PROJEKAT STUDENTSKIH ISTRAŽIVANJA - PSI 2024

i encoding

Aida Branković, Jin Yoon

The Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Health data are characterised by the large numbers of variables categorical in their nature. One-hot encoding is commonly used in health data analysis when dealing with categorical variables. It is a popular technique used in machine learning to represent categorical variables numerically. It converts categorical variables into binary vectors, where each category is represented by a binary value (0 or 1) in a separate feature. While one-hot encoding can be useful in certain scenarios, it can also have an impact on model performance e.g., exacerbate the curse of dimensionality problem and model complexity and interpretability. This project aims to develop novel encoding method using real health data and compare it against the conventional such as one-hot encoding and embedding-based.

Required skills: good coding skills in Python, passed courses in machine learning and AI.