**Análisis de readmisión en hospital de pacientes con diabetes**

J. Berenguel Noguera1, A. Chamizo González2, S. Domínguez González3

1 Estudiante Ingeniería Biomédica, Universidad Rey Juan Carlos, Madrid, España, j.berenguel.2017@alumnos.urjc.es

2 Estudiante Ingeniería Biomédica, Universidad Rey Juan Carlos, Madrid, España, a[.chamizo.2018@alumnos.urjc.es](mailto:m.fernandeza.2018@alumnos.urjc.es)

3 Estudiante Ingeniería Biomédica, Universidad Rey Juan Carlos, Madrid, España, s.dominguezg.2017@alumnos.urjc.es

**Resumen**

*La diabetes mellitus tipo II es una enfermedad cuya prevalencia sigue en aumento en la población mundial y que puede tener consecuencias considerablemente perjudiciales para la salud del paciente si el tratamiento no es planificado y seguido de forma óptima. Como en la última década las técnicas de Machine Learning han estado muy involucradas en el ámbito de diagnóstico y apoyo a la toma de decisiones clínicas con una tasa de éxito muy elevada, se propone la implementación de distintos modelos para realizar un análisis acerca de la readmisión en los hospitales a pacientes diabéticos. Estos modelos serán implementados en scickit-learn y en Pyspark, comparando las prestaciones ofrecidas y la velocidad de cómputo que conlleva previa y posteriormente a la selección de características.*

1. **Introducción**

La diabetes mellitus es un trastorno metabólico hiperglucémico cuya prevalencia mundial en 2021 rondaba el 9.8% de la población entre 20 y 79 años según datos de la International Diabetes Fundation (IDF). Las proyecciones de esta organización sitúan esta cifra en un 11.2% , unos 783.7 millones de personas, para el año 2045. La diabetes tipo II es su forma más común, ya que abarca entre un 85 y un 95% de todos los casos de diabetes en el mundo [1,3]. Entre otras complicaciones, la diabetes mellitus puede causar insuficiencia renal, ceguera, infarto y amputaciones [2], y la mortalidad de esta condición no es despreciable, siendo de 6.7 millones de fallecidos en 2021 [1].

Se trata de una enfermedad cuya peligrosidad se ve gravemente magnificada por conductas y estilos de vida malsanas en cuanto a alimentación y a actividad física, y que puede tener consecuencias desastrosas para el paciente si no se planifica el tratamiento con vistas a futuro y con el seguimiento diario más riguroso posible. Esto es de especial relevancia en el caso de los pacientes con diabetes mellitus tipo II, ya que el efecto radica tanto en la secreción como en la actuación de la insulina. En sus etapas tempranas se evidencia una resistencia de los tejidos a la insulina , mientras que en las etapas tardías aparece, como consecuencia de esto, un déficit en la secreción de insulina [3]. Este proyecto estará enfocado a la planificación de pacientes con diabetes mellitus tipo II.

El personal clínico cuenta con una amplia gama de fármacos para el tratamiento de la diabetes mellitus tipo II, con diferentes indicaciones terapéuticas, posologías, formas farmacéuticas y formas de administración. Uno de los fármacos mejor estudiados y de mayor uso en pacientes de diabetes mellitus tipo II es la metformina [4], así como la repoglinida, la cloropropamida o la glibenclamida, entre otros. Cada fármaco y cada preparación están indicadas en pacientes con diferentes condiciones de hospitalización y seguimiento, pudiendo así aportar información importante para la planificación de su tratamiento.

En la última década, las técnicas de Machine Learning se han visto involucradas en el ámbito del diagnóstico y el apoyo a la toma de decisiones clínicas con una tasa de éxito muy elevada [5]. Estos métodos permiten establecer una línea basal en torno a decisiones médicas (incluidas las relacionadas con la planificación de tratamientos) en forma de predicciones sobre características comunes a los pacientes para los cuales el algoritmo se ha entrenado, que luego puede ser verificada y aprobada por el facultativo. Esto permite escalar las cargas de trabajo en los servicios que emplean estas aplicaciones, así como obtener tasas de error consistentes y modelables.

Dado que:

* La información relacionada con los datos personales, condiciones de hospitalización, derivaciones de especialidad y pruebas de laboratorio de un paciente de diabetes mellitus tipo II aportan información importante que el personal clínico puede utilizar para la planificación de tratamiento de cara a posibles reingresos del paciente en un futuro, pudiéndose utilizar como información a priori.
* Las indicaciones terapéuticas y posología de las distintas medicaciones utilizadas para tratar esta condición guardan relación directa con la duración proyectada para el tratamiento en cuestión.
* Un algoritmo de Machine Learning tiene la capacidad de establecer relaciones causa-efecto entre características asociadas a ejemplos de un conjunto de datos mediante la optimización iterativa de las ponderaciones de dicho algoritmo, posibilitando las predicciones a posteriori sobre nuevas entradas.

Por tanto, la hipótesis del proyecto es que la utilización de un algoritmo de Machine Learning permitirá establecer un modelo predictivo que clasifique a los pacientes entrantes de diabetes mellitus tipo II en función del tiempo estimado de retorno al hospital una vez dados de alta.

**2. Material y métodos**

**2.1. Preprocesado**

La base de datos está compuesta por datos de 100000 pacientes recogidos durante 10 años en 130 hospitales de Estados Unidos. Incluye hasta 50 características (antes del preprocesado) y todos los datos son de pacientes diabéticos de cualquier tipo a los que se les administró medicación. La información sobre las características se puede consultar en la tabla [6].

La base de datos obtenida es una base de datos muy desordenada, por lo que el preprocesado es la tarea más complicada y la que mayor tiempo requiere. Esto se debe a que se ha tenido que hacer un estudio de cada característica para poder ponerlas en contexto, ya que se contaba con características de distinta naturaleza como integers, categóricas, etc. Además, una gran parte de los datos eran valores nulos, NaN, y se ha utilizado MICE para la imputación de los valores obtenidos mediante regresión lineal.

Por otro lado, se ha tenido que realizar un One Hot Encoding para las características ordinales como, por ejemplo, la edad.

Finalmente se obtienen 70 categorías, número superior al inicial, sin embargo, se evita la hiperdimensionalidad.

**2.2. Características**

Al obtener 70 características tras el preprocesado, es conveniente hacer una selección de las mismas para que el entrenamiento y la predicción sean lo mejores posible. Para esta selección de características se utilizan distintos estadísticos como chi cuadrado, información mutua, el test ANOVA o PCA. Tras utilizar estos métodos, se seleccionan las 10 mejores características, excepto en PCA que se seleccionan solamente los 4 primeros autovectores, lo que hace que las prestaciones no mejoren, pero el tiempo de cómputo disminuye considerablemente. Además también se realizan mapas de correlación U de Theil.

**2.3. Modelos**

Se implementan hasta 8 modelos de scikit-learn como esembles, árboles de decisión, regresión logística, Naïve Bayes y perceptrón multicapa; además de 3 modelos implementados en pyspark como random forest, perceptrón multicapa y regresión logística.

Estos modelos se implementan previa y posteriormente a la selección de características para comparar tanto las prestaciones que ofrecen los distintos modelos como su velocidad de cómputo y, como se prioriza la velocidad frente a la performance, se decide entrenar los modelos con las 10 mejores características seleccionadas.

**3. Resultados y discusión**

Los modelos que mejores prestaciones ofrecen de los implementados en scikit-learn son el AdaBoost, el perceptrón multicapa y la regresión logística, mientras que el mejor de los implementados en pyspark es el random forest.

Por otro lado, se puede destacar que el mejor modelo en relación prestaciones y velocidad es el Naïve Bayes, ya que tarda alrededor de 5 segundos en ejecutarse frente a los 30 segundos del random forest de pyspark, y ambos modelos tienen prestaciones muy similares (alrededor de 0.56 de accuracy).

Por último, remarcar que, en general, las prestaciones son considerablemente bajas ya que las prestaciones se encuentran en torno a un 0.6 de accuracy, un 0.7 de AUC, un 0.55 de F1-score y un 0.6 de recall.

**4. Conclusiones y limitaciones**

Los modelos implementados no ofrecen buenas prestaciones y esto se puede deber a distintos motivos como la necesidad de realizar una ingeniería de características más profunda y compleja, implementar una mayor variedad de modelos (como redes neuronales más complejas).

Sin embargo, se puede destacar la gran interpretabilidad en el sentido del valor de este proyecto, ya que es claramente diferenciable qué características son más predictoras y aportan más valor.

Además, la implementación de los modelos no es costosa computacionalmente, ya que hay mucha mejora en velocidad en la implementación de los modelos sin emplear ingeniería de características.

**5. Referencias**

[1] <https://diabetesatlas.org/data/en/world/>

[2] <https://www.idf.org/component/attachments/attachments.html?id=984&task=download>

[3] <https://www.revespcardiol.org/en-treatment-diabetes-mellitus-general-goals-articulo-13037902>

[4] <https://www.ncbi.nlm.nih.gov/books/NBK279506/>

[5] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7141410/>

[6] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996476/table/tab1/?report=objectonly>