CACiC 2023
CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN

Primeras Experiencias en la Identificación de Personas con Riesgo de Diabetes en la Población Argentina utilizando Técnicas de Aprendizaje Automático

Enzo Rucci[1], Gonzalo Tittarelli[2], Franco Ronchetti[1] , Jorge F. Elgart[3], Laura Lanzarini[1], and Juan José Gagliardino[3]

III-LIDI, Facultad de Informática, UNLP – CIC. La Plata (1900), Bs As, Argentina
{erucci,fronchetti,laural}@lidi.info.unlp.edu.ar
Facultad de Informática, UNLP. La Plata (1900), Bs As, Argentina
CENEXA, Facultad de C. Médicas, UNLP-CONICET. La Plata (1900), Bs As, Argentina
{jelgart,jjgagliardino}@cenexa.org

Resumen La detección de Diabetes Tipo 2 (DT2) y prediabetes (PDM) representa un verdadero desafío para la medicina debido a la ausencia de síntomas patogenómicos y a la falta de conocimiento de los factores de riesgo asociados. Si bien existen algunas propuestas de modelos de aprendizaje automático que permiten identificar a personas en riesgo, las características de esta enfermedad hacen que uno que resulte adecuado para una población, no necesariamente lo sea para otra. Este artículo propone desarrollar y evaluar modelos predictivos que permitan identificar personas con riesgo de DT2 y PDM específicos para la población argentina. Partiendo de un cuidadoso preprocesamiento de la base de datos, se generaron dos datasets particulares considerando el compromiso entre cantidad de registros y de variables disponibles. Luego de aplicar 5 diferentes modelos de clasificación, los resultados obtenidos muestran que algunos de ellos obtuvieron muy buenos rendimientos para ambos datasets. En particular, RF, DT y ANN demostraron gran poder de clasificación, con altos valores en las métricas consideradas. Considerando la vacancia de herramientas de este tipo para la población argentina, este trabajo representa el primer paso hacia modelos más sofisticados.

Keywords: salud pública · enfermedad crónica · machine learning

1. Introducción

La Diabetes Tipo 2 (DT2) es una enfermedad crónica caracterizada por niveles elevados de glucemia que se manifiesta cuando el páncreas endocrino es incapaz de producir la cantidad de insulina suficiente que requieren sus tejidos [11]. Debido a su creciente prevalencia en combinación con su elevado costo de atención [4,29], constituye un serio problema de salud pública, por lo que se han realizado grandes esfuerzos por desarrollar estrategias efectivas para su prevención y tratamiento oportuno, así como para evitar el desarrollo de sus complicaciones crónicas.

Resulta importante reconocer que las consecuencias negativas de esta enfermedad comienzan en una etapa previa conocida como prediabetes (PDM), la cual está definida como una elevación en la concentración de glucosa en sangre más allá de los niveles normales pero sin alcanzar los valores diagnósticos de diabetes. La PDM se manifiesta a través de la Glucemia en Ayunas Alterada (GAA), la Tolerancia a la Glucemia

Alterada (TGA) y la combinación de ambas [1]. La PDM implica un riesgo elevado de desarrollar DT2 del orden del 30% [7] y de 70% [9] en los siguientes 4 y 30 años, respectivamente.

El desarrollo de la DT2 es un proceso lento y progresivo condicionado por factores genéticos, ambientales y de comportamiento. Actualmente no existe una cura definitiva para esta enfermedad. Sin embargo, varios estudios han demostrado que se puede prevenir o demorar su aparición en personas con PDM a través de la adopción de un estilo de vida saludable (plan de alimentación y práctica regular de actividad física) y/o asociado con la ingesta de diversos fármacos [7,26]. En Argentina, una iniciativa de este tipo es el Programa Piloto para la Prevención Primaria de Diabetes en la provincia de Buenos Aires (PPDBA) desarrollado por el CENEXA (UNLPCONICET) y financiado por el Ministerio de Ciencia y Tecnología de la Nación, la empresa SANOFI y el CONICET [12].

La detección de DT2 y PDM representa un verdadero desafío para la medicina debido a la ausencia de síntomas patogenómicos y a la falta de conocimiento de los factores de riesgo asociados. Es por eso que frecuentemente una persona pueda pasar meses (o incluso años) sin saber que se encuentra en riesgo. En ese sentido, estadísticas publicadas en el año 2018 por la Federación Internacional de Diabetes muestran que aproximadamente un 50% de la población mundial desconoce su enfermedad [10]. Esto explica la necesidad de contar con un método de detección simple y preciso. En consecuencia, este artículo propone desarrollar y evaluar modelos predictivos basados en aprendizaje automático (AA) que permitan identificar personas con riesgo de diabetes y PDM en la población argentina considerando como base de datos la correspondiente al PPDBA.

El resto del artículo se organiza de la siguiente forma. La Sección 2 introduce el marco referencial para este trabajo. Luego, la Sección 3 describe el procesamiento realizado a la base de datos mientras que la Sección 4 describe y analiza los resultados obtenidos. Finalmente, la Sección 5 presenta las conclusiones junto al trabajo futuro.

2. Marco Referencial

2.1 Factores de Riesgo y Diagnóstico de Diabetes y Prediabetes

El desarrollo de DT2 es un proceso lento y progresivo que se encuentra condicionado por factores genéticos, ambientales y de comportamiento. Entre los factores de riesgo se encuentran, entre otros, el género, el índice de masa corporal (IMC), la circunferencia de cintura, los hábitos de alimentación, la práctica de actividad física, la edad, los antecedentes familiares de diabetes (incluida gestacional), la etnia y los trastornos del sueño. Por su parte, la PDM representa un estado previo a la DT2 y su progresión puede prevenirse e incluso revertirse mediante la adopción de estilos de vida saludables [27]. Si además consideramos que la PDM no es una pre-enfermedad, pues ya presenta disfunciones metabólicas, resulta sumamente importante identificar personas con PDM tanto como lo es poder hacerlo con las que ya tienen DT2 no diagnosticada.

El diagnóstico se realiza por medio de análisis de sangre y cualquier persona con presencia de los síntomas o factores de riesgo asociados debe ser examinada. En Argentina, habitualmente se emplea la Prueba de Tolerancia Oral a la Glucosa (PTOG) para determinar si una persona posee DT2, PDM o ninguna de ellas [1].

2.2 Trabajos relacionados

En la última década, numerosos modelos han sido propuestos para identificar diabetes no diagnosticada y/o PDM utilizando técnicas de AA. Estas propuestas utilizan diversas variables clínicas y de laboratorio asociadas a los factores de riesgo de la enfermedad para establecer la predicción. La mayoría de ellas [15] [2] [13] [16] [19] [21] [16] [33] [6] [28] [22] [24] [18] [34] [30] [25] emplean una base de datos de diabetes conocida como PIMA Indian Diabetes (PID) del repositorio de la Universidad de California Irvine [2], EEUU. Este *dataset* contiene registros de mujeres del pueblo indígena Pima de EEUU y se compone de 8 atributos relacionados con los factores de riesgo de desarrollar la enfermedad. Dispone de un total de 786 registros; sin embargo, el número se reduce a la mitad al eliminar aquellos que poseen valores nulos. Se puede decir que estos trabajos son una *prueba de concepto* más que una implementación real y que, en general, se han orientado a mejorar el rendimiento (*accuracy*) de los algoritmos de clasificación de diabetes.

Son pocos los trabajos que no emplean la base de datos PID. Entre ellos se encuentran [14] [32] [8] quienes propusieron modelos para identificar diabetes y prediabetes usando datos de la Encuesta NHANES de EEUU. Recientemente, [31] también estudió la aplicación de distintas técnicas de AA para la detección de diabetes no diagnosticada en la población estadounidense aunque empleando la base de datos BRFSS. En forma similar, [20] presentó diferentes modelos predictivos de estas enfermedades para la población china usando una base de datos *ad-hoc*. A diferencia de los anteriores, [5] hizo hincapié en el desarrollo y validación de modelos predictivos únicamente para prediabetes.

Como se explicó en la Sección 2.1, el desarrollo de DT2 se encuentra condicionado por factores que pueden variar de una población a otra. Es por lo que un modelo predictivo que resulte adecuado para una población, no necesariamente lo será para otra. Este estudio representa el primer paso hacia modelos predictivos específicos para la población argentina.

3. Implementación

3.1 Conjunto de datos

[1] Esta prueba comienza con una extracción de sangre de la persona en ayunas. Luego, se le pedirá que tome un líquido que contiene una cierta cantidad de glucosa. A continuación, se le tomarán muestras de sangre nuevamente cada 30 a 60 minutos después de ingerir la solución. El examen es costoso económicamente y puede demorar hasta 3 horas.

[2] https://archive.ics.uci.edu/ml/index.php

Descripción Los modelos predictivos serán desarrollados a partir de la base de datos del programa PPDBA [12], la cual cuenta con 1316 registros de personas. Cada registro corresponde a una persona que mediante PTOG fue identificada como diabética, prediabética o sin ninguna de ellas. Además de datos de laboratorio (hemoglobina glicosilada; colesterol total, HDL y LDL; triglicéridos y creatinina), se cuenta con variables clínicas asociadas a los factores de riesgo de estas enfermedades tales como el sexo; la edad; el Indice de Masa Corporal (IMC); la presión arterial; los antecedentes familiares de diabetes; los hábitos alimenticios y de actividad física; entre otros.

Caracterización De los 1316 registros actuales, 80 debieron ser descartados ya que omitían valores requeridos de glucemia para poder calcular el resultado de la PTOG (no es posible determinar la clase). La Tabla 1 presenta una breve descripción estadística de los 1236 registros disponibles. Se puede notar que hay varias variables que presentan nulos, los cuales se analizan con mayor profundidad en la sección siguiente. También se puede observar que: hay más personas del sexo femenino que del masculino; la mayoría de las personas tienen entre 45-64 años; la mayoría de las personas tienen un IMC mayor a 30 kg/m^2; la mayoría de las personas tienen una circunferencia de cintura de más de 102cm y de 88cm para el sexo masculino y femenino, respectivamente; la mayoría de las personas realizan actividad física; consumen vegetales, frutas y hortalizas; no toman medicación para controlar hipertensión; sí le encontraron hiperglucemia en algún control; y alguno de sus familiares (de primer o segundo grado) tiene diabetes; las variables asociadas a glucemia basal, colesterol HDL, triglicéridos y creatinina basal parecieran tener una amplia dispersión; el resto no; en cuanto a la clase, la mitad de las personas no están en riesgo de tener prediabetes o diabetes.

Limpieza

Valores con ruido Para analizar la presencia de ruido de las variables, se utilizó el método de Tukey para identificar los intervalos de valores atípicos leves y extremos. En base a lo anterior, se detectaron valores atípicos leves en las variables edad (4), imc (10), circ___de_cintura (1), glucemia_basal (27), glucemia_pprandial (46), colesterol_ldl (7), colesterol_total (9), colesterol_hdl (11), triglicéridos (21) y creatinina_basal (8). Además, se encontraron valores atípicos extremos en glucemia_basal (51), glucemia_pprandial (6), trigliceridos (12) y creat_basal (6).

Valores nulos En la Tabla 1 se puede observar que no hay valores nulos en las variables cualitativas, a excepción de le_diag___familiar, que presenta 3 registros con valores nulos. No ocurre lo mismo con las variables cuantitativas, donde el faltante de valores es mucho mayor. Afortunadamente, para algunas de esas variables cuantitativas, sí se cuenta con una variable cualitativa asociada que permite conocer el rango en que se encuentra ese valor faltante. Esto ocurre concretamente en los casos de edad, imc y circ__de_cintura. El resto de las variables que presentan nulos son: glucemia_pprandial, colesterol_total, colesterol_ldl, colesterol_hdl, trigliceridos, creat_basal y hem___glicosilada.

Transformaciones

Tabla 1. Breve descripción estadística del dataset

Variable	# nulos	Medida	# casos
sexo	0	Masculino (%)	395 (32%)
		Femenino (%)	841 (68%)
edad	564	Media+DE	57.23±8.8428
rango_edad	0	Menos de 45 años (%)	205 (17%)
		45-54 años (%)	435 (35%)
		54-64 años (%)	429 (34%)
		Mayor de 64 años (%)	167 (14%)
imc	619	Media+DE	31.65±6.314
rango_imc	0	Menor de 25kg/m2	93 (8%)
		25-30 kg/m2	319 (26%)
		Mayor de 30 kg/m2	824 (66%9
circ___cintura	1181	Media+DE	101.3091±13.55
rango___cintura	0	Menos de 94/80 cm (M/F) (%)	56 (4%)
		M: 94-102cm / F: 80-88cm (%)	205 (17%)
		M: Más de 102/88cm (M/F) (%)	975 (79%)
actividad_fisica	0	Sí (%)	915 (74%)
		No (%)	321 (26%)
cons___hortalizas	0	Sí (%)	821 (66%)
		No (%)	415 (34%)
toma___hta	0	Sí (%)	497 (40%)
		No (%)	739 (60%)
le___hiperglucemia	0	Sí (%)	999 (81%)
		No (%)	237 (19%)
le_diag___familiar	3	No (%)	395 (32%)
		Primer grado (%)	412 (33%)
		Segundo grado (%)	426 (34%)
glucemia_basal	0	Media+DE	104.36±27.28
glucemia_pprandial	55	Media+DE	119,59±42,51
colesterol_total	705	Media+DE	198,28±41,1
colesterol_ldl	715	Media+DE	119,79±36,82
colesterol_hdl	706	Media+DE	49,82±14,40
trigliceridos	705	Media+DE	151,4±95,59
creatinina_basal	619	Media+DE	1,117±5,8
hem___glucosilada	635	Media+DE	5.61±0,43
resultado_ptog	0	Sin riesgo (%)	620 (50%)
		Prediabetes (%)	480 (38%)
		Diabetes (%)	136 (12%)

Tratamiento de valores atípicos Se consultó con expertos del dominio médico sobre las ocurrencias de valores atípicos en las determinaciones de laboratorio. Se concluyó que, si bien son valores que estadísticamente pueden ser considerados atípicos, sí se

encuentran dentro de los posibles valores extremos para estas determinaciones. La excepción son 2 valores en creatinina_basal (’85’, ’118’) que efectivamente corresponden a (posibles) errores de carga. Por lo tanto, esos dos valores particulares se reemplazaron con nulos.

Tratamiento de valores nulos Se debe optar entre: (1) eliminar las variables que tienen nulos; o (2) eliminar los registros que tienen variables con valores nulos. La opción 1 permite mantener el tamaño muestral a costa de reducir la cantidad de variables de entrada para los modelos. En sentido opuesto, la opción 2 permite mantener la cantidad de características a costo de reducir el tamaño muestral. En este caso, se decidió seleccionar la opción 2 únicamente para la variable le_diag___familiar, que cuenta con sólo 3 valores nulos. Por otro lado, la opción 1 se aplicó a las variables edad, imc y circ_de_cintura, considerando que hay una variable cualitativa asociada que permite conocer el rango de cada valor. Para el resto de las variables, se discutirá en la Sección 3.2.

Agrupamiento de variable de clase De la Tabla 1 se puede notar que la distribución de clases no está balanceada (variable resultado_ptog). Para minimizar el impacto de esta cuestión, se procede a crear una nueva variable de clase que divida entre personas sin riesgo de tener PDM o DM (registros con valor “Normal”) y las que sí lo tienen (registros con valor “PDM” o “DM”). Como resultado, la variable queda balanceada en cuanto a ocurrencias de cada valor y, a la vez, se simplifica el análisis posterior al pasar a ser ahora un problema de clasificación binaria. Sin embargo, se deberá ser cuidadoso al momento de analizar los resultados, especialmente con lo que se pueda decir sobre predicción de DM, por ser la de menor ocurrencia.

Eliminación de variables resultado_ptog fue descartada por el agrupamiento realizado en la sección anterior, mientras que glucemia_pprandial fue excluida, ya que contar con ese valor implicaría que la persona tuviera que hacerse una PTOG, careciendo de sentido el uso de los modelos propuestos.

Análisis de correlaciones La Fig. 1 muestra la matriz de correlación obtenida sobre el *dataset* inicial disponible. Desde el punto de vista clínico, tiene sentido que exista una correlación lineal débil entre el rango de la circunferencia de cintura y el rango IMC. También resulta razonable que existan correlaciones débiles entre los rangos de edad, IMC y circunferencia de cintura y sus valores asociados en las variables de edad, imc y circ___de_cintura. Por otra parte, el colesterol_total se calcula a partir del colesterol_ldl [3], lo que explica su correlación fuerte. Adicionalmente, la relación entre glucemia_basal, glucemia_pprandial, resultado_ptog y clase tiene sentido, ya que las dos primeras determinan el valor de la tercera, la cual a su vez se agrupa para generar la cuarta de ellas.

3.2 Segmentaciones propuestas

Ante el porcentaje alto de nulidad de las variables de laboratorio, se planteó la posibilidad de generar varios datasets a partir del original, considerando distintos criterios para el tratamiento de registros nulos:

- Dataset Clínica+Laboratorio (DCL-bin). Conjunto de datos al cual se realizó una eliminación de registros completos, derivando en 16 variables con 503 ejemplos (229 ⇒ Sin riesgo, 274 ⇒ Con riesgo). Este dataset mantiene todos las variables

disponible (datos clínicos y de laboratorio) a costa de perder cantidad de registros.

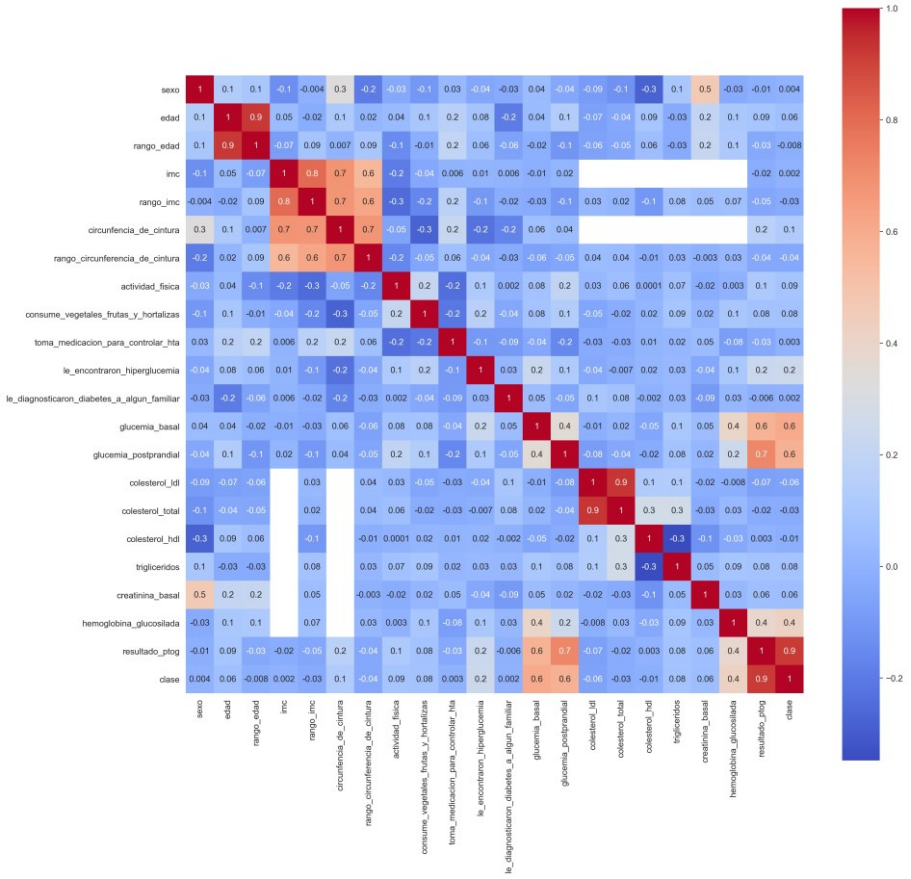

Figura1. Matriz de correlación sobre dataset inicial

- Dataset Clínica+Glucemia basal (DCG-bin). Conjunto de datos que mantiene la información clínica disponible y la única variable de laboratorio que no posee valores nulos (glucemia_basal); el resto de las ellas fueron eliminadas. De esta forma, este dataset cuenta con 10 variables y 1233 ejemplos. En contraposición con DCL aquí se preserva la cantidad de registros frente al valor aportado por el resto de los variables de laboratorio.

4. Resultados Experimentales

4.1 Diseño experimental

Se llevó a cabo un proceso de exploración, preprocesamiento, entrenamiento y evaluación de diversos algoritmos utilizando *scikit-learn*. En particular, se consideraron Logistic Regression (LR), Decision Tree (DT) con max_depth = 5 y class_weight = balanced, k-Nearest Neighbor (kNN) con n_neighbors = 7 y Random Forest (RF) con max_depth = 2. Asimismo, se utilizó TensorFlow para la construcción de modelos de Artificial Neural Networks (ANN) con epochs = 60 y batch_size = 16. En particular, se construyó una ANN de una capa oculta con 100 neuronas y función de activación ReLU, con una regularización L2 de 0,1 para controlar el sobreajuste. Se utilizó un optimizador Adamax con una tasa de aprendizaje de 0,001.

Siguiendo lineamientos habituales del área, el conjunto de datos fue particionado en dos: entrenamiento (70%) y evaluación (30%). Para reducir posibles sesgos, se aplicó la técnica de validación cruzada con muestreo estratificado aleatorio (*StratifiedShuffleSplit, n_splits=50* [23]). Debido a que las variables de entrada poseen magnitudes diferentes, se aplicó una normalización *min-max* para todos los modelos excepto para la ANN donde se utilizó una normalización estándar.

Todas las experimentaciones se realizaron sobre una computadora local utilizando *Jupyter Notebook* [17]. El hardware disponible fue una CPU Intel Core i7 2.6 GHz Quad-Core, 16 GB de memoria RAM y sistema operativo macOS.

4.2 Modelos y resultados para DCL-bin

Tabla 2. Resultados (evaluación) para los modelos aplicados a DCL-bin (clase positiva = "Con riesgo")

Modelo	Accuracy	Precision	Recall	F-score	AUC
RF	94.58 ± 1.51	98.87 ± 1.3	91.1 ± 2.93	94.79 ± 1.53	0.95
DT	93.42 ± 1.59	96.02 ± 2.02	91.73 ± 2.65	93.79 ± 1.53	0.94
ANN	91.13 ± 1.98	92.29 ± 2.68	91.39 ± 2.77	91.79 ± 1.84	0.91
LR	85.56 ± 3.21	90.5 ± 3.11	82.1 ± 4.81	86.01 ± 3.3	0.93
kNN	71.62 ± 3.82	75.53 ± 3.87	70.83 ± 5.75	72.97 ± 4	0.72

La Tabla 2 muestra las métricas de rendimiento para los diferentes modelos aplicados al dataset DCL-bin considerando como clase positiva a "Con riesgo". De los 5 modelos, se puede observar que hay 3 opciones que obtienen valores de *accuracy* superiores al 90%, lo que significa que aproximadamente 9 de cada 10 de los registros evaluados del total, fueron clasificados correctamente (los desvíos son menores al 2%). De estas 3 opciones, RF es el que obtiene la mejor *accuracy*, seguido de DT y finalmente ANN. Mientras que el rendimiento de LR se encuentra cercano a los anteriores con 86% de *accuracy*, el correspondiente a kNN es pobre al alcanzar sólo 72%.

En cuanto a *precision*, los valores se condicen con los de *accuracy*. RF, DT y ANN presentan valores superiores al 90% para la clase de interés ("Con riesgo") lo que significa que, con cualquiera de las opciones, a más de 9 personas a las que se les dice que tiene riesgo, realmente lo tiene. En el mismo sentido, los 3 modelos mencionados valores de *recall* cercanos al 91%, lo que significa que sólo 1 de cada 10 personas que tiene riesgo, no es identificada por ellos.

Por último, al momento de seleccionar un modelo determinado, puede ser interesante examinar los valores de AUC además de la *accuracy*. En este caso, se puede notar que RL desplaza a ANN como uno de los 3 que obtiene los valores más altos de AUC, en comparación a los de mejor *accuracy*.

4.3 Modelos y resultados para DCG-Bin

Tabla 3. Resultados (evaluación) para los modelos aplicados a DCG-bin (clase positiva = "Con riesgo")

Modelo	Accuracy	Precision	Recall	F-score	AUC
RF	93.23 ± 1.12	100 ± 0	86.38 ± 2.25	92.68 ± 1.3	0.93
DT	91.88 ± 1.67	96.68 ± 2.74	86.72 ± 2.2	91.39 ± 1.73	0.92
ANN	91.04 ± 1.63	94.45 ± 2.47	87.17 ± 2.31	90.64 ± 1.71	0.91
LR	75.66 ± 4.37	80.96 ± 4.44	66.7 ± 6.55	73.05 ± 5.3	0.83
kNN	69.97 ± 2.01	70.63 ± 2.32	67.92 ± 3.29	69.2 ± 2.23	0.7

La Tabla 3 muestra las métricas de rendimiento para los diferentes modelos aplicados al dataset DCG-bin considerando como clase positiva a "Con riesgo". Al igual que con DCL-bin, RF, DT y ANN obtienen muy buenos valores de *accuracy*, siendo superiores al 90% (los desvíos son menores al 2%). kNN vuelve a presentar un rendimiento pobre con 70% de *accuracy*, mientras que LR se encuentra entre los 3 anteriores y kNN, aunque con un rendimiento más bajo que con el dataset anterior. En cuanto a *precision*, RF sobreajusta al obtener 100% para la clase "Con riesgo". Por su parte, DT y ANN presentan valores superiores al 90% para la misma clase, estando LR y kNN bastante debajo de ese valor. Al analizar los valores de *recall*, tanto RF como DT y ANN son los que consiguen los mejores valores, estando cercanos al 90% para la clase de interés.

Por último, y a diferencia de DCL-bin, los 3 que obtienen los valores más altos de AUC coinciden con los de mejor *accuracy*.

4.4 Discusión

Algunos de los modelos desarrollados obtuvieron muy buenos rendimientos para ambos datasets. En particular, RF, DT y ANN demostraron gran poder de clasificación, con altos valores en las métricas consideradas. En ese sentido, resulta importante aclarar que los modelos propuestos no pretenden reemplazar a las PTOGs como mecanismo de diagnóstico de DT2 y PDM. Al ser enfermedades de difícil detección precoz, estos modelos buscan identificar aquellas personas de la población argentina que tengan alta probabilidad de tenerlas y desconozcan su condición. Para confirmar el diagnóstico, las personas identificadas deberán realizar eventualmente una PTOG. Los modelos ayudarían a identificar a quienes deben realizarlo y suplirían la ausencia de herramientas de este tipo.

Se puede notar que no hay diferencias significativas entre los mejores valores de *accuracy* y *F-score* conseguidos para ambos datasets. Aunque no es (del todo) correcto comparar resultados de modelos entrenados con datasets diferentes, esta cuestión

podría tener incidencia en el costo de llevar a la práctica los modelos, considerando que conseguir las variables de laboratorio no es gratuito ni sencillo. Para poder dilucidarla, sería necesario contar con un mayor número de registros sin nulos.

Por último, una cuestión importante que debe tenerse en cuenta es el agrupamiento de PDM y DM como clase única, lo cual favoreció al balanceo y simplificó el problema al volverlo de clasificación binaria. El costo es justamente no poder diferenciar entre los casos de PDM y DM. Sin embargo, desde un punto de vista médico, esto no sería tan grave, ya que a fin de cuentas lo que interesa es identificar quienes están en riesgo (no importa de cuál).

5. Conclusiones y Trabajo Futuro

Considerando que DM y PDM son enfermedades de difícil detección, en este trabajo se desarrollaron y evaluaron modelos predictivos específicos para la población argentina a partir de la base de datos del PPDBA. En primer lugar, fue necesario realizar un cuidadoso preprocesamiento de la base de datos, lo que derivó en la generación de dos datasets particulares (DCL-bin y DCG-bin) considerando el compromiso entre cantidad de variables y de registros disponibles. Luego, se aplicaron 5 modelos de clasificación diferentes a cada uno de ellos. Los resultados obtenidos muestran que algunos de los modelos propuestos obtuvieron muy buenos rendimientos para ambos datasets. En particular, RF, DT y ANN demostraron gran poder de clasificación, con altos valores en las métricas consideradas. Debido a limitaciones propias de la base de datos, no es posible afirmar que los resultados sean concluyentes, aunque sí resultan promisorios. Considerando la vacancia de herramientas de este tipo para la población argentina, este trabajo representa el primer paso hacia modelos más sofisticados.

Entre las líneas de trabajo futuro se encuentran:

- Conseguir más registros de la base de datos para mejorar su calidad y al mismo tiempo aumentar su representatividad, para luego replicar el estudio realizado.
- Evaluar el rendimiento de modelos generados a partir de una nueva segmentación que sólo considere datos clínicos. Un modelo de estas características sería más sencillo, sin costo y factible de realizar en cualquier momento, aunque probablemente de menor rendimiento.
- Considerar el desarrollo de modelos de clasificación multiclase para separar los casos de DM y PDM.

Financiamiento. Este estudio fue parcialmente respaldado por PICT-2020-SERIE-A00901.

Referencias

1. Professional practice committee: Standards of medical care in diabetes—2021. Diabetes Care 44 (Supplement 1), S3–S3 (2021). https://doi.org/10.2337/dc21-Sppc, https://care.diabetesjournals.org/content/44/Supplement1/S3

2. Al Jarullah, A.A.: Decision tree discovery for the diagnosis of type ii diabetes. In: 2011 International Conference on Innovations in Information Technology. pp. 303–307 (April 2011). https://doi.org/10.1109/INNOVATIONS.2011.5893838
3. Association, A.H.: What Your Cholesterol Levels Mean. https://www.heart.org/en/healthtopics/cholesterol/about-cholesterol/what-your-cholesterol-levels-mean (2020), accedido: 2022-10-10
4. Bolin, K., Gip, C., Mörk, A.C., Lindgren, B.: Diabetes, healthcare cost and loss of productivity in sweden 1987 and 2005 - a register-based approach. Diabetic medicine : a journal of the British Diabetic Association 26, 928–34 (10 2009). https://doi.org/10.1111/j.14645491.2009.02786.x
5. Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.h., Kang, E.S., Kim, D.W.: Screening for prediabetes using machine learning models. Computational and mathematical methods in medicine 2014 (2014)
6. Dey, S.K., Hossain, A., Rahman, M.M.: Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm. In: 2018 21st International Conference of Computer and Information Technology (ICCIT). pp. 1–5 (2018). https://doi.org/10.1109/ICCITECHN.2018.8631968
7. Diabetes Prevention Program Research Group: Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine 346(6), 393–403 (2002). https://doi.org/10.1056/NEJMoa012512
8. Dinh, A., Miertschin, S., Young, A., Mohanty, S.D.: A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making 19(1), 211 (Nov 2019). https://doi.org/10.1186/s12911-019-0918-5
9. Eddy, D.M., Schlessinger, L., Kahn, R.: Clinical Outcomes and Cost-Effectiveness of Strategies for Managing People at High Risk for Diabetes. Annals of Internal Medicine 143(4), 251–264 (08 2005). https://doi.org/10.7326/0003-4819-143-4-200508160-00006
10. Federación Internacional de Diabetes: Atlas de la diabetes de la fid. Tech. rep., Federación Internacional de Diabetes (2017), https://diabetesatlas.org/resources/2017-atlas.html
11. Gagliardino, J.J., Assad, D., Gagliardino, G.G., Kronsbein, P., Lahera, E., Mercuri, N., Rizzuti, L., Zufriategui, Z.: Cómo tratar mi diabetes. Buenos Aires, Argentina, 3 edn. (11 2016)
12. Gagliardino, J.J., Etchegoyen, G., Bourgeois, M., Fantuzzi, G., García, S., González, L., Elgart, J.F., Ré, M., Ricart, A., Ricart, J.P., Spinedi, E.: Prevención primaria de diabetes tipo 2 en argentina: estudio piloto en la provincia de buenos aires. Revista Argentina de Endocrinología y Metabolismo 53(4), 135 – 141 (2016). https://doi.org/https://doi.org/10.1016/j.raem.2016.11.002
13. Hashi, E.K., Zaman, M.S.U., Hasan, M.R.: An expert clinical decision support system to predict disease using classification techniques. In: 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE). pp. 396–400 (Feb 2017). https://doi.org/10.1109/ECACE.2017.7912937
14. Heikes, K.E., Eddy, D.M., Arondekar, B., Schlessinger, L.: Diabetes risk calculator. Diabetes Care 31(5), 1040–1045 (2008). https://doi.org/10.2337/dc07-1150
15. Ilango, B.S., Ramaraj, N.: A hybrid prediction model with f-score feature selection for type ii diabetes databases. In: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. pp. 13:1–13:4. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1858378.1858391
16. Jayanthi, N., Babu, B.V., Rao, N.S.: Survey on clinical prediction models for diabetes prediction. Journal of Big Data 4(1), 26 (2017). https://doi.org/10.1186/s40537-017-00827
17. Jupyter Notebook Documentation: 7.0.0rc2 documentation. https://jupyternotebook.readthedocs.io/en/latest/index.html, accedido: 2023-04-18

18. Kaur, H., Kumari, V.: Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics (2018). https://doi.org/https://doi.org/10.1016/j.aci.2018.12.004
19. Maniruzzaman, M., Kumar, N., Menhazul Abedin, M., Shaykhul Islam, M., Suri, H.S., El-Baz, A.S., Suri, J.S.: Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. CMPB 152, 23–34 (2017). https://doi.org/https://doi.org/10.1016/j.cmpb.2017.09.004
20. Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q., Liu, Q.: Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences 29(2), 93 – 99 (2013). https://doi.org/https://doi.org/10.1016/j.kjms.2012.08.016
21. Mercaldo, F., Nardone, V., Santone, A.: Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science 112, 2519–2528 (2017). https://doi.org/https://doi.org/10.1016/j.procs.2017.08.193
22. Mir, A., Dhage, S.N.: Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). pp. 1–6 (2018). https://doi.org/10.1109/ICCUBEA.2018.8697439
23. Scikit-Learn: 3.1. Cross-validation: evaluating estimator performance. https://scikitlearn.org/stable/modules/crossvalidation.html,accedido : 2023 − 04 − 11
24. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. Procedia Computer Science 132, 1578 – 1585 (2018). https://doi.org/https://doi.org/10.1016/j.procs.2018.05.122
25. Sneha, N., Gangil, T.: Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data 6(1), 13 (2019). https://doi.org/10.1186/s40537019-0175-6
26. Tuomilehto, J., Lindström, J., Eriksson, J.G., Valle, T.T., Hämäläinen, H., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Laakso, M., Louheranta, A., Rastas, M., Salminen, V., Aunola, S., Cepaitis, Z., Moltchanov, V., Hakumäki, M., Mannelin, M., Martikkala, V., Sundvall, J., Uusitupa, M.: Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. New England Journal of Medicine 344(18), 1343–1350 (2001). https://doi.org/10.1056/NEJM200105033441801
27. Vistisen, D., Kivimäki, M., Perreault, L., Hulman, A., Witte, D.R., Brunner, E.J., Tabák, A., Jørgensen, M.E., Færch, K.: Reversion from prediabetes to normoglycaemia and risk of cardiovascular disease and mortality: the whitehall ii cohort study. Diabetologia 62(8), 1385–1390 (Aug 2019). https://doi.org/10.1007/s00125-019-4895-0
28. Wei, S., Zhao, X., Miao, C.: A comprehensive exploration to the machine learning techniques for diabetes identification. In: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT). pp. 291–295 (2018). https://doi.org/10.1109/WF-IoT.2018.8355130
29. Williams, R., Van Gaal, L., Lucioni, C.: Assessing the impact of complications on the costs of type ii diabetes. Diabetologia 45(1), S13–S17 (Jul 2002). https://doi.org/10.1007/s00125-002-0859-9
30. Wu, H., Yang, S., Huang, Z., He, J., Wang, X.: Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked 10, 100 – 107 (2018). https://doi.org/https://doi.org/10.1016/j.imu.2017.12.006
31. Xie, Z., Nikolayeva, O., Luo, J., Li, D.: Building risk prediction models for type 2 diabetes using machine learning techniques. Preventing chronic disease 16, E130–E130 (Sep 2019). https://doi.org/10.5888/pcd16.190109
32. Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10(1), 16 (2010). https://doi.org/10.1186/14726947-10-16

33. Yuvaraj, N., SriPreethaa, K.R.: Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster. Cluster Computing 22(1), 1–9 (Jan 2019). https://doi.org/10.1007/s10586-017-1532-x
34. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics 9, 515–515 (Nov 2018). https://doi.org/10.3389/fgene.2018.00515