CLASIFICACIÓN DE PACIENTES SEGÚN SU POSIBILIDAD DE ADQUIRIR DIABETES MELLITUS EMPLEANDO ALGORITMOS DE MACHINE LEARNING

Jorge Iván Pincay-Ponce
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Universidad Nacional de la Plata, 1900, La Plata, Argentina.
Email: jorge.pincay@uleam.edu.ec

Diana Alexandra Sánchez-Andrade
Universidad Internacional de La Rioja (UNIR), La Rioja, España.
Universidad de Guayaquil, EC090101 - EC090158, Guayaquil, Ecuador.
Email: diana.sancheza@ug.edu.ec

Ingrid Vanessa Caicedo-Ávila
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Email: vanec027a@gmail.com

David Gabriel Macías-Valencia
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Email: david.macias@uleam.edu.ec

RESUMEN

El presente trabajo tiene como objetivo rediseñar el análisis del conjunto de datos “Pima Indians Diabetes Database”, que fue parte de la investigación titulada “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus” presentada por J. Smith, J. Everhart, W. Dickson, W. Knowler y R. Johannes, en el Proceedings of the Annual Symposium on Computer Application in Medical Care de 1988. Hoy en día el conjunto de datos utilizado por los investigadores, aplicando modelos de detección temprana con redes neuronales, está disponible desde diversos sitios web. El rediseño del análisis original consistió en generar nuevos atributos categóricos a partir de los ocho atributos numéricos de los 768 registros originales, una vez que el modelo se entrenó usando 576 registros, se utilizó Árboles de Decisión y Reglas de asociación, para pronosticar si en otros 192 registros de prueba, las pacientes desarrollarían diabetes mellitus. Los resultados basados en atributos categóricos, como: número de embarazos, prueba de tolerancia a la glucosa, presión arterial diastólica, espesor del pliegue cutáneo de los tríceps, insulina sérica, índice de masa corporal y un indicador de diabetes en familiares cercanos, son comprensibles por un mayor número de personas no especialistas en diabetes. En el escenario expuesto, la probabilidad de predicción correcta con Árboles de Decisión fue del 65% y las 20 primeras Reglas de Asociación generadas van desde el 70% al 90% de confianza como clasificadoras para la predicción de diabetes. La herramienta utilizada en la elaboración del modelo fue Azure Machine Learning Studio (Classic).

Palabras clave: Diabetes Mellitus, Boosted Decision Tree, Azure Machine Learning, Data mining.

Introducción

La diabetes es la enfermedad metabólica más común, por tanto, se ha constituido en un gran desafío mundial, siendo además, la principal causa de muerte en el mundo, e incluso la Federación Internacional de Diabetes estima que en el 2040, 642 millones de personas serán diabéticas (Madmoli et al., 2019). A la fecha, muchos estudios relacionan a los antecedentes familiares con el estado de presencia de diabetes en una persona. (Neel, 1976; Ramesh et al., 2017; Smith et al., 1988).

Por años, diversas organizaciones han recopilado sistemáticamente múltiples conjuntos de datos sobre pacientes diabéticos y muchos de ellos están disponibles desde la web (Breault et al., 2002), en tal sentido, en el presente trabajo se ha obtenido y rediseñado el análisis de un conjunto de datos referido a 768 mujeres de entre 21 y 81 años de la población india Pima, que se localiza cerca de Phoenix, Arizona. Esta población ha sido catalogada en décadas recientes como con fuerte predisposición genética a la Diabetes Mellitus Tipo 2. (Fufaa et al., 2015; Knowler et al., 1991; Olsson & Goedecke, 2020; Pettitt & Knowler, 1988; Savage et al., 1979; Tulloch-Reid et al., 2003).

El mencionado conjunto de datos fue compilado para la investigación titulada “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus” presentada por J. Smith, J. Everhart, W. Dickson, W. Knowler y R. Johannes, en el Proceedings of the Annual Symposium on Computer Application de 1988. Los datos originales consisten en ocho variables predictoras médicas independientes de tipo numérico, en adelante referidas como características, y una variable objetivo dependiente de tipo nominal o texto. Estas características representan factores de riesgo significativos de diabetes para las mujeres Pima. (Smith et al., 1988).

Respecto de los algoritmos escogidos, las reglas de asociación, según (Pincay-Ponce et al., 2020; Wu et al., 2008), han sido incluidas por la IEEE International Conference on Data Mining, entre los diez primeros algoritmos de minería de datos más influyentes en la comunidad de investigación, además, de que, como indican otros autores, se han utilizado en estudios recientes para la clasificación de Diabetes Mellitus, junto con los Árboles de Decisión (Ahouz et al., 2019; Pekel & Özmen, 2020).

El algoritmo específico de árbol de decisión que se seleccionó fue el Two-Class Boosted Decision Tree implementado en Azure Machine Learning Studio, al que en adelante se referirá como Azure ML, para crear un modelo de aprendizaje automático que se basa en el algoritmo de árboles de decisión impulsado, que es un método de aprendizaje conjunto en el que el segundo árbol corrige los

errores del primer árbol, el tercer árbol corrige los errores del primer y segundo árbol, y así sucesivamente, para que en el conjunto de todos los árboles se haga la predicción. (Ríos Canales, 2016). Para calcular las reglas de asociación y el conjunto de elementos frecuentes se utilizó la integración del algoritmo Apriori en Azure ML, desde el paquete Arules disponible el repositorio Comprehensive R Archive Network (CRAN). (Barga et al., 2015).

El objetivo de este artículo es comparar el rendimiento de los clasificadores de árboles de decisión y reglas de asociación en la predicción de la Diabetes Mellitus, en función de su precisión y tasa de verdaderos positivos. La metodología aplicada consistió en analizar y generar con los clasificadores indicados, nuevas características categóricas a partir de las numéricas del conjunto de datos original, justificando cada discretización bibliográficamente. Se ha considerado que los resultados basados en datos categóricos son comprensibles por un mayor número de personas.

Materiales y métodos

Este estudio es de tipo aplicado con detalles descriptivos al recurrir a la discretización de las características originales con base en bibliografía existente. En general, los pasos seguidos fueron cuatro: **1)** Preparación de datos, **2)** Ajuste de los hiperparámetros y examinado de datos con los algoritmos Two-Class Boosted Decision Tree y Apriori, **3)** Cálculo y evaluación de la precisión de los algoritmos, **4)** Uso de servicios web de Azure ML para determinar si un paciente es propenso a tener diabetes en el futuro en función de los datos que se suministren al servicio web.

Para el **Paso 1**, en el Cuadro 1 se muestran las características del conjunto de datos original y sus equivalentes discretizados mediante la aplicación de transformaciones SQL en Azure ML (Mund, 2015, p. 63), junto con la estadística básica de cada característica en el conjunto de datos original. El conjunto de datos original no tiene nulos ni vacíos y al estar depurado por (Smith et al., 1988) la labor de análisis y pre procesado se simplificó considerablemente.

En la discretización, para la concentración de glucosa en plasma, la presión

arterial diastólica, el grosor del pliegue de la piel del tríceps, la insulina sérica de 2 horas y el índice de masa corporal; se establecieron según los rangos determinados en la investigación de (Bashir et al., 2019). La edad en años se la discretizó según las denominadas directrices provisionales sobre clasificaciones de edad internacionales estándar (Affairs, 1982). El número de embarazos se discretizó según las categorías definidas por (Smith et al., 1988), al igual que la medida de influencia genética, que utiliza información de padres, abuelos, hermanos completos y medios, tías y tíos completos y medios, y primos.

La medida de influencia genética documentada en el artículo original crece a mayor número de familiares que desarrollaron la diabetes, también cuando la edad a la que esos familiares desarrollan diabetes es menor y a medida en que el porcentaje de genes que comparten con la paciente aumenta. En contraparte disminuye al ser mayor el número de familiares que nunca desarrollaron diabetes o a medida que es mayor su edad al realizarse el último examen.

Luego, se identificó estadísticamente el poder predictivo de las características en relación con el atributo clase. Esto se hizo con filtros de Azure ML, usando la métrica de información mutua, porque soporta etiquetas y características de texto o numéricas. (Mund, 2015, p. 65), obteniendo los resultados del Cuadro 1. Para finalizar el paso 1, los nuevos datos se dividieron en conjuntos de entrenamiento y pruebas del 70% y 30%, que corresponden a 538 y 230 registros respectivamente.

class	fplas	fmass	fage	fpreg	finsu	fpedi	fpres	fskin
1	0.083771	0.057561	0.045362	0.036228	0.027369	0.01725	0.009376	0.005531

Cuadro 1: Poder predictivo de las características ordenados de mayor a menor según la métrica de Información Mutua.

En el cuadro anterior, también se interpreta como que el grosor del pliegue de la piel del tríceps dice menos con respecto a desarrollar diabetes y que la concentración de glucosa en plasma a 2 horas en una prueba oral de tolerancia

a la glucosa. El siguiente Diagrama de Venn ilustra las relaciones aditivas y sustractivas de varias medidas de información asociadas con las variables correlacionadas X y Y. El área contenida por ambos círculos es la entropía conjunta H (X, Y). El círculo de la izquierda (rojo y violeta) es la entropía individual H(X), siendo el rojo la entropía condicional H(X|Y). El círculo de la derecha (azul y violeta) es H(Y). El área violeta es la información mutua I (X; Y).(Pluim et al., 2003)

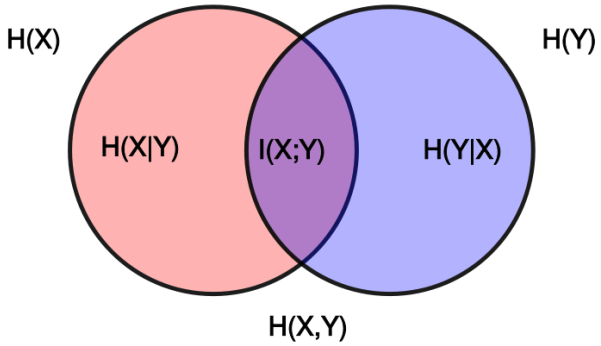


Ilustración 1: Diagrama de Venn para ilustración la métrica de información mutua

Variables	Estadística básica de datos numéricos	Significado de los rangos discretizados
Preg: Número de embarazos	Mínimo 0, Máximo 17 Promedio: 3.845 Des. Estándar: 3.3673 Valores distintos: 17 Valores Únicos: 2	0: Sin embarazo >=1 y <=2: Moderno =>3 y <=6: Post antiguo >=7: Antiguo *Valores Únicos: 4*
Plas: Concentración de glucosa en plasma a 2 horas en una prueba oral de tolerancia a la glucosa	Mínimo 0, Máximo 199 Promedio: 120.895 Des. Estándar: 31.973 Valores distintos: 136 Valores Únicos: 19	<74: Glucosa baja <=74 y <140: Glucosa normal >=140: Glucosa Alta *Valores Únicos: 3*
Pres: Presión arterial diastólica (mm Hg)	Mínimo 0, Máximo 122 Promedio: 69.105 Des. Estándar: 19.356 Valores distintos: 47 Valores Únicos: 8	<64: Hipotensión presión baja >=64 y <85: Presión arterial normal >= 85: Hipotensión presión alta *Valores Únicos: 3*
Skin: Grosor del pliegue de la piel del tríceps (mm)	Mínimo 0, Máximo 99 Promedio: 20.536 Des. Estándar: 15.952 Valores distintos: 51 Valores Únicos: 5	<20: Espesor bajo >=20 y <40: Espesor normal >= 40: Espesor muy alto *Valores Únicos: 3*
Insu: Insulina sérica de 2 horas (mu U / ml)	Mínimo 0, Máximo 846 Promedio: 79.799 Des. Estándar: 115.244 Valores distintos: 186 Valores Únicos: 93	<60: No diabético >=60 y <110: Diabético Etapa1 >=110 y <130: Diabético Etapa2 >=130: Diabético Etapa3 *Valores Únicos: 4*

Variables	Estadística básica de datos numéricos	Significado de los rangos discretizados
Mass; Índice de masa corporal (peso en kg / (altura en m) ^ 2)	Mínimo 0, Máximo 67.1 Promedio: 31.993 Des. Estándar: 7.884 Valores distintos: 248 Valores Únicos: 76	<18.5: Por debajo del peso >=18.5 y <25: Saludable <=25 y <30: Con sobrepeso >=30 y <40: Obeso >=40: Obesidad Extrema *Valores Únicos: 5*
Pedi: Esta medida de influencia genética.	Mínimo 0.708, Máximo 2.42 Promedio: 0.472 Des. Estándar: 0.331 Valores distintos: 517 Valores Únicos: 346	'0 - 0.244' ‘0.245 - 0.525' '0.526 - 0.825' '0.826 - 1.1' '>1.1' *Valores Únicos: 5* Fuente: (Smith et al., 1988)
Age: Edad en años	Mínimo 21, Máximo 81 Promedio: 33.241 Des. Estándar: 11.76 Valores distintos: 52 Valores Únicos: 5	<1: Infante >=1 y <15: Jóvenes >=15 y <25: Joven >=25 y <45: Edad adulta joven >=45 y <65: Edad adulta media >=65: Edad adulta mayor *Valores Únicos: 4*
Class: Atributo de clase (0 o 1)	Tipo: Texto/Clase tested_negative: 500 tested_positive: 268	

Cuadro 2: Variables numéricas originales y sus pares discretizadas

Para el **Paso 2,** los hiperparámetros del árbol de decisión se configuraron como: número de hojas 20, instancias mínimas por hojas 10, Tasa de aprendizaje 0.2, número de árboles de refuerzo 100. Los parámetros de Apriori se definieron como: soporte mínimo 0.1, confianza mínima 0.5, mínimo de elementos en una regla 2, máximo de elementos en una regla 8, orden por confianza, se aplicó poda para las redundancias y en cuanto a los conjuntos de elementos frecuentes se configuraron como 20 al igual que las reglas de asociación. Los **Pasos 3** y **4** corresponden a los resultados.

Resultados y discusión

El modelo, que en Azure ML se denomina experimento resultó tal cual se muestra la Ilustración 1. En la Ilustración 2 se muestra una regla bordeada de color azul para un caso de prueba negativo, que se extrajo de una vista parcial del centésimo árbol de decisión que se generó.

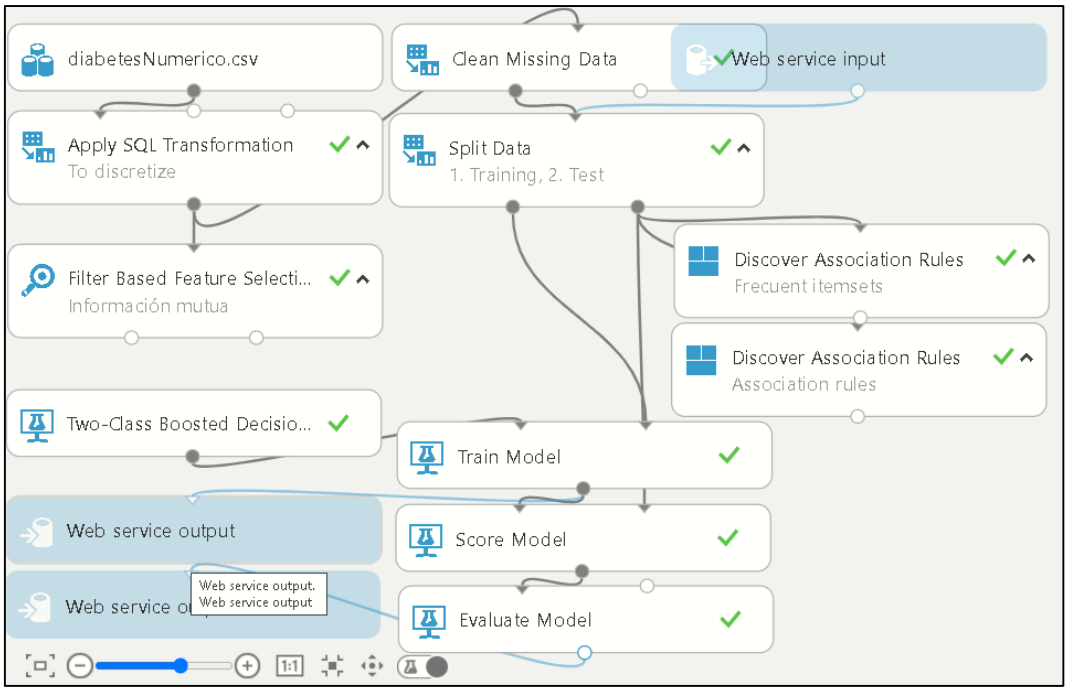


Ilustración 2: Experimento resultante en Azure ML

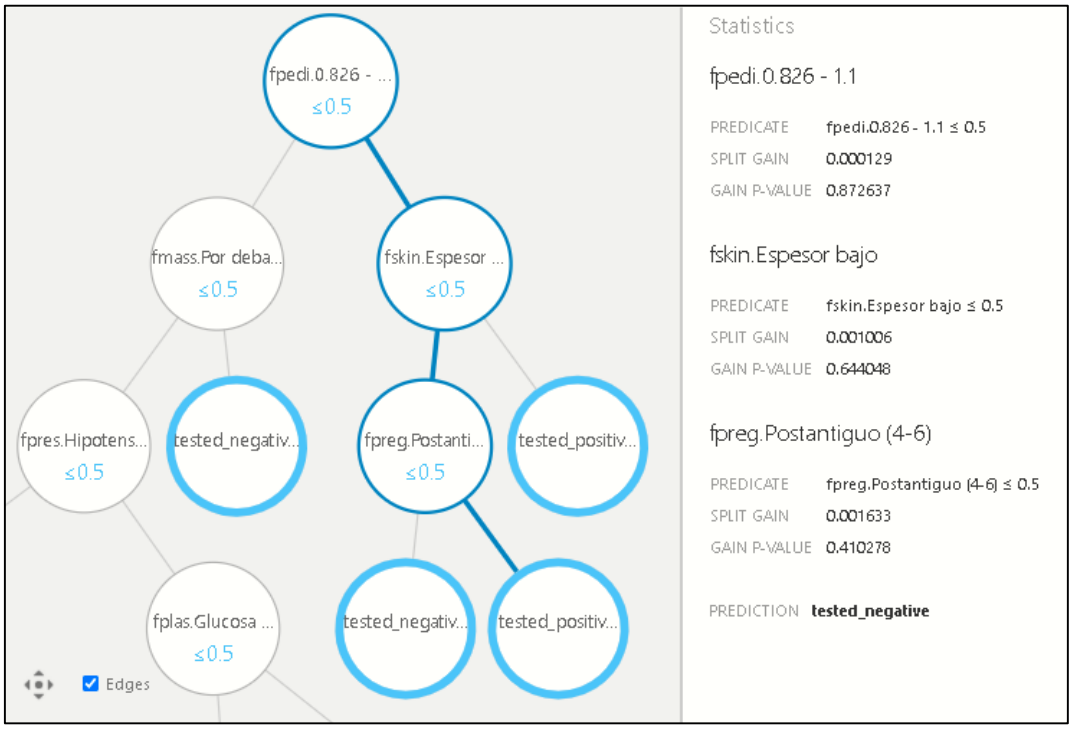


Ilustración 3: Ejemplo de regla con resultados negativos

A modo de ejemplo, en la ilustración 3, se muestran los 5 primeros itemsets más frecuentes, así como las 10 primeras reglas de asociación ordenadas por confianza, indicando para cada regla los antecedentes o left-hand-sides (lhs) y los consecuentes o right-hand-sides (rhs), así como el soporte y la confianza.

id	items	support
1	{fplas=Glucosa normal,class=tested_negative}	0.556522
2	{fplas=Glucosa normal,fpres=Presión arterial normal}	0.430435
3	{fpres=Presión arterial normal,class=tested_negative}	0.408696
4	{fplas=Glucosa normal,finsu=No diabético}	0.4
5	{fplas=Glucosa normal,fage=Edad adulta joven}	0.386957

id	lhs	rhs	support	confidence
1	{finsu=Diabético Etapa 1}	{fplas=Glucosa normal}	0.117391	0.9
2	{fmass=Saludable}	{class=tested_negative}	0.117391	0.9
3	{fmass=Saludable}	{fskin=Espesor bajo}	0.108696	0.833333
4	{fmass=Saludable}	{fplas=Glucosa normal}	0.108696	0.833333
5	{class=tested_negative}	{fplas=Glucosa normal}	0.556522	0.820513
6	{fskin=Espesor bajo}	{finsu=No diabético}	0.334783	0.819149
7	{fage=Edad adulta media}	{fpres=Presión arterial normal}	0.117391	0.818182
8	{fage=Jóven}	{class=tested_negative}	0.256522	0.808219
9	{finsu=Diabético Etapa 1}	{class=tested_negative}	0.104348	0.8
10	{fage=Jóven}	{fplas=Glucosa normal}	0.252174	0.794521

Ilustración 4: Itemsets más frecuentes y reglas ordenadas por confianza. La regla 9 indica que, si está en estado de diabético etapa de 1, entonces la prueba tiene un 80% de probabilidad de dar negativo.

fpreg	fplas	fpres	fskin	finsu	fmass	fpedi	fage	class	Scored Labels	Scored Probabilities
Sin embarazo	Glucosa Alta	Presión arterial normal	Espesor normal	No diabético	Obesidad Extrema	0.245 - 0.525	Edad adulta joven	tested_positive	tested_positive	0.999946
Sin embarazo	Glucosa normal	Presión arterial normal	Espesor bajo	No diabético	Obeso	0 - 0.244	Edad adulta joven	tested_negative	tested_negative	0.024617
Postantiguo (4-6)	Glucosa normal	Presión arterial normal	Espesor normal	No diabético	Obeso	0.526 - 0.825	Edad adulta media	tested_negative	tested_positive	0.999407
Moderno (1-3)	Glucosa normal	Presión arterial normal	Espesor bajo	No diabético	Saludable	0 - 0.244	Edad adulta joven	tested_negative	tested_negative	0.000001
Moderno (1-3)	Glucosa normal	Hipotensión presión baja	Espesor muy alto	Diabético Etapa 3	Obeso	0.526 - 0.825	Edad adulta joven	tested_negative	tested_negative	0.42482

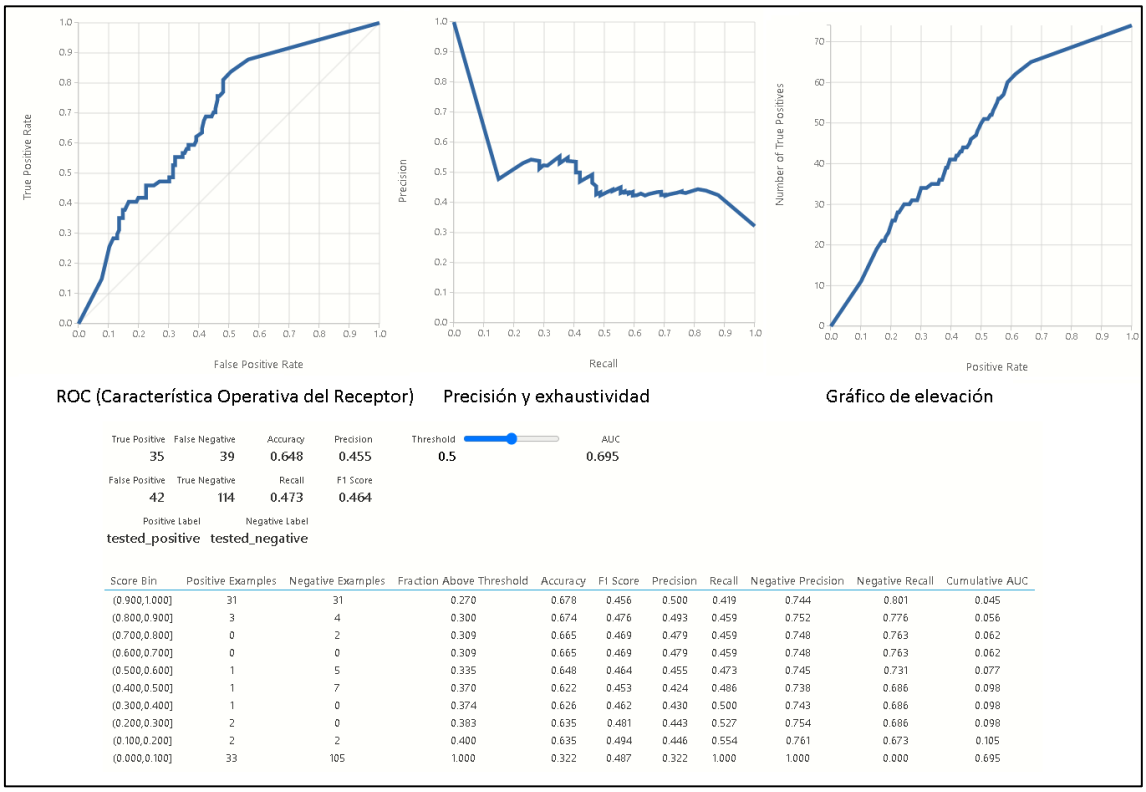
Ilustración 5: Algunas predicciones del árbol de decisión y su probabilidad de ocurrencia encerradas en los recuadros verdes.

En la ilustración 5 se muestran los resultados previstos para un umbral de datos del 50%, destacándose la precisión que alcanza el 64.8%, pero también muchos falsos negativos, lo que puede estar relacionado a la necesidad de aplicar análisis difuso en un problema de este tipo, donde se manejan rangos de valores dispersos entre sí.

Además, la proporción de verdaderos positivos frente a la proporción de falsos positivos expresada en ROC resulta en un área bajo la curva ROC (AUC) de 69.5%. La precisión de 45.5% será entendida como la probabilidad de que la paciente tenga un diagnóstico relevante y la exhaustividad de 47.3% es la probabilidad de que un diagnóstico relevante se suscite. La precisión y la exhaustividad fueron cercanas al artículo original que bordeó el 76%. (Smith et al., 1988). Esto es un desafío pues según (Chauhan & Karvande, 2019;

Kavakiotis et al., 2017), la precisión de detección debe ser aceptable para que el sistema sea confiable, de modo especial en aspectos clínicos.

En este trabajo se logró una exactitud del 65% utilizando árboles de decisión en tanto que, con las reglas que se generaron sobre el conjunto de pruebas, las 20 primeras bordean entre el 70% y 90% de confianza como clasificador para la predicción de diabetes.



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	31	31	0.270	0.678	0.456	0.500	0.419	0.744	0.801	0.045
(0.800,0.900]	3	4	0.300	0.674	0.476	0.493	0.459	0.752	0.776	0.056
(0.700,0.800]	0	2	0.309	0.665	0.469	0.479	0.459	0.748	0.763	0.062
(0.600,0.700]	0	0	0.309	0.665	0.469	0.479	0.459	0.748	0.763	0.062
(0.500,0.600]	1	5	0.335	0.648	0.464	0.455	0.473	0.745	0.731	0.077
(0.400,0.500]	1	7	0.370	0.622	0.453	0.424	0.486	0.738	0.686	0.098
(0.300,0.400]	1	0	0.374	0.626	0.462	0.430	0.500	0.743	0.686	0.098
(0.200,0.300]	2	0	0.383	0.635	0.481	0.443	0.527	0.754	0.686	0.098
(0.100,0.200]	2	2	0.400	0.635	0.494	0.446	0.554	0.761	0.673	0.105
(0.000,0.100]	33	105	1.000	0.322	0.487	0.322	1.000	1.000	0.000	0.695

Ilustración 6: Resultados de la evaluación del Árbol de Decisión Reforzado de Dos Clases

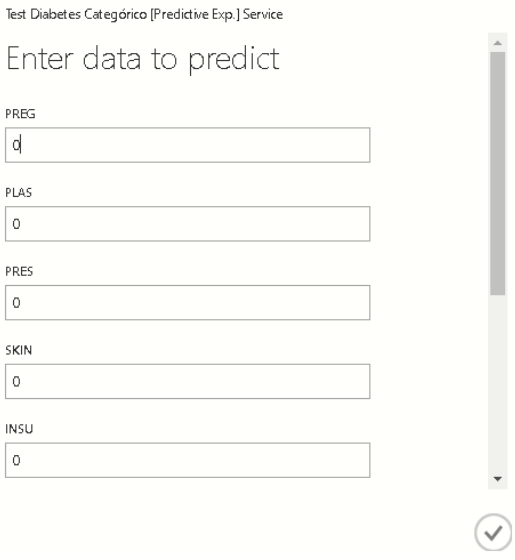


Ilustración 7: Interfaz del servicio web de Azure ML para probar el modelo

Adicional a los resultados numéricos presentados, de cierto modo, hay coincidencia con otros estudios que identifica la obesidad como un factor de riesgo, pero está claro que el riesgo aumenta a medida que aumenta el índice de masa corporal y que el sobrepeso es un factor de riesgo significativo sin la presencia de obesidad. (Dallo & Weller, 2003; Ollila et al., 2017).

Conclusiones

El descubrimiento de conocimientos a partir de bases de datos médicas es importante para realizar un diagnóstico médico eficaz, siendo posible procurar una descripción clara y comprensible de patrones, aun por buena parte de no especialistas como se procuró en este experimento (modelo) implementado en Azure ML, para el diagnóstico de diabetes sobre el conjunto de datos de la población de indias americanas PIMA.

El modelo constó de un Árbol de Decisión Reforzado y Reglas de Decisión generadas con Apriori, del árbol se evaluó la exactitud, la precisión y exhaustividad, la curva ROC, una matriz de confusión. De las reglas se evaluó la confianza de ellas. Los resultados fueron ligeramente mejores con reglas de asociación.

Todos los factores de riesgo incluidos en el análisis presentado, como variables predictoras médicas independientes, tienen una fuerte asociación con la diabetes, sin embargo, si se trata de establecer niveles de incidencia, el orden de mayor a menor sería: la concentración de glucosa, el índice de masa corporal, la edad, el número de embarazos y la insulina en la sangre. En menor medida, pero aclarando siempre que es también en función de los datos disponibles, aparecen la influencia genética, la presión arterial diastólica y el grosor del pliegue de la piel del tríceps.

Referencias bibliográficas

Affairs, D. of I. E. and S. (1982). *Provisional guidelines on standard international age classifications*. United Nations New York. https://tinyurl.com/y2u64erl

Ahouz, F., Sadehvand, M., & Golabpour, A. (2019). Extracting Rules for Diagnosis of Diabetes Using Genetic Programming. *International Journal of Health Studies*, *5*(3). https://tinyurl.com/y2oxp47a

Barga, R., Fontama, V., & Tok, W. H. (2015). Integration with R. In *Predictive Analytics*

with Microsoft Azure Machine Learning (pp. 81–101). Apress. https://doi.org/10.1007/978-1-4842-1200-4_4

Bashir, I., Mariod, A., Banu, R., & Elyas, T. (2019). Significance of Health Related Predictors of Diabetes in Pima Indians Women. *Current Research in Nutrition and Food Science Journal*, *7*, 350–359. https://doi.org/10.12944/CRNFSJ.7.2.05

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*, *26*(1–2), 37–54.

Chauhan, S. C., & Karvande, V. (2019). Improve classification perfomance in diabetes prediction. *Open Access International Journal of Science & Engineering*, *4*(10). https://tinyurl.com/y5oxqq5z

Dallo, F. J., & Weller, S. C. (2003). Effectiveness of diabetes mellitus screening recommendations. *Proceedings of the National Academy of Sciences*, *100*(18), 10574–10579. https://doi.org/10.1073/pnas.1733839100

Fufaa, G. D., Weil, E. J., Nelson, R. G., Hanson, R. L., Bonventre, J. V, Sabbisetti, V., Waikar, S. S., Mifflin, T. E., Zhang, X., Xie, D., Hsu, C., Feldman, H. I., Coresh, J., Vasan, R. S., Kimmel, P. L., & Liu, K. D. (2015). Association of urinary KIM-1, L-FABP, NAG and NGAL with incident end-stage renal disease and mortality in American Indians with type 2 diabetes mellitus. *Diabetologia*, *58*(1), 188–198. https://doi.org/10.1007/s00125-014-3389-3

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104–116.

Knowler, W. C., Pettitt, D. J., Saad, M. F., Charles, M. A., Nelson, R. G., Howard, B. V, Bogardus, C., & Bennett, P. H. (1991). Obesity in the Pima Indians: its magnitude and relationship with diabetes. *The American Journal of Clinical Nutrition*, *53*(6), 1543S-1551S. https://doi.org/10.1093/ajcn/53.6.1543S

Madmoli, M., Madmoli, Y., Taqvaeinasab, H., Khodadadi, M., Darabiyan, P., & Rafi, A. (2019). Some influential factors on severity of diabetic foot ulcers and Predisposing of limb amputation: A 7-year study on diabetic patients. *International Journal of Ayurvedic Medicine*, *10*(1), 75–81. https://tinyurl.com/y48c58fk

Mund, S. (2015). *Microsoft azure machine learning*. Packt Publishing Ltd. https://tinyurl.com/y3k3g44m

Neel, J. V. (1976). Diabetes Mellitus — A Geneticist's Nightmare. In *The Genetics of Diabetes Mellitus* (pp. 1–11). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-66332-1_1

Ollila, M.-M., West, S., Keinänen-Kiukaanniemi, S., Jokelainen, J., Auvinen, J., Puukka, K., Ruokonen, A., Järvelin, M.-R., Tapanainen, J. S., Franks, S., Piltonen, T. T., & Morin-Papunen, L. C. (2017). Overweight and obese but not normal weight women with PCOS are at increased risk of Type 2 diabetes mellitus—a prospective, population-based cohort study. *Human Reproduction*, *32*(2), 423–431. https://doi.org/10.1093/humrep/dew329

Olsson, T., & Goedecke, J. H. (2020). Obesity and type 2 diabetes: understanding the role of ethnicity. *Journal of Internal Medicine*, *288*(3), 269–270. https://doi.org/10.1111/joim.13043

Pekel, E., & Özmen, E. P. (2020). Computational Intelligence Approach for Classification of Diabetes Mellitus Using Decision Tree. In *Computational Intelligence and Soft Computing Applications in Healthcare Management Science* (pp. 87–103). IGI Global. https://doi.org/10.4018/978-1-7998-2581-4.ch005

Pettitt, D. J., & Knowler, W. C. (1988). Diabetes and obesity in the Pima Indians: a cross-generational vicious cycle. *Journal of Obesity and Weight Regulation*. https://tinyurl.com/y2vn3xhs

Pincay-Ponce, J. I., Angulo-Murillo, N. G., Herrera-Tapia, J. S., & Delgado-Muentes, W. R. (2020). Técnicas de minería de datos como soporte para la gestión de un sistema de comercialización de energía eléctrica. *Mikarimin. Revista Científica Multidisciplinaria. e-ISSN 2528-7842*, *6*(2), 19–34. https://tinyurl.com/y47xak9m

Pluim, J. P. W., Maintz, J. B. A., & Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, *22*(8), 986–1004. https://doi.org/10.1109/TMI.2003.815867

Ramesh, S., Caytiles, R. D., & Iyenga, N. C. S. . (2017). A Deep Learning Approach to Identify Diabetes. *Advanced Science and Tech Letter*, *145*, 44–49. https://doi.org/10.14257/astl.2017.145.09

Ríos Canales, V. (2016). Using a Supervised Learning Model: Two-Class Boosted Decision Tree Algorithm for Income Prediction. *Computer Engineering;* https://tinyurl.com/y53afoc3

Savage, P. J., Bennett, P. H., Gordon Senter, R., & Miller, M. (1979). High Prevalence of Diabetes in Young Pima Indians: Evidence of Phenotypic Variation in a Genetically Isolated Population. *Diabetes*, *28*(10), 937–942. https://doi.org/10.2337/diab.28.10.937

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261.

Tulloch-Reid, M. K., Williams, D. E., Looker, H. C., Hanson, R. L., & Knowler, W. C. (2003). Do Measures of Body Fat Distribution Provide Information on the Risk of Type 2 Diabetes in Addition to Measures of General Obesity?: Comparison of anthropometric predictors of type 2 diabetes in Pima Indians. *Diabetes Care*, *26*(9), 2556–2561. https://doi.org/10.2337/diacare.26.9.2556

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/DOI 10.1007/s10115-007-0114-2