Technical University of Denmark

Gas Turbine Emissions: A Classification and Regression Task

02450 Introduction to Machine Learning and Data Mining

Jonas Wiendl, s243543
Linnea Hockauf, s204547
Rami Ishac Hanna, s242507

October 3, 2024

Abstract

This report provides an exploratory analysis of the Gas Turbine CO and NOx Emission Dataset, focusing on statistical analysis, data visualization, and dimensionality reduction using Principal Component Analysis (PCA), as part of a preliminary understanding before further machine learning modeling. The dataset comprises attributes related to environmental conditions, turbine operating parameters, and mechanical pressure control, which influence emissions. Using visualizations and statistical summaries, key relationships and data characteristics were identified, such as correlations among turbine parameters and the distribution of emission levels. PCA was applied to reduce dimensionality and interpret the contributions of different features, demonstrating that the first four components capture most of the variance in the data. These findings are the basis for future classification and regression tasks, aiming to predict emission levels based on operating conditions.

Contribution

Table 1: Contribution

	Jonas Wiendl	Linnea Hockauf	Rami I. Hanna
Description of dataset	25	25	50
Explanation of attributes	25	50	25
Data Visualization & PCA Analysis	50	25	25
Discussion	25	50	25
Exam Problems	33	33	33

Contents

1 Description of dataset

The dataset under analysis focuses on predicting the turbine energy yield (TEY) of gas turbines, utilizing 11 sensor measurements. These sensors capture various environmental and operational attributes such as ambient temperature, pressure, humidity, and turbine inlet and outlet temperatures. In addition to the energy yield, the dataset also allows for the prediction of emissions, specifically carbon monoxide (CO) and nitrogen oxides (NOx), which are critical to understanding the environmental impact of turbine operation.

This dataset was compiled by Heysem Kaya, Pınar Tüfekçi, and Erdinç Uzun from Namık Kemal University in Turkey. The researchers aimed to advance the development of Predictive Emission Monitoring Systems (PEMS) to contribute to environmental sustainability. As the global demand for energy grows, so do concerns over its environmental impacts, such as deforestation and increased gas emissions. Therefore, this dataset serves as a valuable tool for addressing these issues. The dataset, the reference paper, and more details can be accessed online [1, 2].

Previous analyses of the dataset revealed that feature selection based on linear projection weights significantly improves prediction performance, particularly for CO emissions. An analysis of the feature importance weights showed that turbine energy yield (TEY) and compressor discharge pressure (CDP) were the most critical variables for predicting both CO and NOx emissions. Interestingly, the predictive performance for CO was better than for NOx, likely due to stronger correlations within the data distribution [1]. This insight offers a valuable direction for future work on emissions prediction and turbine optimization.

Looking ahead to classification and regression tasks, the goal is to further explore the predictive capabilities of the dataset. In the context of classification, the aim is to categorize turbine emissions into discrete levels, such as high/low emissions of CO, based on key variables like fuel flow, ambient temperature, and turbine load. This could help classify whether emissions exceed regulatory limits. For regression, the primary focus will be on predicting continuous emission levels, specifically NOx concentrations, using attributes such as ambient temperature, humidity, pressure, fuel flow, and turbine output.

To achieve these tasks, data transformation, particularly for classification of CO is required. For the regression task, scaling the continuous variables (e.g., ambient temperature, pressure) will likely improve model performance. Given the environmental and

operational nature of the dataset, regression is expected to be the main machine learning aim, as it provides the most direct method for predicting emission concentrations under varying turbine conditions. By visualizing and analyzing the data, the feasibility of these tasks will become clearer, allowing for more precise models that can inform both energy efficiency and environmental preservation.

Upon inspecting the dataset, no significant data issues, such as missing or corrupted values, have been reported in the accompanying documentation. Initial inspection also reveals a clean dataset without apparent anomalies, providing a solid foundation for both classification and regression tasks. However, further exploration may reveal potential areas for preprocessing, such as normalizing certain attributes to ensure smoother model performance.

2 Explanation of attributes

The dataset includes readings from 11 different sensors on gas turbines. These were collected over a period of 5 years, aimed to identify the important features contributing to CO and NOx emissions. **Table 2** shows the 11 different attributes along with their descriptions, abbreviations, units and data types [1]. Most attributes have a true meaningful zero value, which categorizes their data types as continuous ratios. However, for the temperature attributes, there is no true zero; therefore, these are classified as continuous intervals. Additionally, the dataset contains no missing values.

Variable	**Abbr.**	**Unit**	**Type**	**Role**
Ambient temperature	AT	∘C	Continuous (Interval)	Feature
Ambient pressure	AP	mbar	Continuous (Ratio)	Feature
Ambient humidity	AH	%	Continuous (Ratio)	Feature
Air filter difference pressure	AFDP	mbar	Continuous (Ratio)	Feature
Gas turbine exhaust pressure	GTEP	mbar	Continuous (Ratio)	Feature
Turbine inlet temperature	TIT	∘C	Continuous (Interval)	Feature
Turbine after temperature	TAT	∘C	Continuous (Interval)	Feature
Turbine energy yield	TEY	MWH	Continuous (Ratio)	Feature
Compressor discharge pressure	CDP	mbar	Continuous (Ratio)	Feature
Carbon monoxide	CO	mg/m^3	Continuous (Ratio)	Target
Nitrogen oxides	NOx	mg/m^3	Continuous (Ratio)	Target

Table 2: Gas Turbine Emission Variables

To gain a overview of dataset's the scale and distribution, the summary statistics were examined first. **Table 3** presents the mean, standard deviation, median, minimum and maximum values for each feature.

Feature Name	**Mean**	**Standard Deviation**	**Median**	**Min**	**Max**
AT	17.71	7.447	17.80	-6.2348	37.10
AP	1013	6.463	1012	985.85	1036
AH	77.87	14.46	80.47	24.09	100.2
AFDP	3.926	0.774	3.938	2.087	7.611
GTEP	25.56	4.196	25.10	17.70	40.72
TIT	1081	17.54	1086	1000	1101
TAT	546.2	6.842	549.9	511.0	550.6
TEY	133.5	15.62	133.7	100.0	179.5
CDP	12.06	1.089	11.97	9.852	15.16
CO	2.372	2.263	1.714	0.0004	44.10
NOX	65.29	11.68	63.85	25.91	119.9

Table 3: Summary Statistics (Rounded to 4 Significant Figures)

Investigating **Table 3** highlights differences in magnitude of attribute value. For example, ambient pressure (AP) and turbine intlet temperature (TIT) are several orders of magnitude larger than air filter difference pressure (AFDP) and compressor discharge pressure (CDP). These differences can introduce bias towards large value attributes when applying machine learning to the dataset. One way to mitigate this effect is through standardization of the data, which involves subtracting the mean and dividing by the standard deviation for each attribute. The result is a zero mean and standard deviation of 1, ensuring consistent scaling across units.

3 Data Visulization and PCA analysis

To better understand the data, the following subsections provide data visualizations and describe the implementation of Principal Component Analysis (PCA).

3.1 Data Viszualizations

To start, it is useful to look at the distribution of both attributes and target variables which are shown in **Figure 1**. The attributes can be categorized into three groups, as follows: First, environmental conditions, such as temperature, pressure, and humidity, reflect external factors impacting turbine performance. These attributes show near-normal distributions, suggesting that the turbine operates under typical variations in external conditions that affect the surrounding environment. Secondly, turbine operating parameters, such as inlet and after temperatures, represent internal operational states that fluctuate with different load levels and energy demands. The wider distributions of these attributes indicate that the turbine has been operated across a range of load conditions, leading to variations in parameters like Turbine Energy Yield (TEY) and Turbine Inlet Temperature (TIT).

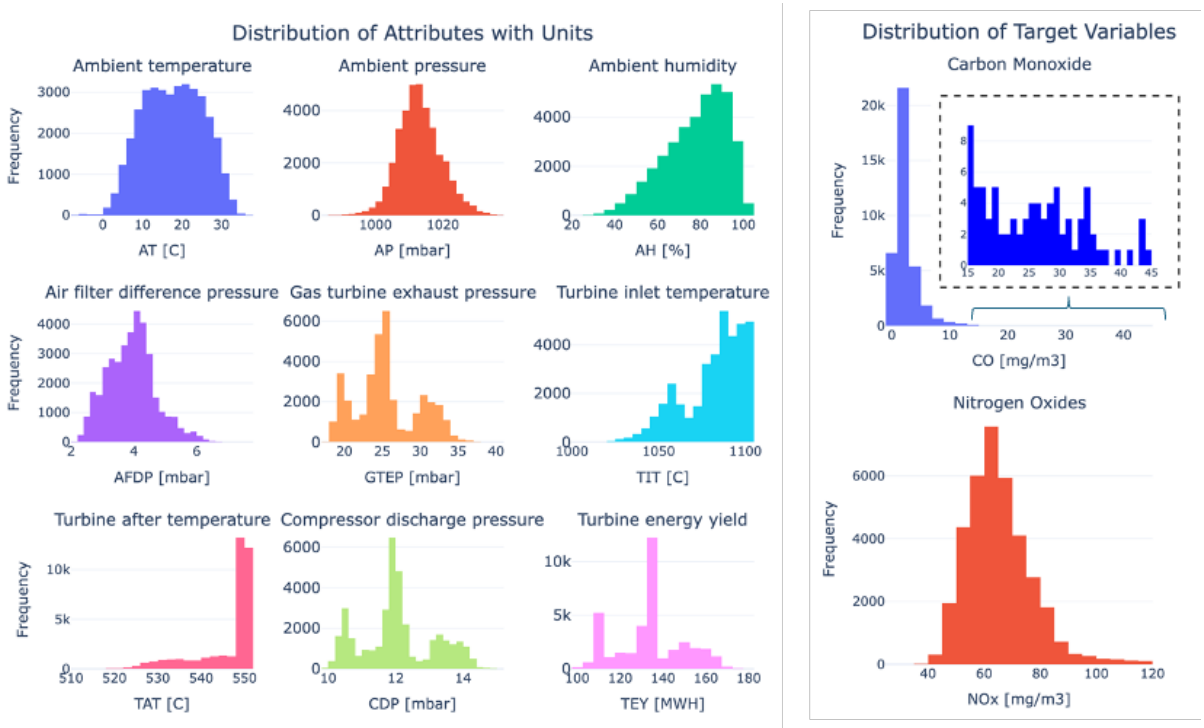

Figure 1: Distribution of Attribute and Target Variables

Lastly, the mechanical pressure control group includes variables related to the tur-

bine's air and pressure management systems. Compressor Discharge Pressure (CDP) shows a stable distribution, reflecting its consistent operation within a specific range. In contrast, Air Filter Difference Pressure (AFDP) displays more variability, likely due to operational changes in airflow or filter conditions.

The target variables, Carbon Monoxide (CO) and Nitrogen Oxides (NOx), represent the emissions produced by the turbine during operation. CO has a highly skewed distribution, with most values clustered within the range 0-8 mg/m^3. This suggests that under normal operating conditions, the turbine produces minimal CO emissions, with occasional spikes due to inefficient combustion or specific load conditions. On the other hand, NOx, which is formed at high combustion temperatures, follows a more normal distribution centered around 60-70 mg/m^3. Based on these observations, both regression for NOx levels and classification for categorized CO levels are feasible.

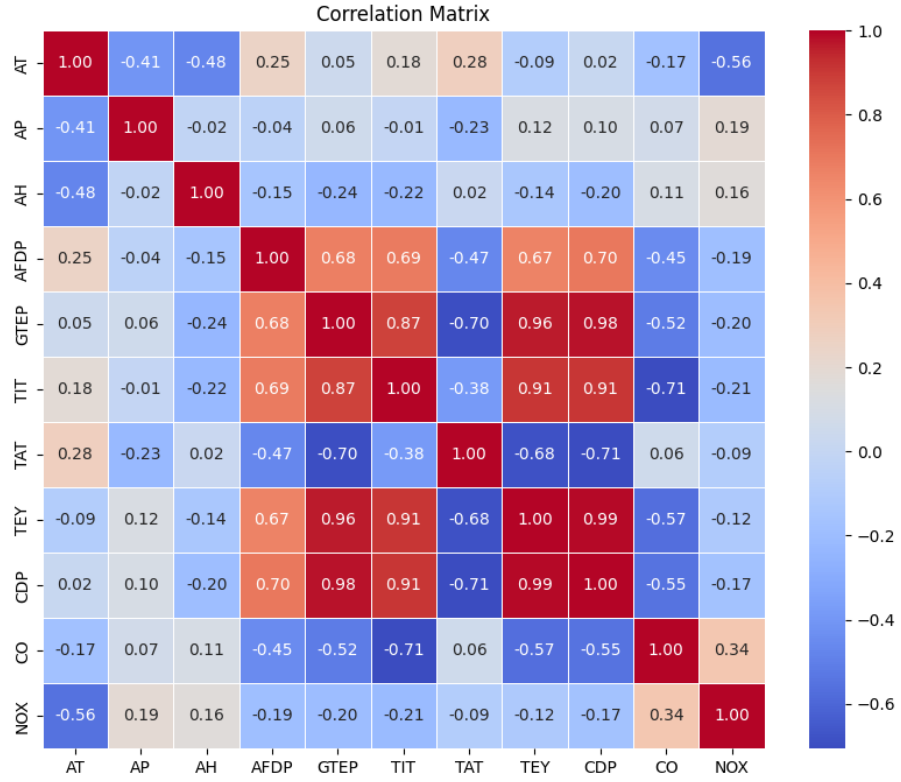

Figure 2: Correlation Matrix of Attribute and Target Variables

Besides the distribution, it is useful to examine the correlation matrix shown in **Figure 2**. It reveals key relationships between turbine operating parameters and emissions. Notably, Turbine Inlet Temperature (TIT) and Turbine Energy Yield (TEY) show strong positive correlations with each other and other performance metrics, while exhibiting

moderate negative correlations with CO emissions, indicating that higher temperatures and energy outputs lead to lower CO levels due to more efficient combustion. In contrast, NOx emissions are positively correlated with TIT and CO, reflecting the increase in NOx production at higher combustion temperatures. Additionally, Ambient Temperature (AT) shows a moderate negative correlation with both emissions, suggesting that lower environmental temperatures may result in higher CO and NOx outputs.

3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was conducted after standardizing the data, which is crucial for fair comparison across attributes with different scales, ensuring that each attribute has a mean of zero and a standard deviation of one. The explained variance plot, (**Figure 3 left**), shows that the first four principal components capture around 90% of the total variance, making them sufficient for further analysis.

The PCA coefficient plot, (**Figure 3 right**), illustrates how each original attribute contributes to the principal components. The environmental conditions (AT, AP, AH), for instance, strongly influence the second principal components. In contrast, the first component is influenced similarly by turbine operating parameters and mechanical pressure control parameters.

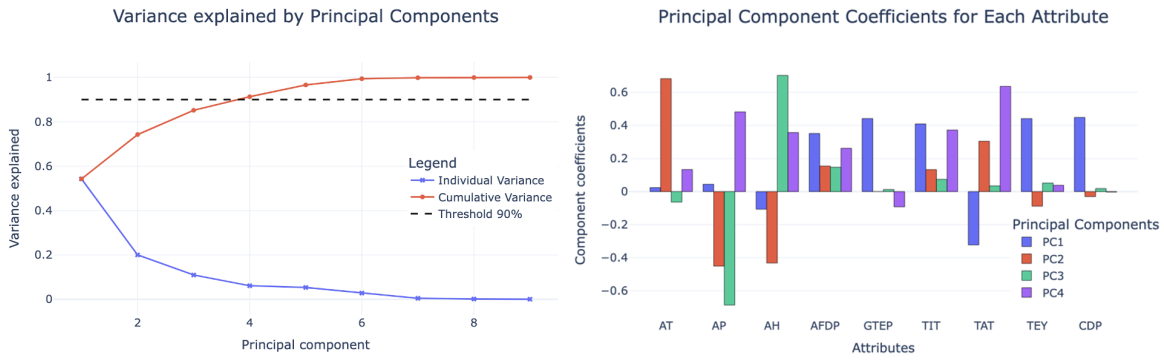

Figure 3: Explained Variance and Component Coefficients

Pairwise PCA scatter plots (**Figure 4**) and (**Figure 5**) show the relationships between the first four principal components, color-coded by NOx and CO values. These plots reveal how emissions are distributed along the principal components. In both plots clusters of different levels of emissions can be identified.

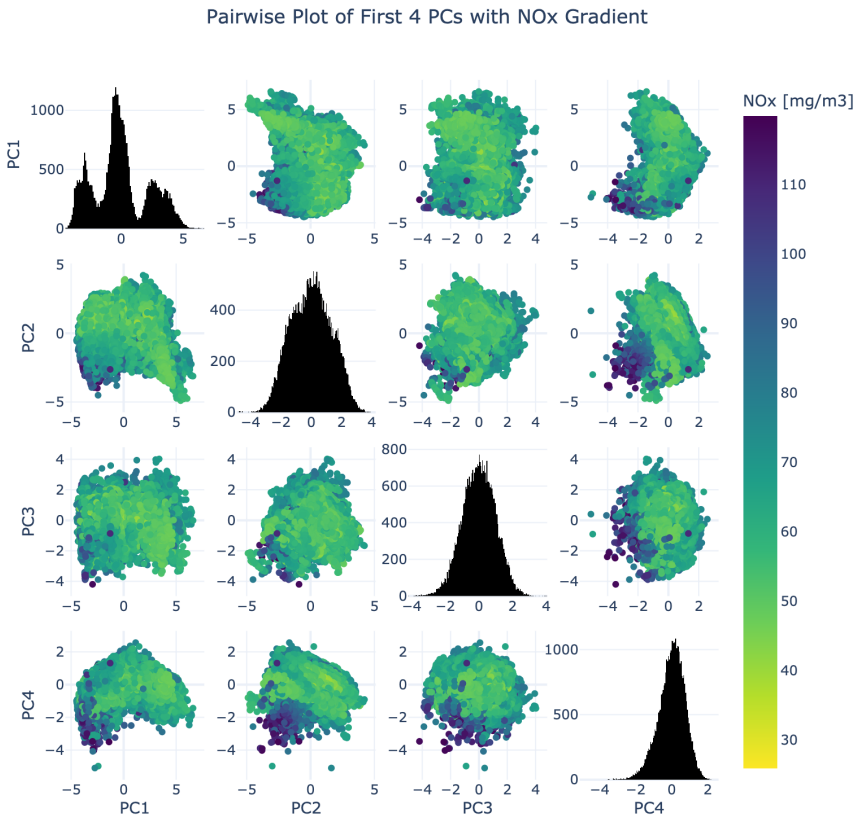

Figure 4: Pairwise Comparison of PCA components - NOx

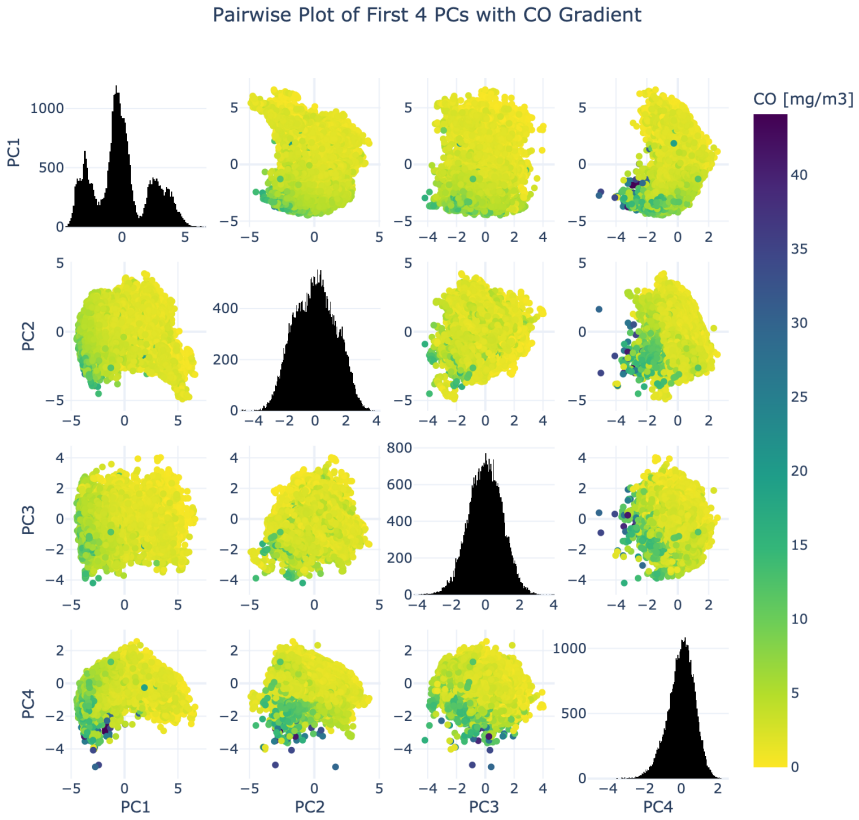

Figure 5: Pairwise Comparison of PCA components - CO

4 Discussion

In **Figure 5**, PC1 clearly separates CO levels with a distinct color gradient, especially alongside PC2, compared to the other PCs, which exhibits less variance. The NOx levels plot in **Figure 4** reveals a similar pattern, where PC1 and PC2, captures most of the variance, although some clustering of high NOx values are also visible for PC3 and PC4. This suggests that both pollutants are primarily driven by PC1, while PC3 and PC4 show more scattered relationships. This is also consistent with the results shown in the accumulative variance in **Figure 3**. With that said, though, it is important to note that the PC1 versus PC4 and PC2 versus PC4 graphs show a more even distribution of variance throughout the plot rather than just a centered oval shape. Looking again at **Figure 4** and **Figure 5**, it can be concluded that PC4's influence on the target variables is still to be considered. The tendency of higher CO and NOx levels to cluster at the plot edges indicates potential outliers, which might reflect unusual operational or environmental conditions.

The attribute coefficients in **Figure 3** reveals how PC1 seems to be dominated by the operational performance of the gas turbine system, indicated by the positive loadings on pressure (GTEP), energy yield (TEY), and inlet temperature (TIT). The negative effect from the turbine after temperature (TAT) could indicate that the higher after-temperatures have a negative impact on combustion efficiency. PC2, in contrast captures the environmental influences, the ambient conditions, suggesting that the PC2 represents the variance from external conditions which can affect the turbine performance.
Combining this, indicates that PC1 reflects the turbine efficiency, with increasing efficiency resulting in lower emissions due to the more complete combustion, looking at **Figure 4 & 5**. PC2 has a lower impact, but captures the environmental factors and demonstrates how these external factors as high temperature, low pressure and humidity additionally can influence the combustion efficiency positively by changing the air density or dynamics.

The data visualization and PCA analysis reveal strong relationships between attributes, while the PCA plots show a good distribution, and the cumulative variance plot indicate feasible dimensionality reduction. Together, these factors suggest that the data is well-suited for our machine learning aims, both for classification and regression tasks.

5 Exam Problems

5.1 Question 1

Considering each type of attribute as the highest level it obtains the the type-hirachy we obtain the following table of attribute types:

Table 4: Attributes Description

No.	Attribute Description	Abbreviation	Attribute Type
x_1	30-minute interval (coded)	Time of day	Ordinal
x_2	Number of broken trucks	Broken Truck	Ratio
x_3	Number of accident victims	Accident victim	Ratio
x_4	Number of immobile buses	Immobilized bus	Ratio
x_5	Number of trolleybus network defects	Defects	Ratio
x_6	Number of broken traffic lights	Traffic lights	Ratio
x_7	Number of run-over accidents	Running over	Ratio
y	Level of congestion/slowdown (low to high)	Congestion level	Ordinal

From **Table 3** it is given that **C)** is correct. *Time of day* and *Congestion level* are ordinal while *Traffic lights* and *Running over* are ratio.

5.2 Question 2

The p-norm distance between x_{14} and x_{18} for the max-norm distance $p{=}\infty$

```
import numpy as np
# Given vectors
x14 = np.array([26, 0, 2, 0, 0, 0, 0])
x18 = np.array([19, 0, 0, 0, 0, 0, 0])
d = x14 - x18 # Finding the distance between x1-y1, x2-y2..
n = np.linalg.norm(d, np.inf) # Finding the maximum absolute distance between
    the 2 vectors
print(n)
# Output 7.0
```

Therefore option **A)** $d_{p=\infty}(x_{14},x_{18}) = 7.0$ is correct

5.3 Question 3

From the diagonal of matrix $\boldsymbol{S}$ we have the singular values of the PCA.

```
import numpy as np
# Singular values
sigma=np.array([13.9,12.47,11.48,10.03,9.45])
var=sigma**2 # Variance (sigma^2)
# For each option
```

```
A=sum(var[0:4])
B=sum(var[2:5])
C=sum(var[0:2])
D=sum(var[0:3])
# Printing each answer, finding the variance proportion
print("First four :",(A/sum(var)))
print("Last three :",(B/sum(var)))
print("First two :",(C/sum(var)))
print("First three :",(D/sum(var)))
#Output
# First four : 0.8667931474824088
# Last three : 0.47985015618178095
# First two : 0.520149843818219
# First three : 0.7167331911605035
```

From this, option **A)** is correct, as the variance explained by the first four principal components are greater than 0.8

5.4 Question 4

Looking at the second column (PC2), of the matrix V, we have:

- Time of Day: low value giving a positive contribution
- Broken Truck: high value giving a positive contribution
- Accident Victim: high value giving a positive contribution
- Defects: high value contribution

Therefore, option D. is correct.

6 References

[1] Kaya, H., Tüfekci, P., & Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(6), 4783–4796. https://doi.org/10.3906/ELK-1807-87

[2] Gas Turbine CO and NOx Emission Data Set [DOI: https://doi.org/10.24432/C5WC95]. (2019).