

Abstract citation ID: ckae144.1156
Predicting diabetes prognosis using machine learning techniques in a Hungarian clinical database

Attila Nagy

A Nagy¹, R Bata¹, A Ghanem¹, M Mór ²

¹Department of Health Informatics, Faculty of Health Sciences, University of Debrecen, Debrecen, Hungary

²Institute of Social and Sociological Sciences, Faculty of Health Sciences, University of Debrecen, Nyiregyh za, Hungary

Contact: nagy.attila@sph.unideb.hu

Background: Chronic conditions such as type 2 diabetes mellitus have great impact on patients' quality of life. Although clinical databases provide a perfect base for research aimed at improving diabetes care, analyzing such databases requires extensive pre-processing, mainly due to the large amount of unstructured data. Our study aimed to present the steps of generating a dataset from a large clinical database, and to apply machine learning-based analytical techniques regarding.

Methods: Data of the Clinical Center of University of Debrecen was used. To structure the unstructured data, regular expressions and natural language processing methods were used. The main machine learning models were as follows: Gradient Boosting Machines to predict the risk of complication development; Long Short-Term Memory Networks to forecast future health outcomes. All analysis and procedures were done using Python.

Results: The database contains approximately 1600 tables, with a total size of 1.9 terabytes, where the largest table is 21.07 gigabytes with approximately 44.64 million rows. The final dataset consisted of 40,332 patients. Most variables originate from the unstructured data, including complications and comorbidities of diabetes, as well as physical and laboratory parameters. Related to laboratory parameters, the number of measurements and the median value for every half-year were available. The diagnosis time of the complications' occurrence is also presented. Machine learning methods were more accurate compared with traditional statistical methods in predicting the prognosis ($p < 0.05$).

Discussion: Our research highlights the importance of clinical data in chronic disease management. There are challenges in pre-processing and managing datasets, but machine learning-based methods are very efficient not only in extracting useful information from unstructured data but also in predicting the prognosis and identifying potential intervention points for better care.

Key messages:

- Natural language processing can be used to obtain useful information from unstructured clinical data.
- Using machine learning techniques on clinical data could improve diabetes care.