

**Instituto Tecnológico y de Estudios
Superiores de Monterrey**
Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I
TC3006C.101


Análisis del contexto y normatividad



Arturo Garza Campuzano A00828096

28 ago 2023

Proyecto seleccionado

El proyecto seleccionado para realizar este análisis de contexto y normatividad es **KNN Algorithm In Machine Learning | KNN Algorithm Using Python | K Nearest Neighbor | Simplilearn**, el cual se puede encontrar como una sección en el siguiente documento: [ImplementacionKNN.ipynb](#). Este proyecto es una implementación del modelo clasificador KNN y se realizó con el sólo propósito de familiarizarse con el modelo. Cabe mencionar que la implementación no es de mi propiedad sino que sólo fue un seguimiento del siguiente tutorial disponible en  **KNN Algorithm In Machine Learning | KNN Algorithm Using Python | K Nearest...**

Normativa asociada al tipo de datos

En este proyecto, se utilizó un conjunto de datos de pacientes del sexo femenino de al menos veintiún años de ascendencia india prima (Mehmet Akturk) para implementar un modelo clasificador KNN para predecir si un paciente tiene diabetes o no. Una normativa asociada al tipo de datos que se usaron es la **Ley Federal de Protección de Datos Personales en Posesión de Particulares**, la cual tiene por objeto la protección de datos personales en posesión de particulares, con la finalidad de regular su tratamiento legítimo, controlado e informado, a efecto de garantizar la privacidad y el derecho a la autodeterminación informativa de las personas (Mendoza Iserte and Orozco Martínez).

Uso de los datos y cumplimiento normativo

Los datos se obtuvieron de la siguiente carpeta de Google Drive: [KNN](#). Esta carpeta se encuentra en la descripción del tutorial fuente de esta implementación. Investigando aparte del verdadero origen de este conjunto de datos se encontró la siguiente información:

- a) Dueños originales: National Institute of Diabetes and Digestive and Kidney Diseases.
- b) Donador de la base de datos: Vincent Sigillito (vgs@aplacen.apl.jhu.edu); Research Center, RMI Group Leader; Applied Physics Laboratory; The Johns Hopkins University; Johns Hopkins Road; Laurel, MD 20707; (301) 953-6231.
- c) Fecha de recuperación: 9 May 1990.

El conjunto de datos cuenta con las siguientes características:

- *Pregnancies*: número de veces que ha estado embarazada.
- *Glucose*: concentración de glucosa plasmática a las dos horas de la prueba de tolerancia oral a la glucosa.
- *BloodPressure*: presión arterial diastólica (mm Hg).
- *SkinThickness*: grosos del pliegue cutáneo del tríceps (mm).
- *Insulin*: insulina sérica de dos horas (mu U/ml)

- *BMI*: índice de masa corporal (peso en kg/(altura en m)²)
- *DiabetesPedigreeFunction*: función de pedigrí de diabetes.
- *Age*: Edad (años).
- *Resultado*: variable de clase (0 o 1).

Estos datos han sido utilizados en el pasado. Por ejemplo, un artículo de Smith et al. (1988) presenta el uso del algoritmo de aprendizaje ADAP para predecir el inicio de la diabetes mellitus. El estudio involucra una variable diagnóstica de valor binario relacionada con la diabetes según los criterios de la Organización Mundial de la Salud. El algoritmo proporciona predicciones de valor real, que se convierten en decisiones binarias utilizando un umbral. El rendimiento del algoritmo se evalúa utilizando métricas de sensibilidad y especificidad en un conjunto de datos de 768 instancias.

Considerando las siguientes observaciones sobre el origen, las características y el uso pasado de los datos se puede constatar que el uso de este conjunto de datos no incumple la normativa establecida:

1. El origen de los datos no es ambiguo y, al parecer, fueron donados por alguien llamado Vincent Sigillito.
2. Si bien, dentro de las características hay información que puede ser considerada sensible para los pacientes, el conjunto de datos presenta un cierto grado de anonimato debido a la omisión de datos todavía más personales.
3. Los datos se han utilizado para alimentar un algoritmo de aprendizaje, por lo que sugiere que el conjunto de datos se puede utilizar para fines académicos.

Sesgo ético de la herramienta

Los datos fueron cargados, divididos en subconjuntos para el entrenamiento del modelo, escalados (para evitar el sesgo de los resultados) y utilizados para elaborar un modelo KNN con la ayuda de `sklearn.neighbors.KNeighborsClassifier`. Al final se probó el modelo con el subconjunto de prueba y se obtuvieron las métricas de desempeño correspondientes.

A lo largo de todo este proceso no se tomó en cuenta la normativa, en cierto grado se asumió que el conjunto de datos era válido y fidedigno para la implementación. En cuyo caso no se procedió con cautela y hay que recordar que el desconocimiento de la ley no exime de su cumplimiento. Por lo tanto, como aprendizaje, siempre que se utilicen datos de cualquier clase se tiene que tomar en cuenta tanto el factor moral como el legal.

Escenarios de falta ética y mal uso

Suponiendo que el conjunto de datos proporcionara información más personal, no se especificara bien el origen de estos o no se haya usado previamente para fines legítimos, se abriría la posibilidad de presentarse escenarios de falta de ética y mal uso como:

- Violación de la privacidad: si los datos se utilizan de manera que sea posible identificar a pacientes individuales y se divulgan públicamente, se estaría violando la privacidad de los pacientes y sus datos personales.
- Comercialización no ética: si los datos se utilizan con fines comerciales sin el consentimiento adecuado de los pacientes, podría considerarse explotación de su información personal para beneficio económico.
- Falta de transparencia: si se utiliza un modelo de aprendizaje automático sin proporcionar información adecuada a los pacientes sobre cómo se tomarían las decisiones basadas en sus datos, podría erosionarse la confianza del sistema médico.

Referencias bibliográficas

1. Mehmet Akturk. "Diabetes Dataset." *Kaggle.com*, 2020, www.kaggle.com/datasets/mathchi/diabetes-data-set.
2. Simplilearn. "KNN Algorithm in Machine Learning | KNN Algorithm Using Python | K Nearest Neighbor | Simplilearn." *Www.youtube.com*, 6 June 2018, www.youtube.com/watch?v=4HKqjENq9OU&t=890s. Accessed 28 Aug. 2023.
3. Mendoza Iserte, Jonathan, and Roberto Orozco Martínez. "La Protección de Datos Personales En Los Expedientes Clínicos." *Iapp*, 4 June 2019, iapp.org/news/a/la-proteccion-de-datos-personales-en-los-expedientes-clinicos/.