5
9	LUNES	2011	426	420	5
9	VIERNES	2011	308	307	1
9	MARTES	2011	364	371	7
9	JUEVES	2011	337	346	9
10	MARTES	2011	562	560	2
10	MIÉRCOLES	2011	501	462	38
10	JUEVES	2011	504	494	10
10	VIERNES	2011	398	401	4
11	LUNES	2011	516	506	10
11	VIERNES	2011	394	392	2
11	JUEVES	2011	452	448	3
11	MIÉRCOLES	2011	495	512	18
11	MARTES	2011	482	478	4
12	MARTES	2011	458	425	33
12	MIÉRCOLES	2011	380	380	1
12	JUEVES	2011	403	395	7
12	LUNES	2011	470	435	35
12	VIERNES	2011	332	329	3
10	LUNES	2011	502	528	26

Tabla 50. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud de San Felipe.

El erro cometido por el modelo en la predicción del año 2011 es del 2,70%.

- **EL VALLE**

Mes	Día semana	año	DATO REAL	GLM	ERROR ABSOLUTO
1	LUNES	2011	396	409	13
1	MARTES	2011	388	389	1
1	MIÉRCOLES	2011	373	374	1
1	VIERNES	2011	300	305	4
1	JUEVES	2011	379	377	2
2	VIERNES	2011	337	336	1
2	JUEVES	2011	436	425	12
2	MIÉRCOLES	2011	436	425	11
2	MARTES	2011	433	423	10
2	LUNES	2011	510	510	0
3	JUEVES	2011	442	445	3
3	VIERNES	2011	351	340	10
3	LUNES	2011	476	456	19
3	MARTES	2011	480	485	5
3	MIÉRCOLES	2011	464	467	3
4	LUNES	2011	469	457	12
4	VIERNES	2011	352	352	1
4	MARTES	2011	409	410	2
4	MIÉRCOLES	2011	388	393	4
4	JUEVES	2011	457	428	29
5	VIERNES	2011	319	317	2
5	JUEVES	2011	440	425	15
5	MIÉRCOLES	2011	427	411	16

5	LUNES	2011	493	472	21
5	MARTES	2011	413	424	11
6	VIERNES	2011	326	317	9
6	JUEVES	2011	405	403	2
6	LUNES	2011	382	378	4
6	MARTES	2011	413	399	14
6	MIÉRCOLES	2011	386	386	0
7	JUEVES	2011	277	274	3
7	MIÉRCOLES	2011	267	268	1
7	VIERNES	2011	224	201	23
7	MARTES	2011	278	278	0
7	LUNES	2011	307	303	4
8	JUEVES	2011	217	220	3
8	LUNES	2011	254	254	0
8	MARTES	2011	228	212	16
8	VIERNES	2011	184	183	1
8	MIÉRCOLES	2011	238	220	18
9	MIÉRCOLES	2011	338	333	6
9	LUNES	2011	393	380	13
9	VIERNES	2011	292	276	16
9	MARTES	2011	334	332	2
9	JUEVES	2011	331	325	6
10	MARTES	2011	428	413	15
10	MIÉRCOLES	2011	434	408	26
10	VIERNES	2011	327	331	4
10	LUNES	2011	400	414	14
11	LUNES	2011	461	448	13
11	VIERNES	2011	334	342	8

11	JUEVES	2011	424	411	13
11	MIÉRCOLES	2011	408	413	5
11	MARTES	2011	444	433	11
12	MARTES	2011	372	359	13
12	MIÉRCOLES	2011	355	345	10
12	LUNES	2011	375	364	11
12	VIERNES	2011	287	270	17
10	JUEVES	2011	428	419	8
12	JUEVES	2011	341	332	9

Tabla 51. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud de El Valle.

El error cometido por el modelo en la predicción del año 2011 es del 2,36%.

- **LA MAGDALENA**

Mes	**Día semana**	**año**	**DATO REAL**	**GLM**	**ERROR ABSOLUTO**
1	LUNES	2011	312	311	2
1	MARTES	2011	313	314	1
1	MIÉRCOLES	2011	257	267	11
1	VIERNES	2011	205	214	9
1	JUEVES	2011	366	352	14
2	VIERNES	2011	260	259	0
2	JUEVES	2011	310	310	0
2	MIÉRCOLES	2011	296	299	3
2	MARTES	2011	319	318	1

2	LUNES	2011	346	349	3
3	JUEVES	2011	339	328	10
3	VIERNES	2011	233	233	1
3	LUNES	2011	314	315	0
3	MARTES	2011	334	327	7
3	MIÉRCOLES	2011	322	313	9
4	LUNES	2011	312	319	7
4	VIERNES	2011	213	227	14
4	MARTES	2011	308	315	8
4	MIÉRCOLES	2011	277	290	12
4	JUEVES	2011	297	298	1
5	VIERNES	2011	235	235	0
5	JUEVES	2011	374	357	17
5	MIÉRCOLES	2011	315	307	8
5	LUNES	2011	335	333	2
5	MARTES	2011	337	333	4
6	VIERNES	2011	216	214	1
6	JUEVES	2011	299	283	16
6	LUNES	2011	303	299	4
6	MARTES	2011	271	272	1
6	MIÉRCOLES	2011	249	236	13
7	JUEVES	2011	196	194	2
7	MIÉRCOLES	2011	155	164	9
7	VIERNES	2011	163	167	4
7	MARTES	2011	217	214	3
7	LUNES	2011	192	196	5
8	JUEVES	2011	165	163	2
8	LUNES	2011	224	214	10

8	MARTES	2011	187	180	7
8	VIERNES	2011	170	155	15
8	MIÉRCOLES	2011	179	145	34
9	MIÉRCOLES	2011	240	239	1
9	LUNES	2011	273	270	3
9	VIERNES	2011	208	208	0
9	MARTES	2011	286	278	8
9	JUEVES	2011	233	233	0
10	MARTES	2011	297	290	7
10	MIÉRCOLES	2011	283	290	7
10	JUEVES	2011	311	312	1
10	VIERNES	2011	260	262	2
10	LUNES	2011	287	287	1
11	LUNES	2011	339	337	2
11	VIERNES	2011	279	293	14
11	JUEVES	2011	320	314	6
11	MARTES	2011	322	322	0
12	MARTES	2011	285	292	7
12	MIÉRCOLES	2011	242	240	2
12	JUEVES	2011	273	262	10
12	LUNES	2011	286	295	9
12	VIERNES	2011	226	220	6
11	MIÉRCOLES	2011	290	282	8

Tabla 52. Aplicación del modelo y comparativa sobre el año 2011 en el centro de salud de La Magdalena.

El error cometido por el modelo en la predicción del año 2011 es del 2,25%, para el centro de salud de La Magdalena.

Como hemos visto al añadir el porcentaje de población atendida por franja de edad y el nivel económico de la zona, el modelo que predice los pacientes que necesitarán una atención por parte del centro de salud, mejora considerablemente disminuyendo el error absoluto en toda la predicción del año 2011 a 2,41% de media. En la tabla 53 mostramos la comparativa del error cometido antes de añadir los valores espaciales y después:

CENTRO DE SALUD	% ERRORES ABSOLUTOS SIN DATOS DE ETARIOS	% ERRORES ABSOLUTOS CON DATOS DE ETARIOS
FUENTEZUELAS	11,75	3,1
VIRGEN DE LA CAPILLA	11,77	2,34
SAN FELIPE	5,32	2,7
FEDERICO DEL CASTILLO	2,54	1,87
EL VALLE	7,16	2,36
LA MAGDALENA	6,67	2,55
BELEN	6,6	1,99

Tabla 53. Resumen del error real cometido por el modelo en la predicción del año 2011 sobre todos los centros de salud.

El error absoluto medio baja del 7,40% al 2,41%. La mejora del modelo es de un 5%. Este error es similar al que teníamos en el modelo general.

5.1.4. Coeficiente estandarizado de regresión

Otro punto importante es conocer el coeficiente estandarizado del modelo para entender cómo afectan estas nuevas variables en el número de visitas. En la tabla 54 se muestran los coeficientes de regresión del modelo GLM con las variables espaciales incluidas.

Atributo	Valor	Coeficiente Estandarizado	Coeficiente
<Interceptar>		0	724,424
DIA_SEMANA	MIÉRCOLES	0,014	4,765
DIA_SEMANA	MARTES	0,075	25,544
DIA_SEMANA	LUNES	0,102	34,749
DIA_SEMANA	VIERNES	-0,162	-55,203
E0_14		0,078	4,781
E15_24		-0,022	-1,973
E25_34		-0,187	-31,21
E35_44		0,011	1,121
E45_54		0,043	9,735
E55_64		-0,123	-15,28
E65_74		-0,03	-3,857
E75_84		0,102	14,448
HUMEDAD_RELATIVA		0,057	0,449
MAS_85		0,059	11,503
MES	4	-0,001	-0,414
MES	5	0,003	1,663
MES	11	0,012	5,839
MES	3	0,022	10,96
MES	10	0,031	15,246

MES	2	0,039	19,22
MES	12	-0,087	-42,953
MES	6	-0,106	-52,551
MES	9	-0,206	-101,654
MES	7	-0,319	-157,451
MES	8	-0,392	-193,725
NIVEL_ECONOMICO		-0,034	-14,541
CALIDAD DE AIRE	1	0,008	3,611
TEMP_MAXIMA		0,099	1,762
TEMP_MEDIA		0,148	2,949
TEMP_MINIMA		-0,051	-1,188
USUARIOS_ADCRITOS		0,258	0

Tabla 54. Coeficientes de regresión estandarizados.

Como se puede observar a mayor nivel económico de la zona (el coeficiente de regresión es negativo), tenemos menos usuarios que usan los servicios públicos de salud. Esto como hemos comentado anteriormente se puede deber principalmente a que exista un mayor número de seguros privados en este sector de la población. Otro posible motivo puede deberse a que al tener mayor nivel económico, teóricamente también tendrá, un mayor nivel de formación lo que puede redundar a que exista mayor número de pacientes expertos.

En cuanto a los rangos de edad podemos observar en la grafica, como los rangos de edad que aumenta la asistencia al médico son: 0-14, 35-44 y de 45-54. Restan pacientes de 55-64, 64-74, mayores de 85 y sobre todo los pacientes de 24 a 34 años.

Finalmente las variables ambientales (temperatura, humedad y calidad del aire) tienen un comportamiento muy similar a los otros modelos generados.

5.2. Optimización de los modelos mediante la interpolación de las temperaturas

Como hemos visto en nuestro estudio, existe una gran relación entre la temperatura y el número pacientes que asisten a los centros de Salud. Los modelos de nuestro estudio se han generado y se han validaron usando la media de la temperatura de las dos estaciones meteorológicas que hay en la ciudad de Jaén, pero hay que tener en cuenta que en la ciudad de Jaén hay 6 centros de salud, ubicados en zonas distintas donde existen importantes diferencias de alturas y proximidad con zona de montaña, lo que provoca que esa temperatura media utilizada tenga un error importante y esto tiene una influencia negativamente en la precisión de los modelos. El objetivo de esta línea de trabajo es valorar la mejora que podría suponer para los modelos el uso de una temperatura y humedad relativa más aproximada a la realidad.

Para determinar con mayor precisión la temperatura que afecta a cada centro de Salud, se han instalado en Jaén 4 estaciones meteorológicas adicionales a las dos existentes, de tal forma que tenemos distribuida por la ciudad 6 estaciones en total. En la Figura 52 podemos ver la distribución de dichas estaciones. En dicha figura también podemos ver la delimitación de las zonas de influencia de los Centros de Salud de Jaén.

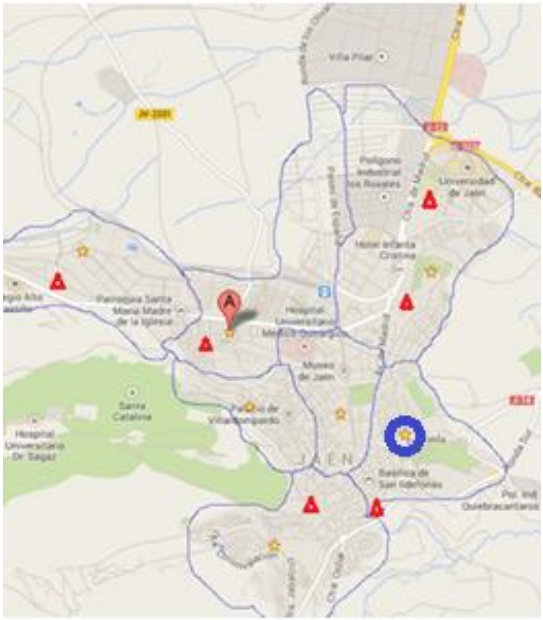


Figura 52. Posicionamiento de las estaciones fijas de Jaén y las nuevas instaladas para medir la temperatura y humedad relativa

Todas las temperaturas y humedades relativas recogidas en las 6 estaciones las insertamos en un SIG y realizamos una interpolación, en la figura 53 podemos ver el resultado de la interpolación para el centro de Salud de Belén (marcado con un circulo azul en la figura 52).

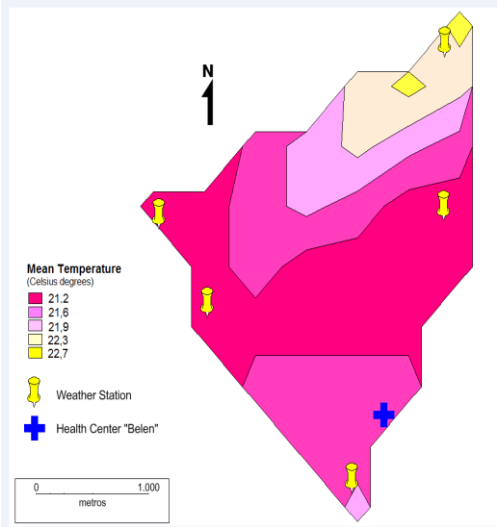


Figura 53. Interpolación de la temperatura media para el Centro de Salud de Belén

Para valorar la mejora en el sistema predictivo que supondría trabajar con temperaturas más precisas, comparamos los resultados de la predicción insertando en el sistema la media de la temperatura y la humedad relativa y por otro lado hacemos el mismo cálculo introduciendo en el sistema la temperatura y humedad relativa interpolada con los datos de las 6 estaciones. Finalmente se compararan los resultados con los datos reales de visitas. Para la comparación usaremos el centro de Salud de belén (marcado en azul en la figura 52).

Al comparar los datos de temperatura de los distintos dispositivos, podemos observar diferencias de hasta 6 grados de unas zonas a otras, como se ha indicado Jaén es una ciudad con una importante diferencia de altura entre zonas. La zona sur tiene mayor altura y está muy próxima a una zona montañosa, mientras que la zona sur es la parte más baja de la ciudad y alinda con campiña, luego son normales estas importantes diferencias de temperatura de unas zonas a otras. En la figura 54 podemos ver las importantes diferencias de medida de unas estaciones a otras en un día concreto (24/09/2014).

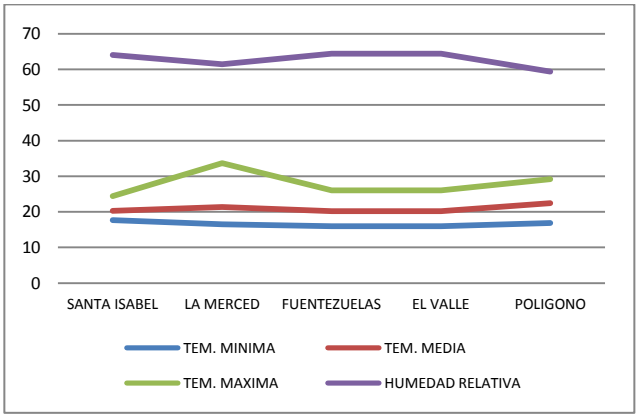
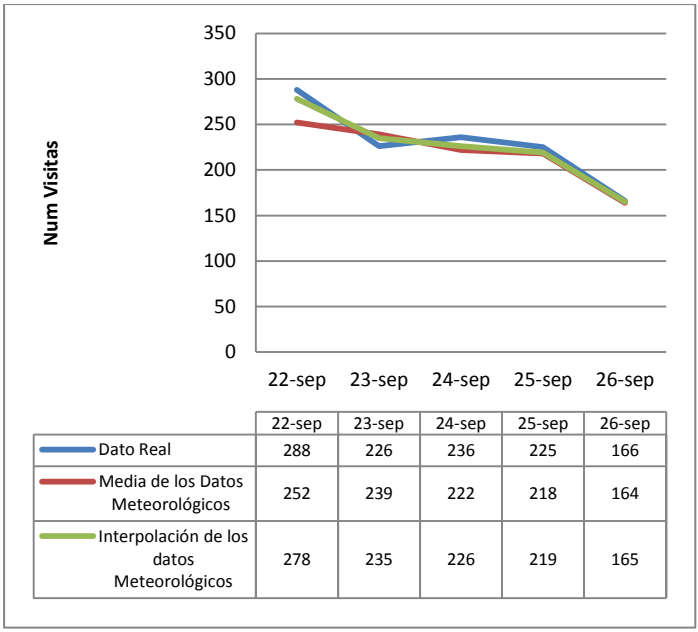


Figura 54. Temperatura y Humedad Relativa medida en las distintas estaciones meteorológicas.

En la Figura 55 podemos ver la comparativa de los resultados obtenidos de los modelos predicativos al trabajar con los valores meteorologías medios y los valores meteorológicos interpolados.



	22-sep	23-sep	24-sep	25-sep	26-sep
Dato Real	288	226	236	225	166
Media de los Datos Meteorológicos	252	239	222	218	164
Interpolación de los datos Meteorológicos	278	235	226	219	165

Figura 55. Comparativa entre el dato real de afluencia de pacientes y resultados de la predicción realizada con la temperatura y humedad media e interpolada.

El error absoluto del modelo predictivo con la media de los datos meteorológicos ha sido de 72 pacientes en una semana, frente a los 36 pacientes de error absoluto del modelo generado con la interpolación de la información. Esto datos indica que el error se ha reducido a la mitad.

5.3. Predicción de las demanda de "Salbutamol" en las farmacias de Jaén

Otra línea importante de nuestros modelos predictivos es que pueden ser usados para predecir el número de demandas que tendremos de un determinado medicamento.

En ciertas épocas del año, en algunas farmacias se agotan ciertos medicamentos, concretamente en febrero del 2014 en Torredonjimeno (pueblo de Jaén) se agoto el Salbutamol. Esto podría evitarse si fuésemos capaces de generar unos modelos predictivos de la demanda que tendrán ciertos medicamentos en función de datos de temperatura, la calidad del aire y el tipo de población atendida por la farmacia. Como hemos visto en nuestro trabajo hay en la literatura numerosas referencias sobre la relación de estos factores con ciertas patologías y también con el uso de fármacos específicos [16, 17, 18, 19].

Al igual que el estudio que hemos realizado para predecir el número de pacientes que asistirán a los centros de salud, hemos realizado otro estudio a modo de prototipo para determinar si es posible predecir la demanda de Salbutamol en una farmacia de Torredonjimeno.

El estudio de minería de datos se ha realizado con información de la demanda de "salbutamol" desde los años 2009 al 2014. El modelo se ha generado con datos del 2009, 2010, 2011 y 2013, y el modelo ha sido validado con los datos del año 2014. Ha sido un estudio supervisado y se han utilizado los algoritmos de MDL para analizar la relación de las variables con el consumo del "Salbutamol" y los algoritmos SVM (con kernel Gaussiano y kernel lineal) y GLM. Los datos utilizados en los modelos han sido los siguientes:

- **Datos de consumo:** número de dispensaciones de "salbutamol". Estos datos han sido proporcionados por varias farmacias.

- **Datos meteorológicos:** Los datos meteorológicos utilizados han sido temperatura máxima, media y mínima, humedad y precipitaciones. Los datos han sido facilitados por la consejería de Medio Ambiente de la Junta de Andalucía.
- **Los niveles de contaminación:** se clasifica como bueno o malo, dependiendo del índice parcial para cada contaminante: sulfuro de dióxido de SO2, partículas, dióxido de nitrógeno NO2, monóxido de carbono CO y O3 ozono. Los datos han sido facilitados por la Consejería de Medio Ambiente de la Junta de Andalucía.
- **El tipo de población que atiende la farmacia**.

El tipo de población que atiende las farmacias se ha obtenido de la misma forma que en nuestro estudio de los centros de Salud, a partir de datos del Instituto Nacional de Estadística (INE, 2014) se ha construido un SIG con una delimitación de la zona de influencia de la farmacia. Finalmente se ha realizado una consulta espacial sobre el SIG para determinar el porcentaje de población atendida por franja de edad. En la tabla 55 podemos ver los resultados de la farmacia estudiada.

Age	Farmacia de Torredonjimeno
0-14	16,85%
14-24	14,60%
25-34	15,04%
35-44	15,99%
45-54	15,55%
55-64	9,52%
65-74	6,51%
75-84	6,48%
> 85	3,45%

Tabla 55. Tipo de población atendida.

Con la ejecución del algoritmo de MDL que calcula el peso de las variables sobre el atributo de target (tabla 56). En esta tabla podemos ver la relación entre la dispensación de "salbutamol" y su relación con el resto de variables (1 representa la relación máxima y 0 ninguna relación).

Variable		Ranking	Peso
Edad	0-14	1	0,631
	15-24	2	0,581
	25-34	3	0,381
	35-44	4	0,331
	45-54	4	0,331
	55-64	4	0,331
	65-74	4	0,331
	75-84	4	0,331
	>85	4	0,331
Mes		5	0,164
Temperatura Minima		6	0,152
Temperatura Maxima		7	0,046
Temperatura Media		8	0,04
Calidad del Aire		9	0,007

Tabla 56. Importancia de atributos

Como se puede observar en la tabla desde un punto de vista estadístico se establece una relación entre ciertos factores ambientales, meteorológicos y el tipo de población que demandan a la farmacia de "salbutamol". Esto es muy importante conocerlo ya que a partir de estas variables podemos hacer una predicción del consumo de este fármaco. También hay que destacar la gran relación que existe entre el consumo del "salbutamol" y el tipo de población atendida por la farmacia, esto es importante ya que hoy en día las ciudades tienen zonas de expansión en las que viven familias jóvenes, por lo que hay más población edad pediátrica y zonas donde viven personas mayores y dependiendo de su localización física variara el consumo de este medicamento.

Finalmente hay que destacar que el algoritmo más preciso en este estudio ha sido GLM, cometiendo un error absoluto del 6% en la predicción. Este dato que a priori no parece muy bueno, es engañoso, ya que el consumo del "Salbutamol" tiene durante el año una gran fluctuación de unos meses a otros, de hasta un 500%, con lo que la predicción del modelo se aproxima mucho a la realidad.

6. DESARROLLO DEL PROTOTIPO DE SISTEMA EXPERTO

Otra parte importante del nuestro trabajo es la implementación de un sistema experto que incorpore todos los modelos generados en este estudio. El objetivo principal es demostrar que este tipo de trabajos teóricos se pueden materializar de forma rápida en productos finales, fáciles de usar y que puedan estar a disposición de los gestores de los recursos sanitarios para ayudarles a tomar las mejores decisiones.

Para nuestro desarrollos usamos SGBD Oracle 11R2. Oracle que nos ofrece todo lo que necesitamos para el desarrollo rápido de este Sistema Experto: la Base de Datos para gestionar la información, Oracle Data Mining que está integrado en la Base de Datos donde tenemos implementados todos los algoritmos de Minería de Datos necesarios y Oracle Aplications Express (APEX) que es un modulo de desarrollo de aplicaciones WEB que también se implementa dentro de la Base de Datos, Apex nos permite desarrollar aplicaciones WEB tanto para PC como para dispositivos móviles. En la figura 56 podemos ver el diseño del sistema.

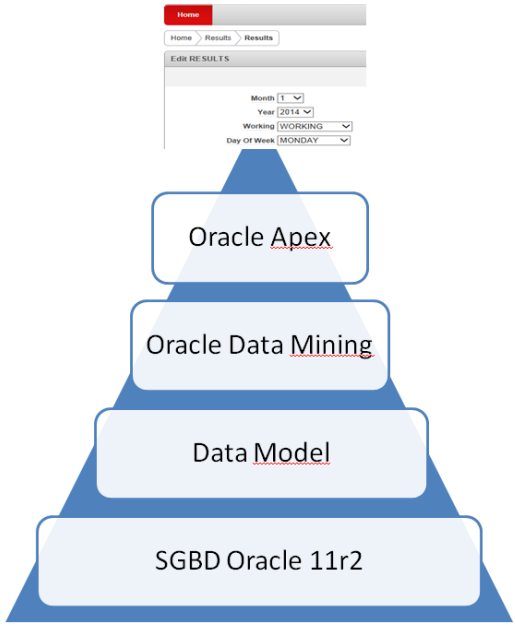


Figura 56. Diseño de las capas del Sistema Experto.

Para realizar el desarrollo de nuestro sistema experto es necesario seguir una serie de pasos que consisten en: Diseño del modelo de datos, Carga de datos para la generación de los modelos Predictivos, Generación de los modelos predictivos con Oracle Data Mining, Desarrollo de la interfaz del sistema experto con Oracle Aplication Expres y diseño y desarrollo del Aprendizaje del Sistema mediante la retroalimentación de datos.

6.1. Diseño del modelo de Datos

Un modelo de datos es solo y exclusivamente un método para diseñar los esquemas de base de datos que posteriormente debemos de implementar en un gestor de bases de datos concreto. Este modelo se representa a través de diagramas y está formado por varios elementos.

Para nuestro sistema el modelo Entidad Relación necesario es muy básico, tan sólo necesitamos dos tablas, Una tabla contendrá los datos históricos de asistencia al médico, datos meteorológicos y calidad ambiental, esta tabla servirá para generar nuestro modelo predictivo. Por otro lado necesitamos otra tabla de resultados para almacenar los parámetros de entrada de la predicción y guardar el resultado. En la Figura 57 mostramos las tablas del modelo y sus atributos.

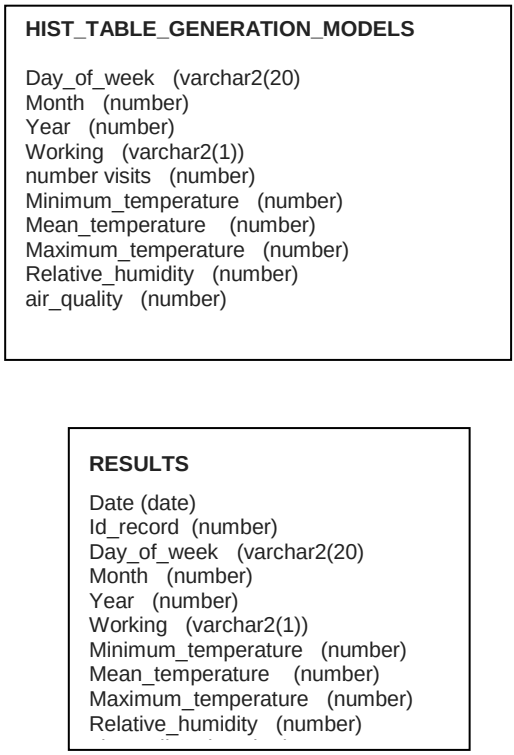


Figura 57. Modelo Entidad-Relación de nuestro Sistema Experto.

6.2. Carga de datos para la generación de los modelos Predictivos

Los datos que usaremos para generar el modelo son los mismos que se usaron para el estudio. Estos datos nos los proporcionan en fichero plano con los campos separados por el carácter "|". Estos ficheros contiene la siguiente información:

- Fichero_datos_visitas.txt.- fecha, número de visitas. Un ejemplo (01/01/2007|1200).

- Fichero_datos_meteorologicos.txt.- fecha, temperatura máxima, temperatura mínima, temperatura media, humedad relativa. Un ejemplo (01/01/2007|3|7|8.2|70).

- Fichero_datos_calidad_aire.txt.- fecha y 0 para días de buena calidad y 1 para días de mala calidad. Un ejemplo (01/01/2007|1).

Para cargar en la Base de Datos esta información usamos una herramienta de Oracle llamada SQL-Loader, esta herramienta permite leer ficheros de texto y cargarlos a las tablas de la Base de Datos. Una vez cargada la información en tablas auxiliares procedemos a agrupar la información siguiendo los criterios que marca el estudio:

- Día de la Semana (Lunes, Martes, ... Domingo).
- Tipo de Día (Laborable, Festivo)
- Mes (1 es Enero, 2 es Febrero, ... Diciembre es 12)
- Año (2007, 2008, 2009, 2010)

6.3. Generación de los modelos con Oracle Data mining

Una vez almacenados los datos, tenemos que generar los modelos con Oracle Data Mining, para ellos usamos la herramienta SQL-Develper.

Todos los modelos son generados de la misma manera con SQL-Developer. Seleccionamos la tabla de Origen (HIST_TABLE_GENERATION_MODELS), aplicamos un filtro para distinguir entre días laborable y no laborable y a continuación seleccionamos el tipo de modelo que queremos generar, en nuestro caso una regresión, seleccionamos el algoritmo (GLM para los días laborable y SVM para los festivos), seleccionamos el atributo target y los atributos que formaran parte del modelo predictivo. En el ejemplo de la Figura 58 mostramos la generación del modelo para las visitas totales en día no laborable, como puede verse es una herramienta totalmente visual.

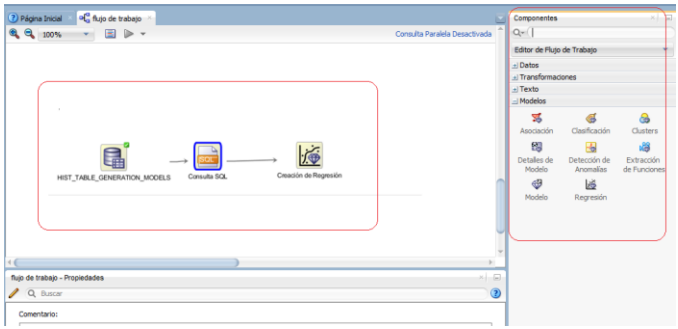


Figura 588. Interfaz de diseño de los modelos de minería de datos.

Una vez generado el modelo con Oracle Data Mining, se almacena en la Base de Datos y puede ser invocado con una simple SQL.

6.4. Desarrollo de la aplicación con APEX

Oracle Application Express o APEX (anteriormente llamado HTML DB) es una herramienta RAD que se ejecuta dentro de la base de datos Oracle. Permite desarrollar aplicaciones WEB de forma segura y rápida. En enero de 2006 el nombre de Oracle HTML DB pasó a ser "Oracle Application Express".

Para desarrollar una aplicación con APEX sólo es necesario abrir un navegador web para conectarnos a APEX. Una vez dentro del entorno de desarrollo podemos acceder a varios asistentes que nos generan páginas web automáticas del tipo REPORT y FORMS basados en las tablas de origen, al seleccionar la tabla nos crea las páginas de tipo FORM con todos los campos y con los botones de insertar, modificar y borrar el registro. En la Figura 59 podemos ver varias pantallas de la interfaz de Desarrollo.

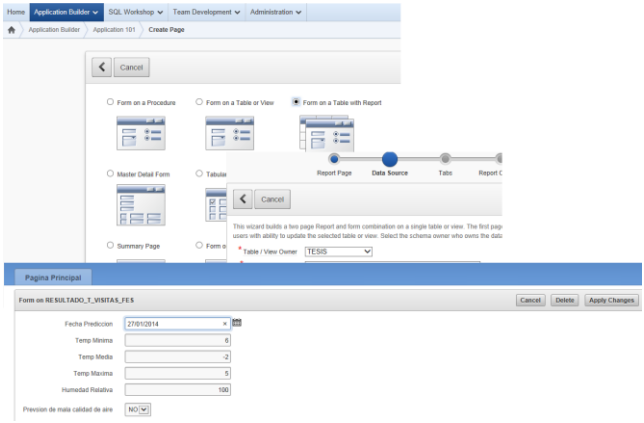


Figura 599. Interfaz grafica de desarrollo de aplicaciones con APEX.

El desarrollo de la interfaz es muy simple y se realiza en pocas horas de trabajo, una vez generada dicha interfaz de manipulación de datos, procedemos a desarrollar la lógica de nuestra aplicación que básicamente consistirá en la invocación de los modelo de data mining. Para ello programamos un trigger en la tabla RESULT de tal forma que cada vez que se haga una inserción o una modificación de un registro, este invoque al modelo y el resultado se almacene en el campo de "Predicción". Aquí mostramos el código fuente del Trigger.

```
CREATE OR REPLACE TRIGGER TGR_MODELO_T_LAB BEFORE
INSERT OR UPDATE ON RESULT FOR EACH ROW BEGIN


  :NEW.FECHA_REALIZADO:= SYSDATE ;

  :NEW.DIA_SEMANA                                    :=
LTRIM(RTRIM(TO_CHAR(:NEW.Date,'DAY')));

  :NEW.MES := TO_CHAR(:NEW.Date,'MM');

  :NEW.AÑO := TO_CHAR(:NEW.Date,'YYYY');

  :NEW.MEDIA_CASOS := NULL;



WITH
"N$10086" as (
       select /*+ inline */
       TO_CHAR(:NEW.DATE,'MM') AS MONTH,
       :NEW.AIR_QUALITY AS AIR_QUALITY
       SYSDATE AS DATE,
       :NEW.Mean_temperature AS Mean_temperature,
       :NEW.Working AS Working,

       LTRIM(RTRIM(TO_CHAR(:NEW.DATE,'DAY')))       AS
Day_of_week,
       :NEW.Relative_humidity AS Relative_humidity,

       TO_CHAR(:NEW.DATE,'YYYY') AS Year,
       :NEW.Maximum_temperature                     AS
```

```
Maximum_temperature,
       :NEW.Maximum_temperature                          AS
Maximum_temperature
                      from
                      DUAL
                      ),
               "N$10073" as (
                                SELECT /*+ inline */

       PREDICTION("TESIS"."REGR_GLM_2_1"   USING   *)
"REGR_GLM_2_1_PRED" INTO :NEW.PREDICTION
               FROM "N$10086" )
               select * from "N$10073";
END;
/
```

Finalmente diseñamos otra pantalla para visualizar el resultado del sistema experto. Para ellos volvemos a usar un asistente de creación de las páginas. Una vez cargados los datos el campo Predicción ya estará cumplimentado (esto se ha calculado automáticamente mediante la ejecución del trigger). Para mejorar la interfaz y ayudar al usuario final en su decisión, desarrollamos una grafica donde se mostraran los datos históricos que nos ha servido para generar el modelo (número de visitas de los pacientes en año anteriores), para ellos solo tenemos que añadir un componente tipo grafico y programar una SQL que agrupe los datos.

El resultado de esa sentencia SQL seria el mostrado en la figura 60.

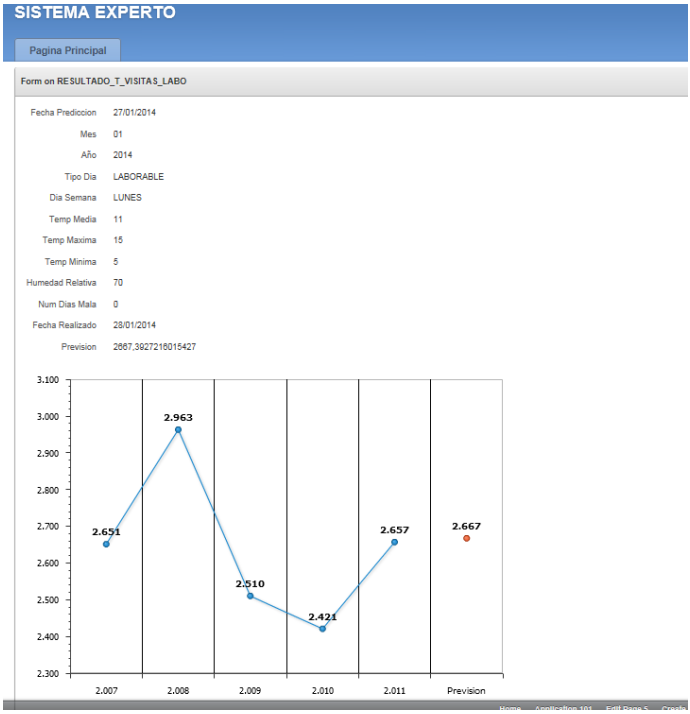


Figura 60. Pantalla de resultados del Sistema Experto desarrollado con APEX.

6.5. Aprendizaje del Sistema mediante la retroalimentación de datos

Es fundamental que nuestro sistema experto pueda aprender para garantizar que se va adaptando a los cambios del uso de los servicios sanitarios, para ellos publicaremos un campo en la aplicación donde el usuario final podrá insertar manualmente el dato real de los pacientes que finalmente asistieron a los centros de Salud, esto garantiza un aprendizaje supervisado inductivo. En la Base de Datos crearemos un trigger que se

ejecutara cuando se modifique este campo (real_data_views), el trigger insertara la información en la tabla: HIST_TABLE_GENERATION_MODELS, que es tabla en la que se basa la generación de nuestros modelos´. Finalmente el trigger ejecutar el procedimiento para volver a generar el modelo con el nuevo dato incorporado.

7. DISCUSIÓN

En este trabajo se ha estudiado la eficacia de los distintos algoritmos de minería de datos para generar modelos predictivos para su uso en los servicios de atención primaria la salud en Jaén.

Como se ha podido comprobar en este estudio el uso de los servicios de salud es muy diferente en función del tipo de día (Festivo o Laborable). En los días festivos sólo se proporcionan servicios de urgencias y el número de visitas es muy pequeño y regular, mientras que los días laborables se atienden a los pacientes con necesidades clínicas y administrativas, lo que hace que los datos tengan una alta estacionalidad, repitiéndose un comportamiento parecido en función del día de la semana (lunes, martes, miércoles, etc.) y el mes.

Para un modelo de predicción óptima es necesario abordar el estudio con los datos separados en dos cohortes independientes (Festivos y Laborables).

Se ha podido observar que es necesario usar algoritmos diferentes para cada cohorte: para los días laborables el mejor algoritmo es GLM, mientras que para los días de festivos el algoritmo SVM con núcleo lineal es el más eficiente.

Otra cuestión importante en el inicio del estudio fue determinar la forma más eficiente de agrupar la información para obtener el modelo más óptimo. En el estudio se ha analizado la información, observándose una gran estacionalidad de los datos. La mejor manera de preparar la información es agruparla por días de la semana (lunes, martes, etc.), mes y año.

Por otro lado, en el estudio era fundamental establecer los atributos que realmente se relacionan con el número de visitas, y hemos observado que ciertos atributos que aparentemente deberían de estar relacionados con la afluencia de pacientes a los centros de Salud y que sin embargo al analizarlos con el algoritmo MDL se ha comprobado que no influyen en las visitas al médico. Estos atributos han sido: La precipitaciones, en teoría parece lógico pensar que los días lluviosos restaría pacientes a los centros, sin embargo ésto no se ha observado en el estudio en la ciudad de Jaén, es posible que en otras ciudades este atributo sí tenga relación.

En este estudio también hemos visto que cuando tenemos un modelo general para toda la ciudad de Jaén y queremos extrapolarlo a un centro de salud concreto, el error del modelo se dispara, ésto es debido a que hay variables locales que modelan la idiosincrasia de los centros. Para ajustar el modelo general se han añadido unas variables espaciales locales como son la edad de los pacientes atendidos, número de pacientes adscritos y el nivel económico. En el estudio se puede observar cómo los factores más influyentes para nuestra predicción son el tipo de población atendida, el nivel económico y el número de pacientes adscritos, es decir los factores locales de cada Centro de Salud, seguido de los factores más generales como: mes del año, temperatura, calidad ambiental, humedad relativa y día de la semana.

En el análisis de los coeficientes estandarizados de regresión, hemos visto qué factores suman o restan pacientes en los Centros de Salud. El tipo de población atendida es fundamental siendo los pacientes pediátricos los que más influyen positivamente en las vistas al médico. Otra franja de edad que suma un buen número de pacientes es la población geriátrica de 75- 84. Por el contrario los pacientes afectan negativamente a las vistas es la población de 25-34 años y de 55-65. Al contrario de lo que se podría pensar, la población mayor de 85 años no influyen positivamente en las visitas al centro de salud.

Otra reflexión importante es que según el coeficiente de regresión se puede deducir que a mayor nivel económico de la zona de influencia del centro de salud tenemos menor número de visitas a los Centros de Salud. También hay que destacar que el factor de calidad del aire malo influye en el aumento de visitas al Centro de Salud.

En este estudio, también hemos evaluado la eficacia de diferentes algoritmos de minería de datos, para predecir la demanda de los servicios administrativos en los centros de salud de atención primaria en Jaén. En el caso de demandas administrativas hay factores ambientales que no afectan a la afluencia de pacientes, tales como la humedad relativa y la precipitación, pero otros como la temperatura juega un papel importante en el sistema, sin embargo, a diferencia de lo que ocurre en otras ciudades, la calidad del aire no tiene peso en el sistema, esto puede deberse a que en Jaén los valores de calidad del aire no experimentan variaciones significativas a lo largo del año. En la predicción de tareas administrativas se generaron dos modelos: uno para las recetas y otro para la emisión de certificados, en ambos casos, los algoritmos que obtienen los mejores

resultados son los modelos lineales. En el caso de la emisión de recetas el algoritmo más eficiente es la SVM con kernel lineal, mientras que el GLM ha funcionado mejor para predecir el número de solicitudes de certificados médicos. Esta predicción puede ser muy útil para mejorar y optimizar los recursos humanos de los centros de salud, separando las agendas en tres bloques de tareas: demanda clínica, renovación de recetas y emisión de certificado. El dimensionamiento de cada bloque de agenda lo determinaría nuestro sistema aplicando los modelos generados a tal efecto.

Un ejemplo práctico de los beneficios que supondría este modelo sería el siguiente: En Jaén un médico dedica un promedio de 1.100 horas al año para pasar consulta; esto significa que con la configuración actual de agenda (intervalos de tiempo de 5 min.) asiste a aproximadamente 13.200 usuarios al año. Con estos datos son necesarios en Jaén 46 médicos para atender a los pacientes (en 2011 603.440 pacientes fueron atendidos).

Si nos centramos en las tareas administrativas con el ajuste actual se necesitan (5 minutos por visita) 1.013.780 minutos, mientras que si fijamos las agendas de receta con 1 minuto y certificados médicos con 3 minutos, el tiempo necesario se reduciría a 353.242 minutos. Esto sería una optimización del sistema en el caso de las recetas de un 169,4% y en el caso de los certificados médicos del 56,4%. Esta mejora significaría que en 2011 la demanda administrativa en Jaén ha sido atendida por 15,3 médicos, mientras con el cambio de modelo podría haber sido atendida por sólo 5,35 médicos, esto significa que 10 médicos se liberan de un total de 46, lo que implica una mejora global del 21,73% del sistema.

Usando estos modelos nos permitirían separar las agendas de los centros de salud entre las tareas administrativas y clínicas y esto podría contribuir a lograr una mejora significativa en la gestión de recursos humanos en el centro de salud. Este ahorro de tiempo puede ser utilizado por los profesionales médicos para realizar otras tareas, como la investigación o mejorar la atención clínica que requieren más esfuerzo y tiempo.

En cuanto al desarrollo del sistema experto hay que indicar que hoy en día, hay muchas herramientas en el mercado que integran los módulos de base de datos y minería de datos que faciliten el desarrollo de tales sistemas. Es esencial para nosotros basamos en este tipos de herramientas de desarrollo rápido para que trabajos de investigación en minería de datos se conviertan en los sistemas expertos que pueden ser utilizados por

los profesionales a los que están destinados. Como se ha visto en este trabajo, con un poco de conocimiento de SQL y las herramientas adecuadas, un sistema experto puede desarrollarse muy rápidamente.

8. CONCLUSIONES

La primera conclusión que podemos sacar de este trabajo es que es posible predecir de forma general los pacientes que acudirán a los centros de salud de Jaén con un error absoluto del 2,29%. Esta información puede ser clave para los responsables de recursos técnicos y humanos de los centros de salud puedan dimensionarlos correctamente.

Se ha podido demostrar que varios atributos meteorológicos influyen de forma importante en el número de pacientes que necesitan una atención médica. Los atributos relacionados son: Temperatura mínima (con un valor de 0,32), Temperatura Media (con un valor de 0,28), temperatura máxima (con un valor de 0,23) y la humedad relativa (con un valor de 0,1). Los valores indicados son reportados por el algoritmo MDL, indicando 1 que tiene la máxima relación y 0 que no tiene ninguna relación con el número de pacientes.

En cuanto a la calidad ambiental, también se ha demostrado su relación. El atributo número de días con mala calidad de aire tiene un valor generado por MDL de 0,03. Este valor es pequeño, pero muy significativo ya que tiene su peso en el modelo a pesar de que Jaén es una ciudad poco contaminada donde predomina los días con buena calidad del aire.

Para el modelo general de pacientes que acudirán a los centros de Salud de Jaén hay que destacar que al analizar sus coeficientes estandarizados de regresión nos indican que a nivel de los días de la semana, los miércoles, jueves y sobre todo viernes resta pacientes a los centros de salud, es decir que tiene un valor negativo (-0,093, -0,173, -0,432 respectivamente), mientras que el lunes añade un número significativo de pacientes (0,108). En la Figura 61 podemos ver una grafica con el coeficiente de regresión y podemos hacernos una idea de cómo afecta al modelo en función del día de la semana.

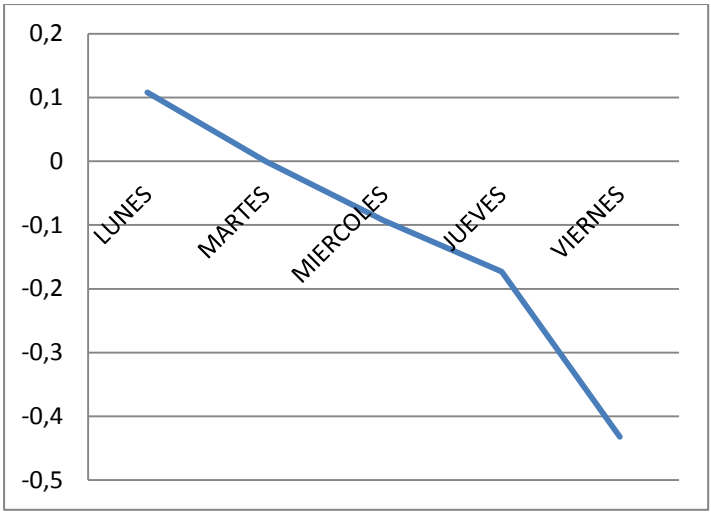


Figura 61. Coeficientes estandarizados de regresión por día de la Semana.

Si analizamos los coeficientes estandarizados por mes, se demuestra que durante los meses de julio, agosto y septiembre, el número de pacientes disminuye (-0,368, -0,61, -0,171). Lógicamente, esto se debe a que coinciden con el período de vacaciones y las altas temperaturas. Por otro lado febrero y octubre son los meses con mayor asistencia de los pacientes a los centros de salud (0,356, 0,314). En la Figura 62 podemos comprobar cómo afectan este atributo a la predicción del número de visitas.

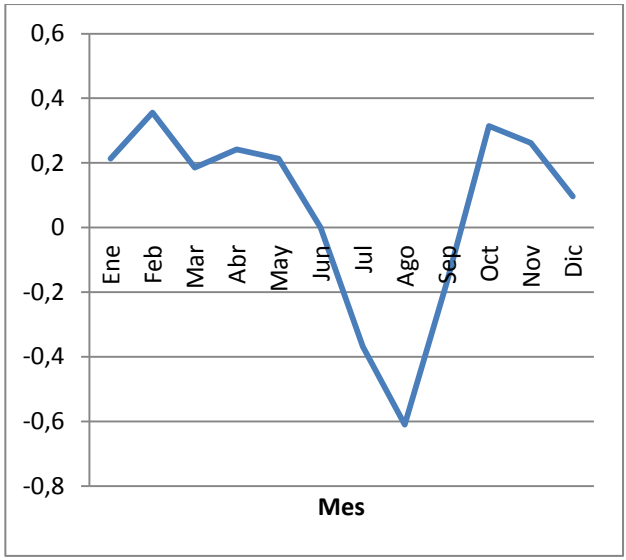


Figura 62. Coeficientes estandarizados de regresión por mes.

A nivel de datos ambientales podemos ver el peso importante que tiene en el número de visitas la temperatura máxima, mínima y media, con coeficientes estándar estimados: -0,175 y 0,752 a 0,168, respectivamente. Finalmente, en el nivel de calidad del aire se ha detectado que es el elemento del modelo con menos influencia, con un coeficiente estimado de 0,068. Sin embargo, durante los días con mala calidad del aire se convierte en un atributo importante.

En cuanto a algoritmos de regresión podemos concluir que para predecir el número de pacientes que necesitan atención médica, el algoritmo más eficiente en el caso de días de laborables es el GLM, en contraste con el caso de los días festivos donde el algoritmo más eficiente es SVM con kernel lineal.

Otro logro importante de este trabajo es que se ha diseñado un modelo fiable, utilizando factores muy básicos de calidad ambiental y meteorológica, siendo precisamente estos atributos la entrada del sistema experto. Estos datos básicos están disponibles en varias páginas web con

una antelación de hasta 10 días. Esto es muy importante ya que si hubiésemos utilizado otros factores más complejos hubiera supuesto que nuestro sistema se quedara en plano meramente teórico, ya que si consideramos un atributo que no es posible conocer con antelación, el usuario final del sistema experto sería incapaz introducirlos en el sistema para obtener la predicción, con lo cual sería un sistema teórico pero imposible de utilizar en la práctica.

Al analizar los coeficientes estandarizados de regresión para los días festivos (Figura 63), podemos sacar varias conclusiones. El día que más pacientes van a urgencias son aquellos cuyos días festivos coinciden en lunes, con un coeficiente de 0.188, esto es lógico ya que el Sábado y el Domingo no se presta servicio de atención primaria y muchos pacientes el Lunes se ve obligada a ir a urgencias. En cuanto a los días del fin de semana (sábados y domingo) observamos que el sábado tiene un coeficiente mayor que el domingo, esto también se puede explicar ya que en domingo muchos usuarios enfermos prefieren pedir cita y esperar al lunes para ir a su médico de cabecera. En la Figura 58 podemos ver cómo afecta el día de la semana a la afluencia de pacientes en días festivos, estos datos pueden ser muy importantes para los gestores de los recursos sanitarios ya que pueden ayudarles a entender mejor el funcionamiento de sus servicios a nivel de demanda por parte de los pacientes.

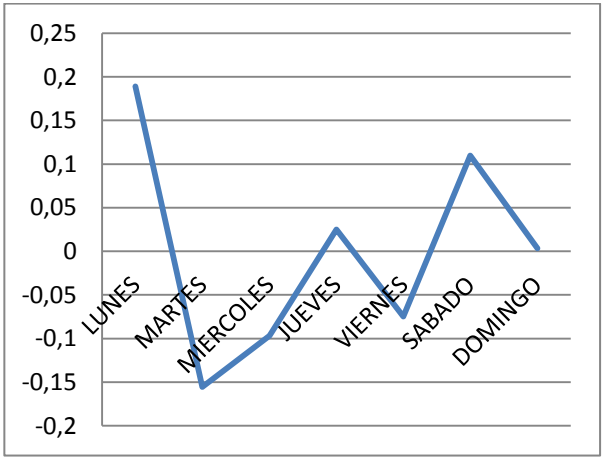


Figura 63. Coeficientes estandarizados de regresión por día de la semana para Festivos.

En cuanto al análisis del modelo de los días Festivos por meses, podemos observar que al igual que en días laborales los meses de Julio, Agosto y Septiembre tienen un coeficiente negativo, mientras que los meses de Febrero y Enero son los que tienen mayor coeficiente positivo. En la Figura 64 podemos ver la grafica de los coeficientes por meses.

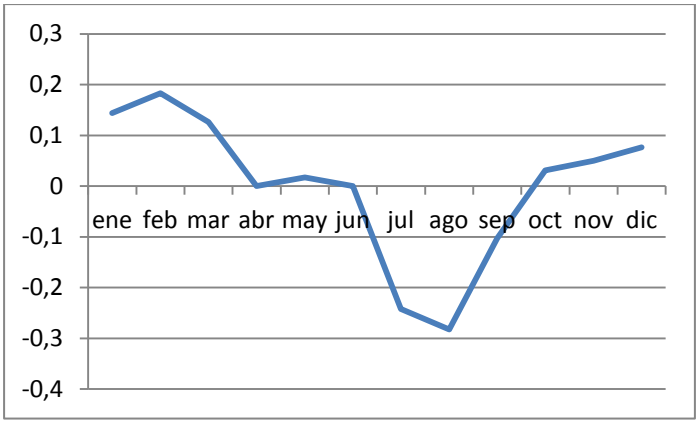


Figura 60. Coeficientes estandarizados de regresión por mes para Festivos.

También podemos afirmar que para poder realizar predicciones en un centro de salud concreto es necesario tener en cuenta variables locales que modelen la idiosincrasia de la zona que atiende dicho centro de Salud, en nuestro estudio hemos analizado rango etario y datos económicos del entorno, y al calcular el peso de estas variables sobre el modelo, con el algoritmo MDL, se ha podido comprobar que las variables espaciales del estudio son las que tiene más peso: Usuarios adscritos, porcentaje de pacientes por edad y nivel económico.

En cuanto al tipo de población atendida se observa que cuanto mayor es el número de población de 0-14 años inscrita a un Centro de Salud, mayor número de visitas atiende, por el contrario el rango de población que resta visitas es 25 a 34 años usan menos los servicios médicos. En la Figura 65 podemos ver de forma grafica los rangos de edad que restan y suman pacientes.

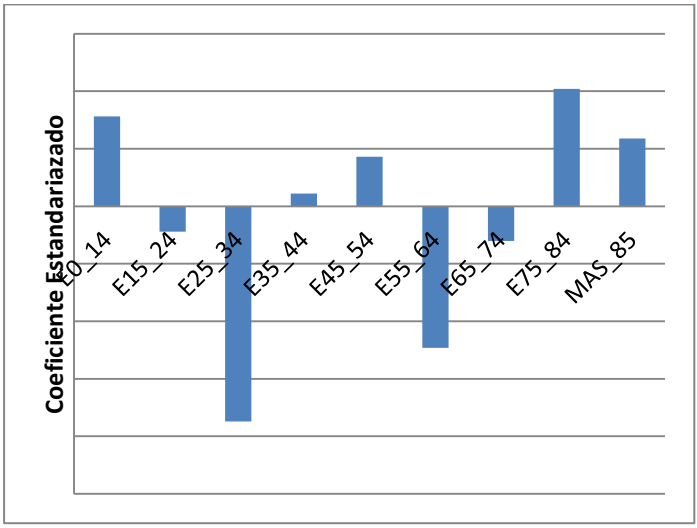


Figura 61. Coeficientes estandarizados de regresión por rango etario.

Se ha concluido también que a mayor nivel económico hay menos visitas, al analizar el coeficiente estandarizado obtenemos un valor negativo (-0,034). Esto como hemos comentado anteriormente puede ser lógico debido a que este tipo de población puede tener un mayor número de seguros privados y otro posible motivo puede ser que al tener mayor nivel económico, teóricamente también tendrá un mayor nivel de formación lo que puede redundar a que exista mayor número de pacientes expertos.

Otro logro importante de este estudio ha sido generar 2 modelos que son capaces de predecir la demanda administrativa de los centros de salud, con un error absoluto del 4,46%. Teniendo en cuenta que hay variaciones significativas en el uso de este tipo de demanda de hasta un 350%, este modelo predictivo podría ser una herramienta útil para la correcta gestión y la planificación de las agendas de cita. Al comparar los resultados de nuestro modelo con los datos reales de 2011, hemos demostrado que podemos obtener una optimización del 21,73% de los recursos médicos.

También se ha demostrado que en todos los modelos el atributo de precipitaciones no se ha considerado ya que contrariamente a lo que se podría pensar que un día lluvioso restara pacientes a los centros de salud. Esto no se ha podido demostrar y en todos los modelos el algoritmo MDL ha indicado que no hay relación entre las visitas y las precipitaciones.

Otra conclusión importante es que es fundamental poder avanzar e investigar en la mejora de la calidad de las variables que utilizan los modelos predictivos, en este caso en nuestro trabajo hemos visto que al trabajar con datos de temperatura y humedad relativa de varias estaciones y utilizar algoritmos de interpolación (en vez de una simple media), nos permite disminuir el error absoluto a la mitad. Es fundamental tener modelos predictivos fiables, pero no es menos importante tener calidad en las variables que intervienen, por eso es importante realizar trabajos e investigaciones que vayan en la línea de mejorar la calidad de las variables que intervienen en los modelos de minería de datos.

También se ha podido demostrar que nuestros modelos pueden ser útiles para predecir el consumo de ciertos principios activos que se dispensan en las farmacias. En nuestro trabajo hemos realizado un estudio piloto con el consumo del "Salbutamol", demostrándose la relación del consumo de este principio activo con las variables utilizadas en nuestro estudio (tipo de población atendida, datos meteorológicos y calidad ambiental). El error absoluto en la predicción ha sido del 6% con el algoritmo GLM. Hay que destacar que la variación de la demanda de este principio activo varía de unos meses a otros en un 500%, lo que hace que el dato de error del modelo sea mínimo en la predicción.

En la parte final de nuestro trabajo hemos desarrollado un sistema experto, en este desarrollo hemos necesitado muy pocas horas de trabajo y como hemos visto en su desarrollo no se necesita un gran conocimiento en minería de datos. Las herramientas que hay disponibles hoy en día en el mercado nos permiten simplificar y acortar los tiempos de este tipo de desarrollos. El sistema ha sido desarrollado en 16 horas de trabajo, es decir, en menos de 3 días laborables.

El uso de estas herramientas y modelos puede ser de gran ayudar para mejorar la gestión de los centros de salud. Primero desde el punto de vista económico ya que los gestores de los centros de salud pueden

dimensionarlos en función de la demanda real que tendrán, evitando que haya sobre dimensionamiento de recursos. Otro aspecto importante es que la aplicación de los modelos pueden ayudar a aumentar la satisfacción del paciente, ya que si se detecta con antelación que va a haber un aumento en la demanda, superando la capacidad del centro de salud, los gestores pueden actuar y evitar estas situaciones que provocaría que un paciente no pudiera obtener una cita para ser visitado por su médico. Finalmente el uso de este modelo podría reducir las visitas a los servicios de urgencias para casos no urgentes, ya que en caso de desbordamiento de los servicios de atención primaria, el paciente no conseguirá cita viéndose obligado a ir a los servicios de urgencias.

En definitiva que el uso de estas herramientas puede ayudar a dimensionar adecuadamente los servicios de atención primaria de salud ayudando a que el sistema sanitario público sea más eficiente y sostenible.

9. FUTURAS LÍNEAS DE TRABAJO

Este trabajo es un primer pilar para seguir avanzando en mejorar los sistemas santuarios basándonos en la minería de datos y en los sistemas de información geográficos. Actualmente quedan abiertas varias líneas de este trabajo que detallo a continuación:

- **Predecir el número de patologías más comunes.** En nuestro estudio hemos trabajado para generar unos modelos globales, pero actualmente en Diraya se codifican todas las patologías en Diagnósticos CIE 9 y CIE10. Una mejora seria obtener las patologías más comunes y generar unos modelos específicos para predecirlas. Esto puede ser de gran ayuda para los gestores de los recursos ya que hay ciertas patologías que conllevan revisiones, pruebas diagnosticas y prescripción de ciertos fármacos. Toda esta información puede ser muy valiosa para la gestión óptima de recursos técnicos y humanos.

- **Avanzar en el estudio para predecir los principales principios activos que se serán solicitados por los pacientes.** En nuestro estudio hemos realizado una prueba sobre la predicción del "Salbutamol", este es uno de los principales principio activos que son dispensados por las farmacias en ciertas épocas del año. Una línea que queda abierta es profundizar en este estudio y hacerlo extensivo a otros principios activos.

10. DIFUSIÓN DEL TRABAJO

Nuestro trabajo ha sido difundido en revistas y congresos. A continuación en la tabla 57 se relaciona el tipo de trabajo y el estado de los mismos a 21 de Noviembre de 2014.

Titulo	Tipo	Revista	Estado
An improvement in the appointment scheduling in primary health care centers using data mining.	Articulo	**Journal of Medcal Systems (Factor Impacto en 2013: 1.372)**	Publicado
Importancia de los CRM Sanitarios en las Pandemias y alertas Sanitarias	Articulo	**Journal Atención Primaria (Factor Impacto en 2013: 0.894)**	Publicado
An Expert System based on GIS and Data Mining to predict the flow of patients to primary health care centers.	Articulo	**Information Systems Frontiers (Factor de Impacto en 2013: 0,761)**	Bajo Revisión
Designing a model to predict patient flow in primary health care centers.	Articulo	**Health Informatics Journal (Factor de Impacto en 2013: 0,787)**	Bajo Revisión
Geospatial factors to model the prediction of patient flow to health centers. A case study in Jaen, Spain	Articulo	**Geospatial Health (Factor de Impacto en 2013: 1,65)**	Bajo Revisión
Use of meteorological, environmental and spatial variables to predict drug use	Poster	**Congreso Internacional. Medical Informatics Europe 2015**	Pendiente de Aceptación

Using geographic information systems to improve the accuracy of data mining models to predict the flow of patients to the health centers	Comunicación	**Congreso Internacional. Medical Informatics Europe 2015**	Pendiente de Aceptación

Tabla 57. Descripción de la difusión del trabajo.

11. BIBLIOGRAFÍA

1. DIRAYA
   http://www.juntadeandalucia.es/servicioandaluzdesalud/principal/documentosacc.asp?pagina=pr_diraya
   (accessed 17 oct 2013)
2. Parvatiyar, Atul and Jagdish N. Sheth (2001), Customer Relationship Management: Emerging Practice, Process, and Discipline," Journal of Economic & Social Research, 3 (2), 1-35.
3. Czepiel, John A. (1990), "Service Encounter and Service Relationships: Implication for Research," Journal of Business Research 20, 13-21.
4. Cubillas, J. J., Ramos, M. I., Feito, F. R., González, J. M., Gersol, R., & Ramos, M. B. (2014). Importancia de los Customer Relationship Management (CRM) sanitarios en las pandemias y alertas sanitarias. Atención Primaria.
5. Selker HP: Coronary care unit triaje decision aids: Hew do we know they work? Am J Marl 1989;87:491-492.
6. Weingarten SR, Ermann 8, Riedinger MS, et al: Selecting the best triaje rule for patients hospitalized with chest pain. Am J Mad 1989;87:494-500.
7. Elena Fernández Valdivieso, Susana Montesinos Sanz, M. José de Miguel Peláez, Margarida Alié Xufré (2008). Papel de enfermería en el triaje de urgencias en atención primaria. Atención Primaria.
8. Haux R (2006) Introduction of health information system to progress in the organization of health care. International Journal of Medical Informatics. 75(3-4): 268-281
9. Rozenblum RJ, Donzé PM, Hockey, Guzdar E, Labuzetta MA, Zimlichman E et al (2013) International Journal of Medical Informatics. 82(3): 141-158

10. Lichtner V, Venters W, Hibberd R, Cornford T, Barber N. The fungibility of time in claims of efficiency: The case of making transmission of prescriptions electronic in English general practice. Int. J. Med. Inform. 2013;82(12):1152-1170.

11. Starfield B. Primary care and health. A cross-national comparison. JAMA 1991;266: 2268-2271.

12. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Production and Operations Management. 12(4): 519-549

13. Bailey NTJ (1952) A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. In: Journal of the Royal Statistical Society. B14: 185–199

14. Kaandorp GC, Koole G (2007) Optimal outpatient appointment scheduling. Health Care Management Science. 10: 217–229

15. Ho C, Lau H (1992) Minimizing total cost in scheduling outpatient appointments. Management Science. 38(12): 1750–1764

16. Dawson J, Weir C, Wright F et al (2008) Associations between meteorological variables and acute stroke hospital admissions in the west of Scotland. Acta Neurologica Scandinavica. 117: 85–89

17. Oiamo TH, Luginaah IN, Atari DO, et al. Air pollution and general practitioner access and utilization: a population based study in Sarnia, 'Chemical Valley,' Ontario. Environ. Health 2011;10:71.

18. Donaldson G.C, Goldring J.J, and Wedzicha J.A. Influence of Season on Exacerbation Characteristics in Patients With COPD. Chest 2012;141:1:94-100.

19. Ferrari U, Exner T, Wanka ER, et al. Influence of air pressure, humidity, solar radiation, temperature, and wind speed on ambulatory visits due to chronic obstructive pulmonary disease in Bavaria, Germany. Int. J. Biometeorol 2012;56:1:137-143.

20. Tseng CM, Chen YT, Ou SM, et al. The Effect of Cold Temperature on Increased Exacerbation of Chronic Obstructive Pulmonary Disease: A Nationwide Study. PLoS One 2013;8:3.

21. Keatinge WR, Donaldson GC, Bucher K, et al. Cold exposure and winter mortality from ischemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. Lancet 1997;349:1341-1346.

22. Dawson J, Weir C, Wright F, et al. Associations between meteorological variables and acute stroke hospital admissions in the west of Scotland. Acta Neurol Scand 2008;117:2:85-89.

23. Rothwell PM, Wroe SJ, Slattery J, et al. Is stroke incidence related to season or temperature?. Lancet 1996;347:9006:934-936.

24. Harper, P. R., and Gamlin Adfaf, H. M., Reduced outpatient waiting timeswith improved appointment scheduling: a simulation modeling approach. OR Spectrum. 25:207–222, 2003.

25. Bourdais, S., Galinier, P., and Pesant, G., Hibiscus: A constraint programming application to staff scheduling in Health Care. In: Rossi, F. (Ed.), CP 2003, LNCS 2833. Springer, Berlin Heidelberg, pp. 153–167, 2003.

26. Cayirli, T., Veral, E., and Rosen, H., Designing appointment scheduling systems for ambulatory care services. Health Care Manage Science. 9:47–58, 2006.

27. Kaandorp, G. C., and Koole, G., Optimal outpatient appointment scheduling. Health Care Manage Science. 10:217–229, 2007.

28. Glowacka, K. J., Henry, R.M., andMay, J. H., A hybrid data mining/simulation approach for modeling outpatient no-shows in clinic scheduling. Journal of the Operational Research Society. 60:1056–1068, 2009.

29. Liu, N., Ziya, S., and Kulkarni, V. G., Dynamic scheduling of outpatient appointments under patient no-

shows and cancellations. Manufacturing & Service Operations Management. 12(2):347–364, 2010.

30. Greenfield S, Blanco DM, Elashoff RM, Ganz PA. Patterns of care related to age of breast cancer patients. JAMA 1987; 257(20): 2766–70.

31. Chu J, Diehr P, Feigl P, Glaefke G, Begg C, Glicksman A, Ford L. The effect of age on the care of women with breast cancer in community hospitals. J Gerontol 1987; 42(2): 185–90.

32. Mor V, Guadagnoli E, Silliman RA, Weitberg A, Glicksman A, Goldberg R, Cummings R, Masterson-Allen S. Influence of old age, performance status, medical, and psychosocial status on management of cancer patients. In: YancikR, YatesJW, editors. Cancer in the elderly: approaches to early detection and treatment. New York, NY: Springer, 1989: 127–46.

33. Ganz PA, Lee JJ, Sim M, Polinsky ML, Schag C. Exploring the influence of multiple variables on the relationship of age to quality of life in women with breast cancer. J Clin Epidemiol 1992; 45(5): 473–85.

34. Fu Zetiana, Xu Fengb, Zhou Yunb, Zhang XiaoShuana (2005). "Pig-vet: a web-based expert system for pig disease diagnosis" Expert Systems with Applications, Volume 29, Issue 1, July 2005, Pages 93–103.

35. Mohamed, A. (2001). "Knowledge based approach for productivity adjusted construction schedule." Expert Systems with Applications, 21(2), 87-97.

36. Boussabaine, A. and Duff, A. (1996). "An expert-simulation system for construction productivity forecasting." Building Research & Information, 24(5), 279-286.

37. Moselhi, O. and Nicholas, M. (1990). "Hybrid Expert System for Construction Planning and Scheduling." Journal of Construction Engineering and Management, 116(2), 221-238.

38. R http://www.r-project.org/ (accessed 17 sep 2013).

39. Weka http://www.cs.waikato.ac.nz/ml/weka/ (accessed 17 sep 2013).

40. MySql http://www.mysql.com/ (accessed 17 sep 2013).

41. Oracle http://oracledocs.shu.ac.uk/oracle/B28359_01/install.111/b32006.pdf (accessed 17 sep 2013)

42. Oracle http://oracle.com (accedido 17 sep 2013)

43. REDIAM. http://www.cma.junta-andalucia.es/medioambiente/site/web/rediam (accessed 17 sep 2013)

44. http://es.wikipedia.org/wiki/Sistema_de_informaci%C3%B3n_geogr%C3%A1fica (accedido 17 de Sep 2013)

45. http://www.mapinfo.com/ (accedido en Marzo 2014)

46. Zhang R, Zhang S, Lan Y, Jiang J (2008) Network Anomaly Detection Using One Class Support Vector Machine. In: International multiconference of engineers and computer scientists

47. Grünwald P. Advances in Minimum Description Length: Theory and Applications. In: Jae Myung, Mark A. Pitt, Peter D. Grunwald, eds. MIT Press. Retrieved 2010-07-03.

48. Allen DM, Cady FB (1982) Analyzing Experimental Data by Regression. In CA: Lifetime Learning Publications. Belmont

49. Belsley DA, Kuh E, Welsch RE (1980) Regression Diagnostics: In John Wiley & Sons. New York

50. Cameron AC, Trivedi PK (1988) Regression Analysis of Count Data. In: Cambridge University Press. Cambridge

51. Corinna C and Vapnik V. Support-Vector Networks. J. Mach. Learn. Res. 1995; 20:3:273-297.

52. Press WH, Teukolsky SA, Vetterling WT, et al. Section Support vector machines. In Press WH, Teukolsky SA, Vetterling WT and Flannery BP, eds. Numerical recipes:

The Art of Scientific Computing. New York: Cambridge University 2007:16.5

53. Cristianini N and Shawe-Taylor J. An introduction to support vector machines and other kernel based methods. In Cristianini N and Shawe-Taylor J. Cambridge: Cambridge University Press 2000: 6.

54. Dobson AJ. An Introduction to Generalized Linear Models. In Chatfield C and Zidek J, eds. Texts in Statistical Science Series. Chapman & Hall/CRC 2002:90-100.