# Principal Component Analysis (PCA)

**To address the extreme multicollinearity problem,** Principal Component Analysis (PCA) will be carried out.

To make a ML model that can accurately predict WQI scores, which is the end goal of this project, feature reduction must be carried out on the data. Feature reduction techniques must be conducted before training the ML model to prevent problems of overfitting and to improve prediction accuracy and generalisation ability.

Principal component analysis (PCA) is an unsupervised feature reduction technique that constructs principal components from the relevant features in a dataset. It does so by linearly transforming correlated variables into a smaller number of uncorrelated variables which capture the majority of the variance. [1]

Code from Kaggle was utilised to conduct PCA analysis on our dataset. [2]

**Step 1: Standardise the data**

**Step 2: Perform PCA**

The principle components will be determined.

**References**

[1] B. Mwangi, T. S. Tian, and J. C. Soares, ‘A review of feature reduction techniques in neuroimaging’, *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, Apr. 2014, doi: 10.1007/s12021-013-9204-3.

[2] ‘Principal Component Analysis’. Accessed: Oct. 12, 2025. [Online]. Available: https://kaggle.com/code/ryanholbrook/principal-component-analysis