Predicción de diabetes y el análisis estadístico para encontrar datos no esenciales en la predicción dentro de distintos segmentos de datos

Fabian A. Trigo Faúndez
Estudiante de Licenciatura en Física de la Universidad de Valparaíso

Preparado como propuesta de proyecto de clase para LFIS 419: Inteligencia Artificial. Entregado el 19/04/23 al profesor Jorge Arévalo.

Resumen
Se predice la presencia de diabetes mediante redes neuronales y se encuentran los datos de entrada menos importantes para cada grupo definido por variables fáciles de observar sin equipo médico, como el BMI y la edad.
Se dividen los datos utilizando "k-nearest neighbors" y se realiza la predicción dentro de cada uno de estos grupos; se entrenan modelos con ciertas neuronas de entrada apagadas y se compara su predicción; se utilizarán los datos de un instituto de salud, ver (3) y (4).
Al finalizar se mostrará que grupo de datos son suficientes para predecir la diabetes con un puntaje de aciertos por sobre el 95% dentro de un grupo especificado por el BMI, la edad y la cantidad de embarazos.

Introducción
La diabetes es una "pandemia", afectando a cada vez más personas, actualmente unos 537 millones de adultos (1), la gravedad del problema aumenta por el hecho que solo el 50% de ellos saben que la padecen, se proyecta que habría unos 30 millones de personas con diabetes en América del sur y casi 50 millones con una proyección para e 2040 (2). El ser capaces de predecir la diabetes con costo mínimo en tiempo y dinero puede significar una optimización esencial para la batalla contra la enfermedad.

Se intentará resolver cuales son las variables esenciales, ósea las que son capaces de predecir la diabetes con un 95% de aciertos; las cuales podrían ser o no distintas para cada grupo definido por edad, BMI y cantidad de embarazos.

Datos y Métodos
Este conjunto de datos proviene del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, ver (3) y (4); se utiliza para predecir si una paciente tiene diabetes basándose en medidas diagnósticas. Las instancias del conjunto de datos fueron seleccionadas con restricciones de una base de datos más grande, y todas las pacientes son mujeres de al menos 21 años, de ascendencia india Pima; temporalmente entre abril del 2016 y abril del 2022.

Se encuentran en formato CSV, con variables como cantidad de embarazos, nivel de glucosa en sangre, presión sanguínea, grosor de piel, nivel de insulina en sangre, índice masa corporal, porcentaje de diabetes producto de genética, edad y si tiene diabetes o no.

El análisis previo será separar los datos en grupos definidos por variables fáciles de medir, en este caso el BMI, la edad y la cantidad de embarazos; datos capaces de obtenerse en una conversación; esta separación se realiza con algoritmos K vecinos cercanos implementada en *scikitlearn*. Para la predicción se utilizará una red neuronal simple, construida en *Pytorch*, a la cual se le encontrarán los hiperparametros ideales para predecir la diabetes teniendo todos los datos de manera iterativa. Finalmente se hará una comparación de predicción con las variables no esenciales utilizando un árbol de decisión que tenga y que no tenga acceso a estos datos censurables al modelo.

Las variables esenciales y las no esenciales en la predicción se comparan con aquellas obtenidas por el uso de PCA (Análisis de componentes principales), ósea aquellas que expliquen mayor la correlación.

Las simplificaciones son en cuanto a la población utilizada, de menos de 1000 datos para mujeres de descendencia indias Pima con más de 21 años; pues esto ya significa una muestra más especifica que intentar la población mundial.

Referencias

(1) Facts & figures. (2022). Idf.org. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
(2) Sapunar Z., J. (2016). EPIDEMIOLOGÍA DE LA DIABETES MELLITUS EN CHILE. *Revista Médica Clínica Las Condes*, *27*(2), 146–151. https://doi.org/10.1016/j.rmclc.2016.04.003
(3) Akshay Dattatray Khare. (2022). *Diabetes Dataset*. Kaggle.com. https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download
(4) 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. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.