Predicción de probabilidad de diabetes en pacientes

Ana Estefanía Henao Restrepo, Juan José Gil Hoyos

Especialización en Analítica y Ciencia de Datos

Departamento de Ingeniería de Sistemas

Universidad de Antioquia, Colombia

aestefania.henao@udea.edu.co

jjose.gil@udea.edu.co

A. Comprensión del problema de aprendizaje automático

1. La diabetes se está convirtiendo en una enfermedad cada vez más común. Existe evidencia científica abundante que establece que los casos de diabetes están incrementando dramáticamente en el mundo. Tanto así, que se la ha catalogado como una epidemia para la humanidad por parte del foro económico mundial durante el año 2020 [1] [2].

Factores como el consumo excesivo de alcohol, una dieta desbalanceada, poca actividad física y fumar, tienen una relación estrecha con la probabilidad de contraer diabetes tipo 2 [1].

¿Qué implicaciones tiene para la salud en general la diabetes?

- Si tiene diabetes es mucho más probable que desarrolle una enfermedad cardiaca o derrame cerebral que aquellas que no la padecen [3].
- La asociación americana de enfermedades cardiovasculares, (AHA por sus siglas en inglés), establece que la diabetes disminuye los niveles de colesterol HDL (también conocido como colesterol bueno), aumenta los niveles de triglicéridos y los niveles de colesterol LDL (también conocido como colesterol malo). Estos últimos incrementan el riesgo de padecer enfermedades cardiovasculares y derrames cerebrales [4].

Pero la diabetes no solo ataca la salud física sino también tiene grandes consecuencias sobre la salud mental:

- Hay medicinas que pueden alterar las células beta o alterar el funcionamiento de la insulina, como los medicamentos para enfermedades psiquiátricas [5].
- Los enfermos con esquizofrenia tienen de 2 a 4 veces más riesgo de presentar diabetes tipo 2. Los antecedentes familiares de diabetes son más comunes en pacientes con esquizofrenia [6].
- Las personas con diabetes tienen entre 2 y 3 veces más probabilidades de presentar depresión que las personas sin diabetes [6].

Se pretende hacer uso del dataset extraído de la plataforma Kaggle para predecir si un paciente padece o no diabetes, basados en atributos relacionados con hábitos de alimentación y condición de salud. Por lo tanto, el problema a abordar será de clasificación.

El objetivo principal de nuestra variable de respuesta es determinar si un paciente tiene o no diabetes a partir de los datos dados. El link para consultar la fuente de datos es el siguiente: https://rb.gy/df76r [7]

Como métrica de negocio, se propone hacer uso del modelo predictivo que se desarrolle, para que, a partir de un conjunto de datos recolectados de un paciente sea posible ofrecerle al médico tratante una herramienta de apoyo al diagnóstico temprano de diabetes.

2. A continuación, se mencionan los principales hallazgos de los artículos consultados:

a. Sisodia et al., describieron en su artículo, *Prediction of Diabetes using Classification Algorithms,* el proceso que llevaron a cabo para predecir la diabetes en pacientes haciendo uso de diferentes modelos de clasificación sobre el dataset extraído del repositorio UCI machine learning, el cual se conoce como *Pima Indians Diabetes Database (PIDD)*. Fueron tres los algoritmos implementados: Árbol de decisión, Máquina de Soporte Vectorial (SVM, por sus siglas en inglés) y el modelo de Naive Bayes. Los algoritmos antes mencionados utilizan método de validación cruzada con 10 doblajes o folds y las métricas que permitieron cuantificar su desempeño fueron: precission, acurracy, F-measure y recall.

El resultado más destacado de las experimentaciones llevadas a cabo fue el 76.30% en accuracy, (el más alto

obtenido), así como el F Measure de 76%. Estos resultados se alcanzaron haciendo uso del modelo Naive Bayes. Adicionalmente, los resultados fueron verificados a través de la construcción de las *Reciever Operating Characteristics curves* (ROC, por sus siglas en inglés) [8].

b. Rosales realizó la investigación sobre: *Predicción de diabetes mellitus tipo 2 utilizando atributos médicos del Policlínico Leo SAC de San Juan de Lurigancho mediante el enfoque de Machine Learning.* Se desarrollaron y se ajustaron 13 modelos de Machine Learning de tres categorías: modelo clásico (regresión logística, árbol de decisión, Naive Bayes, KNN y SVM), red neuronal (perceptrón multicapa), y modelo ensemble (Random Forest, AdaBoost, LogitBoost, Gradient Boosting, XGBoost, LightGBM, y CatBoost) para predecir la diabetes mellitus tipo 2 por medio del método de validación cruzada (10 veces). Los modelos con hiperparámetros óptimos se evaluaron mediante el accuracy, precisión, sensibilidad, especificidad, F1-score, tasa de clasificación errónea y el AUC en el conjunto de datos de entrenamiento y de prueba.

De los 13 modelos de Machine Learning, el modelo que superó consistentemente a los demás fue LightGBM en las siete métricas de rendimiento: Accuracy (datos de entrenamiento: 0.9963, datos de prueba: 0.96), precisión (datos de entrenamiento: 0.9999, datos de prueba: 0.9298), sensibilidad (datos de entrenamiento: 0.9865, datos de prueba: 0.96), especificidad (datos de entrenamiento: 0.9999, datos de prueba: 0.9720), F1-score (datos de entrenamiento: 0.9932, datos de prueba: 0.9298), tasa de clasificación errónea (datos de entrenamiento: 0.0038, datos de prueba: 0.04) y el área bajo la curva característica operativa del receptor (AUC, por sus siglas en inglés) (datos de entrenamiento: 0.9999, datos de prueba: 0.9872), mientras que el modelo con el rendimiento más bajo en la mayoría de las siete métricas fue el árbol de decisión y Random Forest en la parte de entrenamiento y de prueba respectivamente [9].

c. Sánchez et al., realizaron la investigación sobre: Predicción de la diabetes mellitus tipo 2 en pacientes adultos mediante regresión logística binaria. Se utilizó la regresión logística binaria, con el apoyo del SPSS 25 por medio del método de validación cruzada. Como resultado, se logró predecir la probabilidad de que la población se enferme de DM2 (valor $P<0.05$). Como conclusión se evidenció que quienes tuvieron la predisposición genética e Hipertensión Arterial presentaron más riesgo de enfermar de DM2, no ocurriendo lo mismo con la edad y el sexo, cuyas relaciones no fueron significativas [10].

d. Maniruzzuaman et al., realizaron una investigación que fue descrita en el artículo *Classification and prediction of diabetes disease using machine learning paradigm*. El dataset utilizado, proviene de la *National Health and Nutrition Examination Survey* que se evaluó entre los años 2009 al 2012. Los algoritmos implementados durante la experimentación fueron de clasificación, entre los que se mencionan: Regresión logística, el modelo de Naïve Bayes, el algoritmo de Árbol de decisión, Adaboost y el método Random Forest.

El método de validación que se utilizó para evaluar el desempeño de los modelos implementados es el método de validación cruzada, que consistió en utilizar otro dataset que proviniera de datos reales. Para este caso en particular, se utilizaron los datos de la Universidad de California, Irvine, la cual cuenta con un repositorio de datos asociado con la información de pacientes provenientes de la India que fueron diagnosticados con diabetes. Las métricas de desempeño que permitieron hacer una evaluación de la validez de los modelos utilizados fueron el accuracy global y el área bajo la curva. El máximo valor alcanzado de accuracy global fue 94.25%, mientras que el máximo valor de área bajo la curva y de F-Measure fueron de 0.95 y 96.88% con un protocolo de partición de 10. [11].

B. Entrenamiento y evaluación de modelos

Esta fuente de datos proviene de la limpieza de los resultados de una encuesta realizada por *Centers for Desease Control and Prevention (CDC)* por sus siglas en inglés, cuyo nombre fue *Behavioral Risk Factor Surveillance System*, la cual inicialmente contenía 441.456 registros y fue llevada a cabo durante el año 2015 [12]. Esta información se adaptó para utilizarse en una de las competiciones que se generan dentro de la herramienta Kaggle. Su lanzamiento dentro de esta plataforma se produjo en noviembre de 2022.

El dataset consta de tres archivos en formato csv: **el primero** de estos corresponde a los datos que servirán como insumo para entrenar el modelo predictivo que se proponga, **el segundo** servirá para poner a prueba el modelo y **finalmente**, un tercer archivo de prueba para llevar a cabo el cargue de las respuestas del reto (submission file). Dentro del dataset, existen 18 columnas, las cuales se dividen en 17 variables características y una sola variable de respuesta. El archivo que permitirá entrenar el modelo tiene un total de 80.692 registros, el archivo de puesta en marcha tiene 20.000 registros, al igual que el archivo de cargue de respuestas.

Como **métricas de desempeño,** para este caso de estudio la mejor opción es el F1 score, sin embargo, hay que recordar que es necesario también analizar la matriz de confusión. Se selecciona un F1 score aproximado de 75% como mínimo. Finalmente, se propone utilizar como métrica de validación la pérdida logarítmica, la cual se calculará una vez se tenga alguna predicción del modelo propuesto.

Como **métrica de negocio**, se propone hacer uso del modelo predictivo que se desarrolle, para que, a partir de un conjunto de datos recolectados de un paciente sea posible ofrecerle al médico tratante una herramienta de apoyo al diagnóstico temprano de diabetes.

C. Resultados y discusión

Para la determinación del modelo de clasificación más adecuado en la predicción de la variable de respuesta, se evaluó el desempeño de siete modelos diferentes. Estos fueron: regresión logística, KNeighbors Classifier, clasificador de Naive Bayes para modelos Bernoulli multivariable, Random

forest, Support Vector Classifier, XGBoosting Classifier y Adaptive Boost Classifier.

Inicialmente, se utilizó el algoritmo de búsqueda de rejilla, (Grid search), para determinar los mejores hiper parámetros de cada uno de los modelos mencionados. Durante este proceso, se hizo uso del 80% de la base datos de entrenamiento con y sin eliminación de datos atípicos a través del algoritmo LOF para los modelos de: regresión logística, KNN, BernoulliNB y random forest. Para los modelos faltantes, (SVC, XGBoost y Adaboost), debido a su alto costo computacional, se optó por aplicar el Grid search con 4000, 8000 y 40000 filas del total de la base datos con y sin eliminación de datos atípicos.

La búsqueda de rejilla realizada arrojó los siguientes valores de hiper parámetros óptimos para cada uno de los modelos, así como la métrica f1 score de prueba más alta:

Tabla 1. Hiper parámetros óptimos y métrica de desempeño

Modelo	Hiper parámetros	Valores óptimos		Métrica de desempeño evaluada	
				mean_test_f1_score	
		Sin eliminación de datos atípicos	Con eliminación de datos atípicos	Sin eliminación de datos atípicos	Con eliminación de datos atípicos
Regresión logística	**C**	0.1	0.1	75.75%	76.76%
	Penalty	l1	l1		
	Solver	Saga	liblinear		
KNN	**n_neighbors**	93	21	75.24%	76.12%
	metric	manhattan	manhattan		
Bernoulli Naive Bayes	**Binarize**	0,35	0,35	72.82%	73.43%
	Alpha	0	0,5		
	Force_alpha	True	False		
Random Forest	**n_estimators**	500	1000	75.27%	75.96%
	max_depth	5	5		
	criterion	Gini	gini		
Support Vector Classifier	**Registros del dataset utilizados durante el Grid Search y el cross validation**	40.000	40.000	76.10%	77.33%
	C	1	10		
	Kernel	rbf	rbf		
	gamma	Auto	auto		
XGBoost Classifier	**Registros del dataset utilizados durante el Grid Search y el cross validation**	40.000	40.000	76.08%	77.38%
	learning_rate	0.06	0.06		
	n_estimators	500	500		
	max_depth	3	3		
Adaboost Classifier	**Registros del dataset utilizados durante el Grid Search y el cross validation**	40.000	40.000	75.68%	76.81%
	learning_rate	0.3	0.1		
	n_estimators	500	500		
	algorithm	SAMME.R	SAMME.R		

Posteriormente, se implementó un modelo de validación cruzada haciendo uso de todos los modelos evaluados con sus respectivos hiper parámetros óptimos, y tomando un 10% adicional de los datos del dataset que no se utilizaron durante la búsqueda de rejilla. Para el caso del escenario sin eliminación de datos atípicos, el mayor valor de la métrica F1 score de prueba es del 77.35% y se alcanzó con el modelo de SVC. Por otra parte, para el escenario con eliminación de datos atípicos, el máximo valor de F1 score de prueba es del 78.58% y se obtuvo con el modelo de XGBoost.

Finalmente, se utiliza el 10% de los datos del dataset que no fueron implementados ni en la búsqueda de rejilla ni en la validación cruzada para simular en producción haciendo uso de los modelos de SVC para el escenario sin eliminación de datos atípicos y el modelo XGBoost para el caso de eliminación de datos atípicos. Para el caso del modelo SVC, el valor de F1 score que se obtuvo durante la predicción de la clase 0 en la variable de respuesta, (la cual representa el hecho de que un paciente no padece de diabetes), fue de 72%, mientras que para la clase 1, (la cual representa el hecho de que un paciente sí padece de diabetes), fue del 76%. Por otra parte, para el modelo XGBoost, el valor de F1 score que se obtuvo para la predicción en la clase 0 fue del 74%, mientras que para la predicción en la clase 1 fue del 77%. Estos resultados nos permiten concluir que la eliminación de datos atípicos a través del algoritmo LOF permitió obtener el valor mínimo de la métrica que se propuso alcanzar (75% en promedio). Sin embargo, no existe una diferencia significativa entre el valor que se alcanzó luego de implementar este mecanismo de eliminación controlada de datos atípicos con respecto a cuando no se utilizó.

Si bien se logró un mejor desempeño con los modelos SVC y XGBoost, su alto costo computacional los hace poco prácticos, razón por la cual se opta por emplear en futuras versiones de este proyecto el modelo de regresión logística, el cual logró el F1 score esperado durante la aplicación de la búsqueda de rejilla. Sin embargo, se hará una validación para comprobar que este dataset cumple con los supuestos iniciales para poder aplicar este tipo de algoritmos.

Comparando los resultados obtenidos en la **Tabla 1** con los resultados de los artículos, se evidencia que los modelos implementados arrojan resultados similares, ya que permiten obtener una métrica de desempeño (F1 score) muy cercana a la obtenida en el artículo a, mientras que los demás artículos obtuvieron resultados con métricas muy cercanas a 1, lo que puede deberse a la implementación de modelos más robustos, y que son más acordes con el contexto de la base de datos, como por ejemplo modelos de redes neuronales de perceptrón multicapa, los cuales podrían conducir a mejores resultados en cuanto a la predicción en comparación con un modelo de machine learning supervisado como los que se utilizaron durante las experimentaciones de este proyecto.

D. Bibliografía

[1] O. Folorunso y O. Oguntibeju. "The Role of Nutrition in the Management of Diabetes Mellitus". IntechOpen - Open Science Open Minds | IntechOpen. https://www.intechopen.com/chapters/42086 (accedido el 15 de marzo de 2023).

[2] Centro Nacional para la Prevención de Enfermedades Crónicas y Promoción de la Salud, División de Diabetes Aplicada. "La diabetes y su corazón". Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/spanish/resources/features/diabetes-and-heart.html#:~:text=Modo%20en%20la%20diabetes%20

afecta%20el%20corazón&text=La%20presión%20arterial%20alta%20aumenta,el%20riesgo%20de%20enfermedad%20cardiaca (accedido el 15 de marzo de 2023).

[3] C. M. Story. "A Guide to Living with Diabetes and High Cholesterol". Healthline. https://www.healthline.com/health/high-cholesterol/treating-with-statins/guide-to-diabetes-and-high-cholesterol#:~:text=Diabetes%20and%20high%20cholesterol%20often%20occur%20together&text=The%20American%20Heart%20Association%20(AHA,for%20heart%20disease%20and%20stroke (accedido el 15 de marzo de 2023).

[4] Instituto Nacional de la Diabetes y las Enfermedades Digestivas y Renales. "Síntomas y causas de la diabetes - NIDDK". National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/sintomas-causas (accedido el 15 de marzo de 2023).

[5] S. B. YP. "DIABETES, TRASTORNOS PSIQUIÁTRICOS E INTERACCIONES ENTRE AMBAS ENTIDADES". Bienvenidos a siicsalud. https://www.siicsalud.com/dato/resiiccompleto.php/130807#:~:text=Los%20enfermos%20con%20esquizofrenia%20tienen,en%20los%20pacientes%20con%20esquizofrenia (accedido el 15 de marzo de 2023).

[6] Centro Nacional para la Prevención de Enfermedades Crónicas y Promoción de la Salud, División de Diabetes Aplicada. "La diabetes y la salud mental". Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/spanish/living/mental-health.html (accedido el 15 de marzo de 2023).

[7] TFUG Chandigarh. "Diabetes Prediction Competition(TFUG Chd Nov 2022) | Kaggle". Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/competitions/diabetes-prediction-competitiontfug-chd-nov-2022/data (accedido el 15 de marzo de 2023).

[8] D. Sisodia y D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Comput. Sci., vol. 132, pp. 1578–1585, 2018. Accedido el 9 de mayo de 2023. [En línea]. Disponible: https://doi.org/10.1016/j.procs.2018.05.122

[9] "Predicción de diabetes mellitus tipo 2 utilizando atributos médicos del Policlínico Leo SAC de San Juan de Lurigancho mediante el enfoque de Machine Learning", *TecnoHumanismo*, vol. 2, n.º 4, 2022. Accedido el 9 de mayo de 2023. [En línea]. Disponible: https://doi.org/10.53673/th.v2i4.123

[10] B. Sánchez Martínez, V. Vega Falcón y N. Gómez Martínez, "Predicción de la diabetes mellitus tipo 2 en pacientes adultos mediante regresión logística binaria.", Dilemas contemporáneos: Educación, Política y Valores, mayo de 2021. Accedido el 9 de mayo de 2023. [En línea]. Disponible: https://doi.org/10.46377/dilemas.v8i3.2675

[11] M. Maniruzzaman, M. J. Rahman, B. Ahammed y M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm", *Health Inf. Sci. Syst.*, vol. 8, n.º 1, enero de 2020. Accedido el 9 de mayo de 2023. [En línea]. Disponible: https://doi.org/10.1007/s13755-019-0095-z

[12] National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. "CDC - 2015 BRFSS Survey Data and Documentation". Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2015.html (accedido el 15 de marzo de 2023).