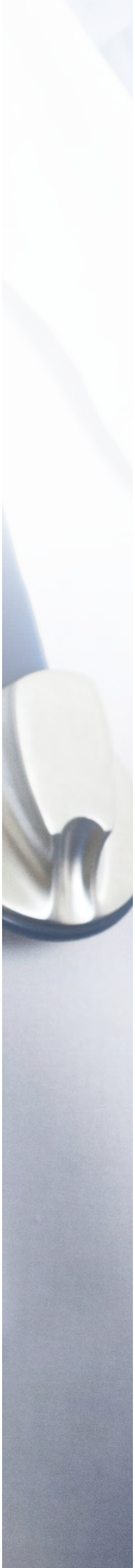
instituto
de ingeniería
del conocimiento

Optimizando la sanidad: segmentación y predicción en el entorno Salud

Autor: Ángela Fernández Pascual

Data Scientist del IIC en Health & Energy Predictive Analytics



Optimizando la sanidad: segmentación y predicción en el entorno Salud

Abstract

El futuro de la sanidad está orientado hacia el análisis de grandes cantidades de datos de salud recogidos de diversas fuentes. Las tecnologías *Big Data* y los estudios observacionales basados en datos reales obtenidos de la práctica clínica diaria (*Real World Data, RWD*) permiten tener una visión más global y tienen el potencial de mejorar la toma de decisiones en presencia del paciente, y reducir así la incertidumbre en el diagnóstico. En este trabajo se presentan dos de las tareas básicas para optimizar y mejorar el sistema sanitario: la segmentación de pacientes y la predicción de reingresos.

Palabras clave:

Big Data en salud; Real Word Data; Segmentación de pacientes; Predicción de reingresos; Asistencia sanitaria; Medicina personalizada.

Autor: Ángela Fernández Pascual

Data Scientist del IIC en Health & Energy Predictive Analytics

instituto de ingeniería del conocimiento

Optimizando la sanidad: segmentación y predicción en el entorno Salud

instituto de ingeniería del conocimiento

Autor: Ángela Fernández Pascual

Introducción

Big Data es un término muy presente y relacionado con el sector Salud. En esta área existe una gran cantidad de información relacionada con pacientes, enfermedades y centros sanitarios que, bien tratada y organizada, puede aportar mucho valor a los profesionales y a los gestores de centros sanitarios, y permitiría aumentar la calidad del servicio, favorecer una mejora del sistema y, en definitiva, beneficiar a los pacientes y a su salud.

Hoy en día existen numerosas fuentes que recogen información: las historias clínicas electrónicas, los aparatos de telemedicina que se utilizan para diagnosticar pacientes, los análisis clínicos o los *wearables* que registran datos sobre constantes vitales de salud como la tensión o la actividad física. Más allá de la información proporcionada por los ensayos clínicos aleatorizados exploratorios (ECA), los nutricionales o la genómica, existen además otros datos que aportan valor añadido a esta información, como los datos epidemiológicos o los datos procedentes de la práctica clínica real, conocidos como *Real World Data (RWD)*. Este tipo de datos se relaciona con el concepto de **medicina personalizada** y podrían por ejemplo ayudar a predecir si una persona tiene riesgo de padecer cáncer, diabetes o una enfermedad del corazón.

El cambio tecnológico que subyace tras la **historia clínica electrónica** y la capacidad de procesamiento de grandes cantidades de datos hace posible responder a lagunas de conocimiento que los ensayos ECA no permitían resolver y que, por el contrario, se están resolviendo actualmente con el análisis de los RWD.

El análisis de los RWD puede beneficiar a distintos ámbitos (García López et ál., 2014):

- Pueden identificar anticipadamente a los pacientes crónicos en riesgo de descompensación. De forma simplificada, es equivalente a **estratificar la población** identificando a aquellos individuos que deban ser incluidos en programas específicos de atención. Nos acercamos, por tanto, a la medicina de precisión, cuya idea intuitiva es que no se espera a que el paciente acuda enfermo al profesional, sino que se prevé esta situación, siendo los profesionales los que van a buscar al paciente antes de que enferme para ser tratado.

- Facilitan la toma de **decisiones clínicas en tiempo real**, analizando casos similares y proponiendo alternativas proactivas. De esta forma se consigue reducir la variabilidad en la práctica médica, se podría comparar la calidad de la atención recibida por pacientes en distintos centros o por diferentes médicos. Sería posible definir indicadores de medida de calidad que, tras su evaluación, permitan detectar necesidades y ayuden a la planificación de estrategias de mejora que aporten calidad y sostenibilidad al sistema sanitario.
- Permiten trasladar información directamente a los pacientes. Con esto se pretende **empoderar al paciente**, pasando este a tener un rol más activo en sus propios cuidados y, quizás, una mayor efectividad para modificar estilos de vida, controlar factores de riesgo y mejorar la adherencia a los tratamientos.

El gran reto del *Big Data* en salud reside en ser capaces de analizar este gran volumen de datos heterogéneos, procedentes de distintas fuentes, para obtener conocimiento que pueda reportar beneficios. Hoy en día es posible trabajar con estas grandes cantidades de información a gran velocidad gracias al progreso de la tecnología. Más aún, hoy en día es posible obtener valor de estos datos de forma que ayuden a **optimizar la sanidad**. Esto es gracias a técnicas de minería de datos y aprendizaje automático.

El sector sanitario necesita **nuevas herramientas tecnológicas** de *Big Data* o de inteligencia de negocio y nuevas capacidades funcionales. En el Instituto de Ingeniería del Conocimiento (IIC) se dispone de un amplio recorrido utilizando este tipo de técnicas en distintos ámbitos profesionales. Gracias a ello, somos capaces de aplicar técnicas propias de análisis de la información que permiten realizar tareas que ayudan a una mejor gestión, así como a obtener un mejor tratamiento y atención del paciente. Cuanto mejor se gestione la salud y la enfermedad, menos se gastará y, además, la población estará más sana.

La predicción de ingresos, la hiperfrecuentación médica o la segmentación de pacientes son algunas de estas tareas que, en definitiva, permiten crear sistemas de alertas, de predicción de necesidades o de generación de recomendaciones que faciliten a los profesionales su trabajo, y ayuden a una **optimización del sistema sanitario** en todos los sentidos.

Este trabajo se centra en dos de estas tareas básicas para optimizar y mejorar el sistema sanitario: la segmentación de pacientes y la predicción de reingresos.

Segmentación de pacientes

Actualmente la población está envejeciendo, la esperanza de vida es mayor y, por tanto, el modelo sanitario está cambiando, pues prevalecen las **enfermedades crónicas** y la **pluripatología** frente a las enfermedades puntuales.

Sin embargo, el sistema sanitario español está diseñado para atender principalmente episodios agudos de enfermedades, por lo que, al aumentar los casos crónicos, en muchas ocasiones no se atiende a estos pacientes convenientemente, y quedan expuestos a riesgos como la duplicación u omisión de servicios, o la falta del control requerido para su correcta evolución.

Para mejorar la capacidad de respuesta y la asistencia a este tipo de pacientes es fundamental detectar a estos individuos y sus diversas patologías, de manera que se pueda llevar a cabo una **atención multidisciplinar**, detectar sus futuros reingresos y planear los diversos controles necesarios. Con este tipo de estrategias se conseguiría, además, reducir el gasto hospitalario, al realizar una mejor planificación, evitando la duplicación innecesaria de servicios y asignando a cada paciente los recursos necesarios en función de sus necesidades reales.

Una de las posibles respuestas a este problema consiste en la correcta estratificación o **segmentación de los pacientes** en función de niveles de riesgo. Tener la población clasificada permite al gestor sanitario incrementar la toma de decisiones proactivas.

Además, el análisis de la información obtenida desde distintos **perfiles médicos** es clave para conseguir adelantarse a las necesidades tanto de los pacientes como de los centros de salud. Segmentar

la población, por ejemplo, por patologías crónicas, estableciendo niveles entre los distintos pacientes, permite dirigir programas de actuación preventivos sobre el grupo de mayor riesgo.

Este tipo de análisis no es útil solo en hospitales y centros de salud, también puede suponer una gran ventaja para servicios socioasistenciales. En este ámbito, la segmentación de individuos permite identificar, por ejemplo, a aquellas personas con riesgo de cronificación de pobreza o exclusión social que necesitarían ayuda con más urgencia. También permitiría detectar a aquellas personas con riesgo de dependencia de las ayudas públicas.

La **estratificación de la población** no es un proceso fácil. Entre otros factores, no es un proceso estático, pues las características de los pacientes evolucionan con el paso del tiempo, y con ellas cambia su nivel de riesgo. Por ello, las clasificaciones obtenidas deben reajustarse periódicamente.

Para realizarlo adecuadamente y de forma óptima, de manera que sea un proceso eficaz y repetible, es necesario aplicar **técnicas de aprendizaje automático**. En concreto, para solucionar este problema será necesario aplicar técnicas de *clustering* o análisis de grupos.

Existen algunas aplicaciones en el mercado para realizar segmentación, como la diseñada por la Universidad Johns Hopkins en Estados Unidos: *Adjusted Clinical Groups (ACG)*. Esta herramienta ha sido probada, entre otros, por la Comunidad Autónoma del País Vasco (Orueta et ál., 2013) para segmentar la población de sus hospitales, con resultados satisfactorios. En esta herramienta se crean grupos en función de la probable duración o recurrencia de una enfermedad, así como grupos por diagnóstico, lo que permite reducir costes y **mejorar la atención al paciente.**

Por su parte, el Instituto de Ingeniería del Conocimiento (IIC) es capaz de analizar todas estas características, así como los factores de utilización y demanda de los distintos servicios hospitalarios o asistenciales. Ofrece un servicio personalizado, adaptado a las necesidades del demandante, que permite segmentar la población en grupos con características y niveles de riesgo similares. El IIC cuenta con una amplia experiencia en el uso de técnicas y algoritmos de aprendizaje automático, lo que le permite realizar un **análisis óptimo y eficaz de datos** provenientes de distintas fuentes de información y obtener a partir de ellos variables relevantes para clasificar a la población.

Predicción de reingresos

Los reingresos hospitalarios de pacientes se han considerado como un problema de **calidad asistencial** y tienen un impacto económico importante en el sistema de prestación de servicios de salud. Se dice que un paciente reingresa cuando se produce una nueva entrada en un hospital en un periodo de tiempo muy breve (inferior a 30 días) desde el ingreso original. Puede tratarse de un ingreso programado o no y puede ser en el mismo hospital en el que ingresó por primera vez o en otro diferente. Según el Servicio de Medicina Interna del Hospital de Navarra (Alonso Martínez et ál., 2001), podemos **clasificar estos reingresos** en cuatro tipos:

- reingreso por complicaciones del ingreso previo,
- recurrencia de la enfermedad,
- adherencia al tratamiento
- o una nueva enfermedad.

Cada una de estas causas tiene un **nivel de riesgo** distinto y puede considerarse un problema de calidad o no. Por ejemplo, si el paciente reingresa por una nueva enfermedad, el servicio prestado en la visita anterior no se considera de mala calidad. Además, a la hora de clasificar y evaluar un reingreso se debe tener en cuenta a los enfermos crónicos, los cuales requieren un criterio especial.

Predecir un reingreso se puede traducir en una reducción de costes, una mejora de la asistencia sanitaria y un aumento en la calidad del servicio, siempre que se evite el reingreso del paciente. El objetivo primordial de este tipo de estudios es, por tanto, identificar aquellos pacientes que tienen un mayor riesgo de agudización y, por tanto, de hospitalización.

Para poder abordar este problema, se debe contar con un buen repositorio de datos médicos que permita entrenar un buen **modelo de predicción.** Gracias a la digitalización de los datos, hoy en día se cuenta con bases de datos cada vez más completas y unificadas entre hospitales.

Para trasladar el problema de **predicción de reingresos** al campo del aprendizaje automático, se debe reformular, en este caso, como un problema de clasificación binaria. Se va a definir la variable objetivo como aquella que nos indica si el paciente reingresa o no en menos de 30 días.

Antes de poder decidir el modelo más adecuado para la predicción se deben definir las variables de entrada. Las variables más utilizadas hasta ahora para solventar este problema (Futoma et ál., 2015) han sido: la edad, el sexo, la raza, si es un paciente de la sanidad pública, los días de estancia en el hospital y el número de ingresos en el último año. Por otro lado, aunque no se considere como una variable de forma directa, también es importante tener en consideración el **diagnóstico del paciente**, pues este marcará de manera importante el riesgo de reingreso. Según Futoma et ál. (2015), esta variable se utiliza para definir modelos independientes según el tipo de diagnóstico.

Asimismo, antes de **entrenar un determinado modelo**, suele ser conveniente realizar un preproceso de los datos. En esta línea se podrían filtrar los datos para, por ejemplo, eliminar del modelo aquellos casos de pacientes que acabaron en defunción. También sería interesante poder combinar los casos de solape de reingresos en distintos hospitales, normalmente debido a traslados de uno a otro.

En el citado artículo, se presenta una completa y acertada comparativa de modelos con los que abordar el problema de **predicción de reingresos.** Entre ellos, se propone hacer una clara división entre aquellos modelos aplicados a toda la muestra disponible o aquellos aplicados a grupos de diagnóstico, como se ha indicado anteriormente. La batería de modelos probados comprende:

- **Regresión Logística (LR)** (Duda et ál., 2000): Este método es el más usado en la literatura existente acerca de la predicción de reingresos, por ser un método sencillo de aplicación directa.
- **Regresión Logística con selección de variables multipaso (LRVS)**: Esta modificación del modelo anterior, donde se realiza una selección previa de variables utilizando una heurística, se presenta en He et ál. (2014).
- **Regresión Logística Regularizada (RLR)**: Dentro de estos métodos, en los que se añade un término de regularización al término de error, se ha probado *Ridge Regression* (Hoerl & Kennard, 1970), *Lasso* (Tibshirani, 1996), y *Elastic Net* (Zou & Hastie, 2005), que combina los dos anteriores.
- **Bosques Aleatorios (RF)** (Hastie et ál., 2009): Este método de clasificación se halla dentro del paradigma del aprendizaje basado en conjuntos y parte de entrenar un gran número de árboles binarios de clasificación que después combina, siendo capaz de encontrar separaciones no lineales en los datos. Son modelos sencillos y fáciles de parametrizar.
- **Máquinas de Vectores Soporte (SVM)** (Schölkopf & Smola, 2002): Este tipo clásico de modelos no lineales se ha aplicado en la literatura tanto con kernel lineal como polinómico (Yu et ál., 2013). A pesar de ser modelos clásicos y muy explotados, asignar bien el valor a los parámetros involucrados sigue siendo un tema controvertido.

- **Redes Profundas (DL)** (Bengio, 2009): Son métodos que encuentran fronteras no lineales en datos complejos. Su principal problema es que son pesados de entrenar y tienen muchos parámetros cuyo valor es difícil de fijar.

Para determinar el mejor método, el siguiente paso es definir una buena medida de bondad de los modelos probados. Al tratarse de un problema de clasificación en el que, además, interesa **evaluar el riesgo de reingreso** de cada paciente, parece que lo más adecuado es trabajar con curvas ROC *(Receiver Operating Characteristic*, o Característica Operativa del Receptor). Este tipo de curvas es una representación gráfica del ratio de verdaderos positivos (sensibilidad) frente al ratio de falsos positivos (1-especificidad), que varía según el umbral de discriminación.

Para tener una medida de bondad de cada método se utilizará el área bajo la curva ROC (AUC-ROC), medida estándar que nos permitirá ordenar los distintos modelos desarrollados por la confianza validada.

En base a estos criterios y según el estudio realizado por Futoma et ál. (2015), parece que actualmente los modelos más apropiados para abordar el problema de la predicción de reingresos son los bosques aleatorios, la regresión logística regularizada y las redes profundas. En concreto, las redes profundas se presentan como los **modelos más prometedores** para solucionar este problema, aunque como ya se ha comentado, tienen como desventaja su difícil parametrización y uso, al tratarse de modelos complejos aún en estudio y desarrollo. Asimismo, en el artículo de Futoma se destaca que estos modelos ofrecen mejores resultados cuando se aplican de forma local, es decir, se entrenan modelos específicos para cada grupo de pacientes con una misma enfermedad, frente a aquellos que se aplican sobre toda la muestra de manera indiscriminada.

El Instituto de Ingeniería del Conocimiento domina estas áreas y modelos, y ofrece un **servicio de predicción de reingresos** adaptado a las necesidades y datos disponibles en los distintos hospitales.

Conclusiones

El *Big Data* presente en salud supone un gran reto: optimizar la sanidad y mejorar las condiciones de salud, tanto de cara al paciente como pensando en el conjunto de la población. Con este objetivo en mente pueden usarse los datos disponibles en las distintas bases de datos hospitalarios para realizar, por ejemplo:

- **La segmentación de pacientes**, que permitiría mejorar la capacidad de respuesta y la asistencia médica a pacientes crónicos o personas con necesidades socioasistenciales, entre otros.
- **La predicción de reingresos** de pacientes, que permitiría una reducción de costes, una mejora de la asistencia sanitaria y un aumento en la calidad del servicio.

Estas dos tareas son posibles gracias a la aplicación de metodologías de **aprendizaje automático y minería de datos.** El Instituto de Ingeniería el Conocimiento domina estas áreas y ofrece en el entorno Salud un amplio abanico de soluciones adaptadas a las necesidades y datos disponibles en cada centro.

Agradecimientos

Me gustaría agradecer a Julia Díaz y Ana González la ayuda prestada y las sugerencias realizadas sobre este trabajo, así como al equipo de Gestión de Contenidos del IIC por la revisión y mejora del mismo.

Referencias

- Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*. 56, 229-238. Elsevier.
- García López, J. L., del Llano Señarís, J. E., del Diego Salas, J., & Recalde Manrique, J. M. (2014). *Aportación de los "Real World Data (RWD)" a la mejora de la práctica clínica y del consumo de recursos de los pacientes.* Fundación Gaspar Casal. Madrid.
- He, D., Matthews, S. C., Kalloo, A. N., & Hutfless, S. (2014). Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association*. 21, 272–279. OUP Journals on behalf of AMIA. UK.
- Orueta, J. F., Mateos, M., Barrio, I., Nuño, R., Cuadrado, M., & Sola, C. (2014). Estratificación de la población en el País Vasco: resultados en el primer año de implantación. *Atención Primaria.* 45 (1), 54–60. Elsevier.
- Yu, S., Esbroeck, A.V., Farooq, F., Fung, G., Anand, V., & Krishnapuram, B. (2013). Predicting readmission risk with institution specific prediction models. *ICHI '13, 2013 IEEE International Conference on Healthcare Informatics.* Proceedings. 415-420. Philadelphia, PA, USA.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Springer.*
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning.* 2 (1), 1-127. Y. Bengio, Canada.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 67 (2), 301-320. Wiley. New York, USA.
- Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press. Cambridge, MA, USA.
- Alonso Martínez, J. L., Llorente Díez, B., Echegaray Agara, M., Urbieta Echezarreta, M. A., & González Arencibia, C. (2001). Reingreso hospitalario en Medicina Interna. *Anales de Medicina Interna.* 18 (5), 28-34. Aran Ediciones. Madrid.
- Duda, R., Hart, P., & Stork, D. (2000). *Pattern Classification* (2nd Ed.). Wiley. New York, USA.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 58 (1), 267-288. Wiley. New York, USA.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 12 (12), 55-67. ASA. UK.

©ADIC

Siguenos en:

C/ Francisco Tomás y Valiente, nº 11
EPS, edificio B, 5ª planta
UAM Cantoblanco
28049 Madrid, España.

Tel.: (+34) 91 497 2323
Fax: (+34) 91 497 2334
iic@iic.uam.es
www.iic.uam.es