На сьогоднішній день аналіз великої кількості даних займає ключове місце у виявленні прихованих закономірностей і тенденцій, які не відразу видно з узагальнених даних. Так як дані мають складні структури та великі розміри, науковці працюють над розробкою методів зменшення розмірності великих вибірок даних. Зменшення розмірності як етап попередньої обробки машинного навчання є ефективним у видаленні нерелевантних і надлишкових даних, підвищення точності навчання та покращення зрозумілості результату, за допомогою візуалізації розмірності. А також, процес зменшення кількості аналізованих випадкових величин шляхом отримання набору основних змінних. Проте зменшення розмірності має недолік, пов'язаний з втратою даних. Дуже важливо зменшити розмірність набору даних без втрати будь-якої інформації з цих наборів даних. У статті розглянуто існуючі методи зменшення розмірності великих вибірок даних, а саме: аналіз головних компонент (Principal Component Analysis), лінійний дискримінантний аналіз (Linear Discriminant Analysis), аналіз головних компонент ядра (Kernel Principal Component Analysis), багатовимірне масштабування (MDS), t-розподільного стохастичного вбудовування сусідів (t-SNE) та аналіз незалежних компонент (Independent Component Analysis).

Кожен з методів має свої переваги та недоліки, для вибору найбільш оптимального методу зменшення розмірності великих вибірок даних було проведено їх порівняльний аналіз. На наборі даних ініціативи з нейровізуалізації хвороби Альцгеймера та на наборі даних про щитоподібну залозу було протестовано кожен з розглянутих методів.

Результати порівняльного аналізу методів було представлено у вигляді графічних зображень.

Ключові слова: вибірка; данні; розмірність; аналіз; дослідження; класифікація; візуалізація.

Today, big data analysis occupies a key place in finding hidden patterns and trends that are not immediately visible from generalized data. Since data has complex structures and large sizes, scientists are working on developing methods to reduce the dimensionality of large data sets. Dimensionality reduction as a pre-processing step in machine learning is effective in removing irrelevant and redundant data, increasing training accuracy, and improving the comprehensibility of the result by visualizing the dimensionality. It is also a process of reducing the number of analyzed random variables by obtaining a set of basic variables. However, dimensionality reduction has the disadvantage of losing data. It is very important to reduce the dimensionality of a dataset without losing any information from these datasets. This article discusses the existing methods for dimensionality reduction of large data sets, namely Principal Component Analysis, Linear Discriminant Analysis, Kernel Principal Component Analysis, Multidimensional Scaling, t-distributed stochastic neighborhood embedding and Independent Component Analysis.

Each of the methods has its advantages and disadvantages, and a comparative analysis was performed to select the most optimal method for reducing the dimensionality of large data sets. Each of the considered methods was tested on the Alzheimer's Disease Neuroimaging Initiative dataset and the thyroid dataset.

The results of the comparative analysis of the methods were presented in the form of graphic images.

Keywords sample; data; dimensionality; analysis; research; classification; visualization.