RV College of Engineering®

IEEE

FIU FLORIDA INTERNATIONAL UNIVERSITY

Fachhochschule Dortmund
University of Applied Sciences and Arts

8th International Conference on Computational Systems and Information Technology for Sustainable Solutions

Paper ID: 397

"Dimensionality Reduction via Graph-Based Feature Selection"

By

Hrishik Sai Bojnal
Parnika Singh
Dr. Anitha J

INTRODUCTION

- According to Cisco, the global internet usage in **2016** was **1** zettabyte. By **2025**, it is estimated to reach **175** zettabytes. This is depicted in Fig. 1.
- Data growth is unprecedented, almost reaching up to **2.5 Quintillion bytes**.
- Data handling dependents on the sector, tools in hand, development etc.
- All data consists of redundancies. The estimated amount of redundancies vary from **33% to 85%**
- According to a study by IBM, the cost of poor data quality for businesses in the US alone is estimated to be around **$3.1 trillion** per year.

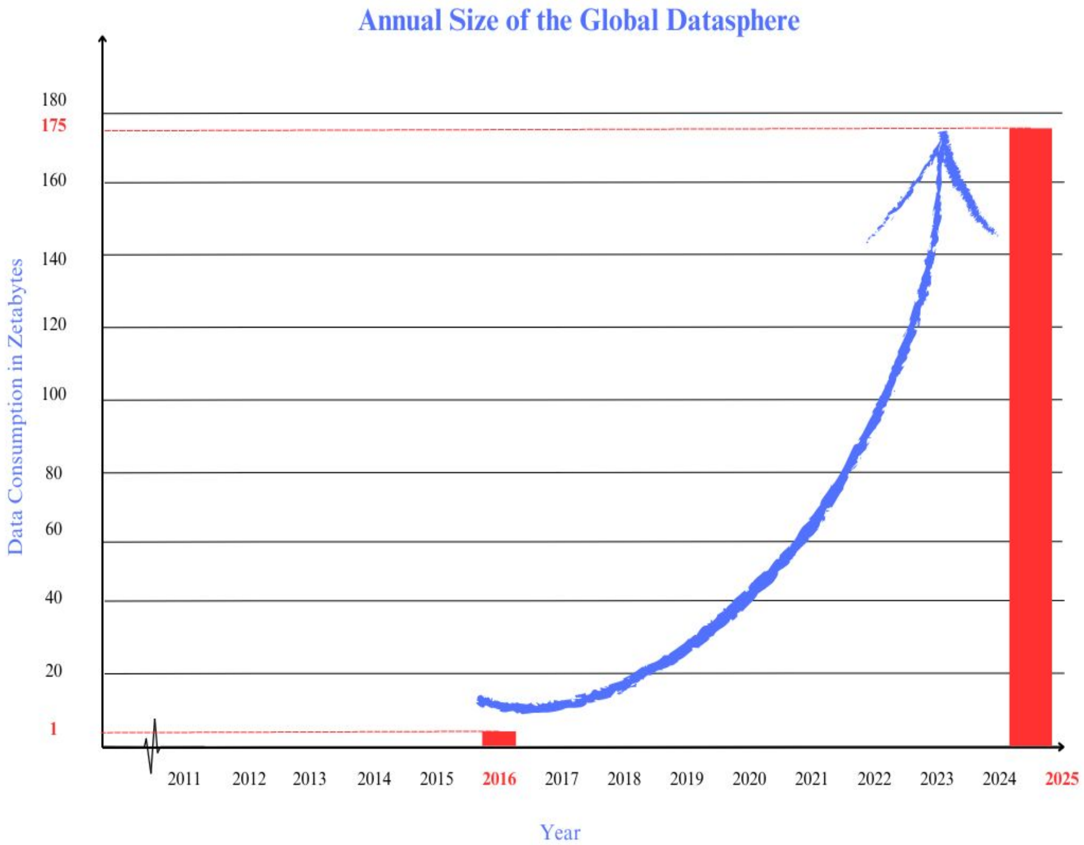


Fig. 1. Global Annual internet usage.

LITERATURE SURVEY

"Data with more features than observations cause a strain on the computational costs."

[1] S. Feng and H. Wang, "Comparison of PCA and LDA Dimensionality Reduction Algorithms based on Wine Dataset," in 33rd Chinese Control and Decision Conference (CCDC), 2021
Compares LDA with PCA – LDA is better at classification.

[2] A. Kazemipour and S. Druckmann, "Nonlinear Dimensionality Reduction Via Polynomial Principal Component Analysis," in 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2018.
Proposes Poly-PCA – Works on synthetic polynomial data. May not work for real life datasets.

[3] X. Sun, L. Liu, C. Geng and S. Yang, "Fast Data Reduction With Granulation-Based Instances Importance Labeling," IEEE Access, vol. 7, pp. 33587-33597, 2019.
Proposes a Granular Computing approach – Reduction rate at 46%. Can be further improved!

LITERATURE SURVEY

[4] R. N. S. Widodo, H. Abe and K. Kato, "Hadoop Data Reduction Framework: Applying Data Reduction at the DFS Layer," IEEE Access, vol. 9, pp. 152704-152717, 2021.
Introduces Hadoop Data Reduction Framework – Enables addition of own data reduction schemes to Hadoop.

[5] C. Gakii, P. Mireji and R. Rimiru, "Graph Based Feature Selection for Reduction of Dimensionality in Next-Generation RNA Sequencing Datasets," Algorithms, vol. 15, 2022.
Uses Maximal Clique – MSE at par with PCA at best, and sometimes outperformed by PCA.

[6] H. Zhang and M. Gabbouj, "Feature Dimensionality Reduction with Graph Embedding and Generalized Hamming Distance," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018.
Uses Graph Embedding – Achieves better performance than PCA, albeit slightly.

WORKFLOW

1. *Feature Redundancy*

- τ-redundancy: Defined between F_1 and F_2 if —> $R^2(F_1, F_2) \geq \tau$
- If two features are F_1 and F_2 are τ-redundant, we can approximate F_2 in terms of F_1 as:

$$F_1 \approx \beta_0 + \beta_1 \cdot F_2$$

- When one feature is linearly related to another, we can get rid of this feature altogether, without affecting the result of any machine learning algorithm.
- We will now prove this for neural networks.

FEATURE REDUCTION

The activation of a neuron is given by

$$activation = \phi(f_1 w_1 + f_2 w_2 + f_3 w_3 + \ldots + f_n w_n + b)$$

FEATURE REDUCTION

The activation of a neuron is given by

$$activation = \phi(f_1 w_1 + f_2 w_2 + f_3 w_3 + \ldots + f_n w_n + b)$$

If $\mathbf{f_1 \approx \beta_0 + \beta_1 \cdot f_2}$, then

$$activation = \phi(f_1 w_1 + (\beta_0 + \beta_1 \cdot f_1) \cdot w_2 + f_3 w_3 + \ldots + f_n w_n + b)$$

FEATURE REDUCTION

The activation of a neuron is given by

$$activation = \phi(f_1 w_1 + f_2 w_2 + f_3 w_3 + \ldots + f_n w_n + b)$$

If $f_1 \approx \beta_0 + \beta_1 \cdot f_2$, then

$$activation = \phi(f_1 \cdot (w_1 + w_2 \cdot \beta_1) + f_3 w_3 + \ldots + f_n w_n + (b + w_2 \cdot \beta_0))$$

Here, $(w_1 + w_2 \beta_1)$ and $(b + w_2 \beta_0)$ can be learnt as new parameters and respectively. Thus, the feature can be discarded altogether.

RV College of Engineering

IEEE

FIU FLORIDA INTERNATIONAL UNIVERSITY

Fachhochschule Dortmund
University of Applied Sciences and Arts

8th International Conference on Computational Systems and Information Technology for Sustainable Solutions

REMOVING REDUNDANT FEATURES

- **Data is visualised as a correlation graph**
- **The nodes represent the features and the edges represent the correlation between them.**
- **Articulation points are chosen carefully to avoid affecting the connectivity of the graph.**
- **The non-articulation points are dropped. These are our non-essential or redundant columns.**
- **The articulation points represent the basic structure of the graph.**
- **This ensures that the dataset integrity is maintained while reducing the number of redundant columns and the essential columns are retained.**
- **In Fig. 2,**
  - **By removing node *F24*, the 2 other components are disconnected.**
  - **Therefore, *F24* is an essential column.**
  - **By removing node *F26*, the graph connectivity is not affected.**
  - **Therefore, *F26* is not a redundant column.**

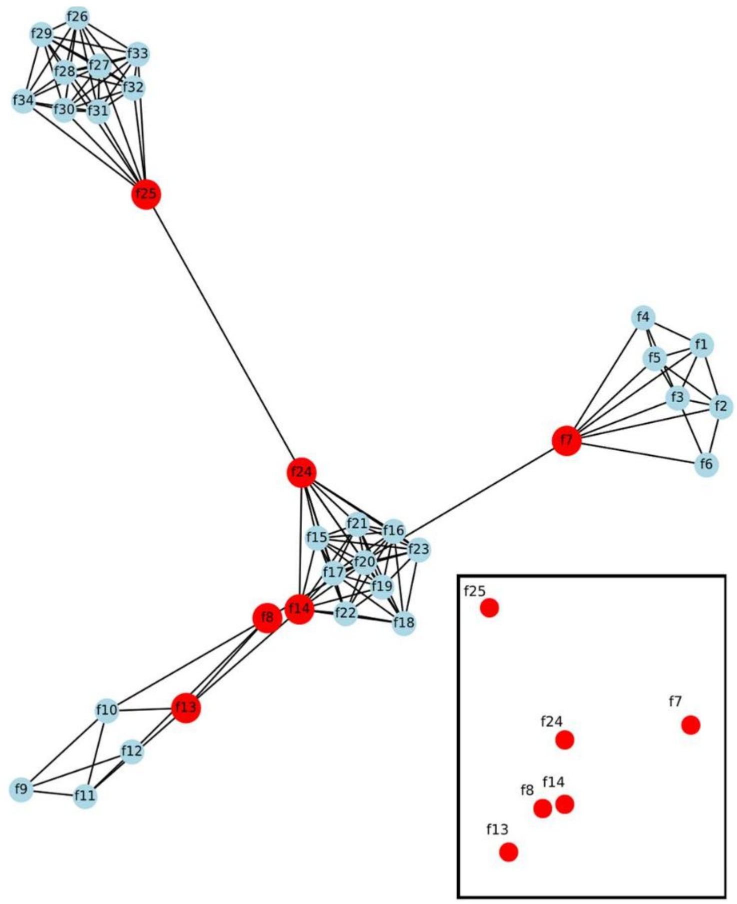
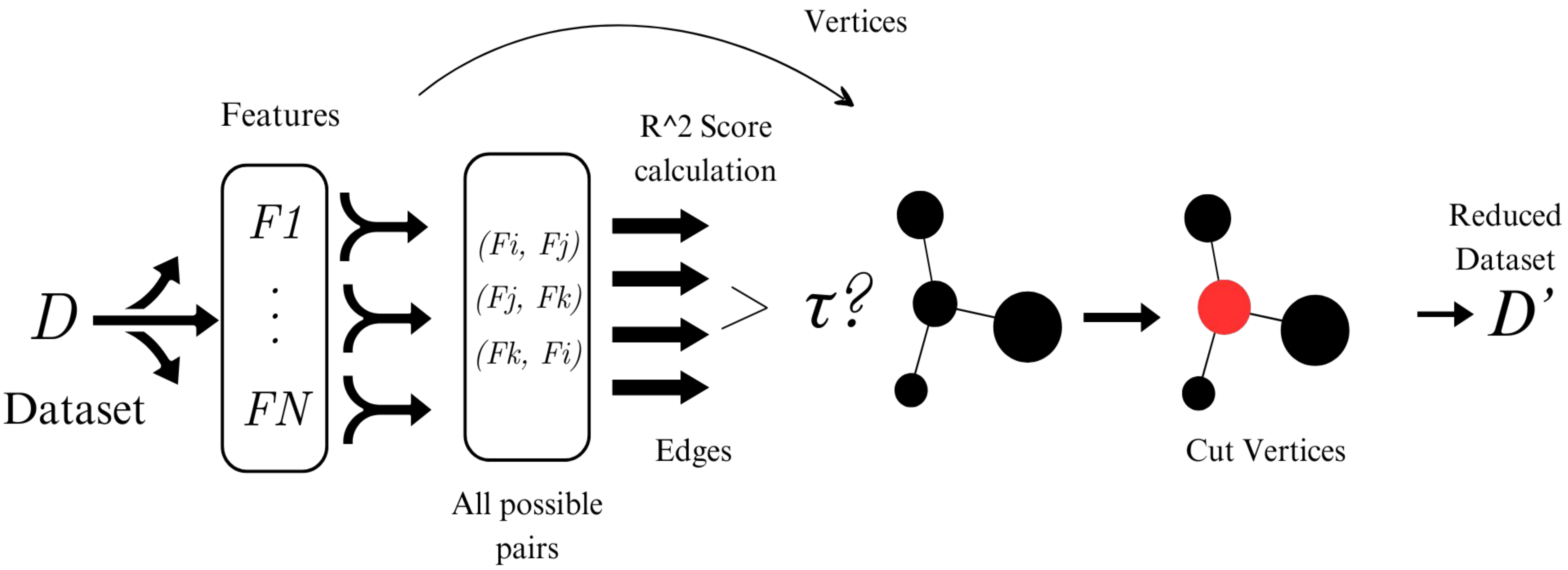


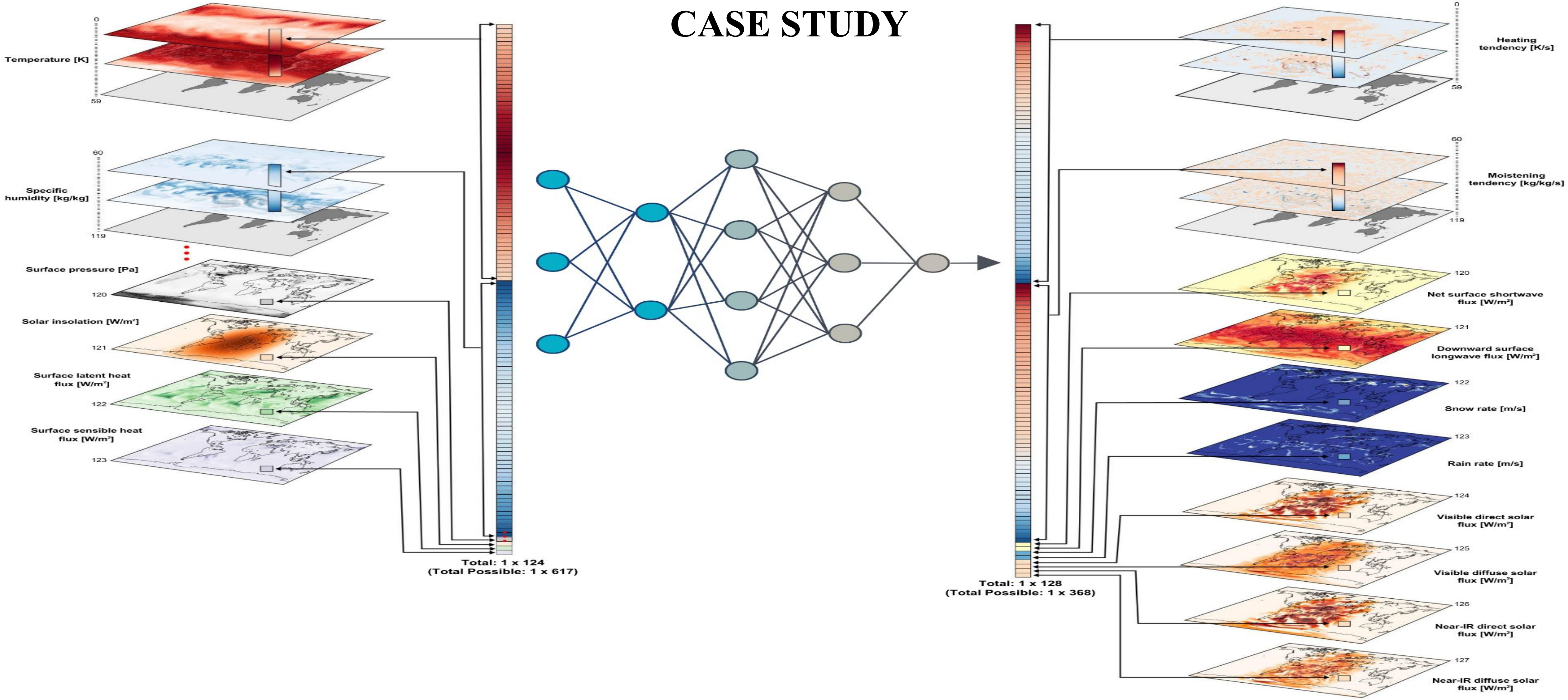
Fig. 2. Articulation points (shown in red) summarize the general structure of the graph.

ALGORITHM


Vertices
Features
R^2 Score
calculation
D
Dataset
F1
FN
(Fi, Fj)
(Fj, Fk)
(Fk, Fi)
All possible
pairs
Edges
> τ?
Cut Vertices
Reduced
Dataset
D'

8th International Conference on Computational Systems and Information Technology for Sustainable Solutions

CASE STUDY



RV College of Engineering

IEEE

FIU FLORIDA INTERNATIONAL UNIVERSITY

Fachhochschule Dortmund University of Applied Sciences and Arts

8th International Conference on Computational Systems and Information Technology for Sustainable Solutions

CASE STUDY

- **ClimSim** - Energy Exascale Earth System Model (E3SM)-Multiscale Modelling Framework (MMF) model.
- It consists of 60 levels, 556 columns corresponding to 25 input variables and 368 columns corresponding to 14 target variables.
- The dataset is widely used in improving climate prediction systems.
- Due to its high granularity and versatility, it adversely affects the computational costs in these high level processes.
- It includes various features over 60 environmental layers as depicted in Fig. 3.
- The data is split in a **80-20 ratio** and a Random Forest Regressor is applied over the First **1000 rows**.
- We now check for results **w.r.t 3 cases** :
  - ***Without any reduction technique applied.***
  - ***By applying PCA.***
  - ***By applying the proposed technique.***

Albedo for diffuse longwave radiation	Albedo for direct longwave radiation	Albedo for diffuse shortwave radiation	Albedo for direct shortwave radiation	Nitrous oxide volume mixing ratio
Air temperature	Specific humidity	Cloud liquid mixing ratio	Cloud ice mixing ratio	Zonal wind speed
Meridional wind speed	Surface pressure	Solar insolation	Surface latent heat flux	Surface sensible heat flux
Zonal surface stress	Meridional surface stress	Cosine of solar zenith angle	Upward longwave flux	Sea-ice areal fraction
Land areal fraction	Ocean areal fraction	Snow depth over land	Ozone volume mixing ratio	Methane volume mixing ratio

Fig. 3. Features present in the ClimSim dataset (The essential attributes are represented in green).

RV College of Engineering®

IEEE

FIU FLORIDA INTERNATIONAL UNIVERSITY

Fachhochschule Dortmund
University of Applied Sciences and Arts

8th International Conference on Computational Systems and Information Technology for Sustainable Solutions

CASE STUDY

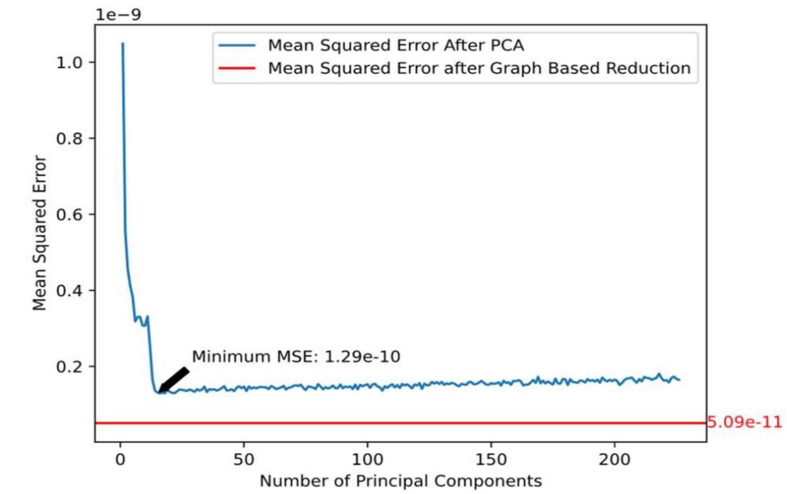


Fig. 4. A comparison of the mean squared errors of the dataset after PCA reduction (shown in blue), against the proposed method.

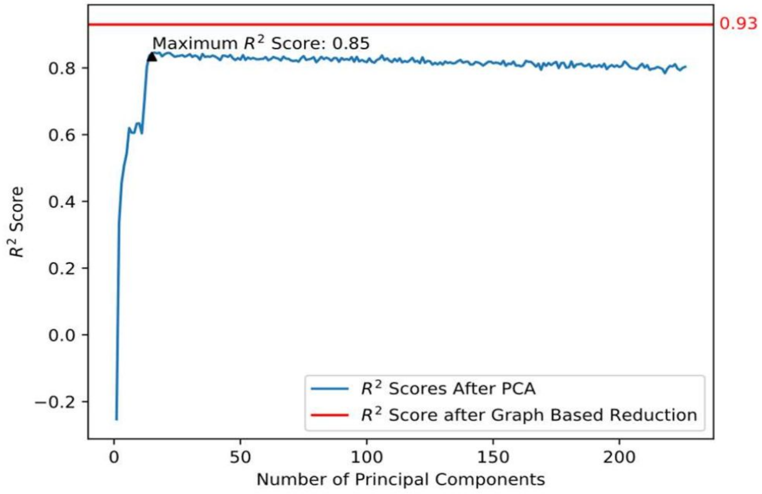


Fig. 5. A comparison of the R-squared scored of the dataset after PCA reduction (shown in blue), against the proposed method.

Method	Reduction	R-Squared Value
Random Forest Regressor	**-**	**0.979**
PCA	**15 components at best.**	**0.85**
Proposed Graph based reduction technique	**48.56%**	**0.939**

Fig. 6. The results after applying various methods over the dataset.

CONCLUSION

- With pre-existing technologies outperforming each other in specific scenarios. Our work proves to yield better results compared to PCA over a hugely versatile and granular dataset.

- Our approach is comparatively faster.
  - If **N = no. of features.**
  - **Worst case = O(N^2)**

- We also achieve a 48.56% reduction in the dataset while a good R-squared score of 0.939. The method also avoids causing a huge loss over the dataset trainability.

- It is evident that our work continues to be advantageous over other options that are available.

CONCLUSION

- Artificial intelligence is a field that is continuing to bloom. It has developed extensively in the past decade and is the foundation of many different services out there.

- This paves the path towards improvement of various possible fields. It will continue to prove the intelligence of human thinking since the earliest recorded time.

- Often times, we tend to forget about the basic processes and factors that can ultimately adversely affect the goal of our work.

- For example, by carefully examining the quality of the data we use, we can improve the output of our model.

- These forgotten processes are what we aim to improve. With this work just being the beginning of it.

Thank You!