Fachhochschule Südwestfalen
University of Applied Sciences

Machine Learning

Project Task 2 Report
within the context of the Machine Learning module
Summer Semester 2024

Professors:	Prof. Dr. Andreas De Vries Prof. Dr. Stefan Böcker
Submitted by:	Seyedehsadaf Edalat Matriculation number: 30372928
	Aziz Fahmi Matriculation number: 30461097
Deadline:	2024-07-08

Wir geben Impulse

Table of Contents

List of Figures 2
List of Abbreviations 3
1 Description of the Problem 4
2 Description of the Solution 5
  2.1 Data Collection and Preprocessing 5
  2.2 Principal Component Analysis 5
  2.3 Linear Regression Model 6
3 Results and Discussions 7
  3.1 Principal Component Analysis Results 7
  3.2 Essential Variables and Feature Importance 8
  3.3 Linear Regression Performance 9
  3.4 Assumptions and Limitations 10
4 Conclusion 11
Bibliography 12
Appendix 13

List of Figures

Figure 1: Source Code and Scree Plot 6
Figure 2: 3D Scatter Plot with Regression Plane 7
Figure 3: Biplot 8
Figure 4: Heatmap 9

List of Abbreviations

MSE Mean Squared Error

PCA Principal Component Analysis

1 Description of the Problem

Our project's goal is to evaluate and enhance an apartment building's solar-thermal water heating system, which supports space heating through a hot water tank. By using solar energy, the system heats a buffer tank that provides heat for space heating and water heating. A gas condensing boiler is used to complement the heating in cases where solar energy is insufficient.

We concentrate on examining some measured factors (x) and their effects on gas consumption (y) in order to comprehend and enhance the performance of this system. These variables include sunshine duration, outdoor temperature, solar yield, solar pump operation times, and the usage of a 3-way valve that manages the heating distribution. Weekly averages of these variables have been recorded and are kept in a CSV file. Specifically, we aim to address the following questions:

- Principal Component Analysis (PCA): How many principal components of the measured variables should be considered to effectively predict gas consumption?
- Essential Variables: What are the key variables that significantly impact gas consumption in this heating system?
- Linear Regression: Can a linear regression model using the principal components of the measured variables accurately predict gas consumption ?

We will use PCA to reduce the dimensionality of the data and find the most important components in order to accomplish these goals. The goal variable in our linear regression model will be gas consumption, which will be constructed using these components. Our goal in analysing the data was to evaluate the effectiveness of the heating system. The analysis aimed to provide insights into its operational performance.

2 Description of the Solution

PCA and Linear Regression are two structured machine learning approaches that we will use to analyse and optimise the apartment complex's solar-thermal heating system. Here, we go into detail about the approach and actions done to meet our goals.

2.1 Data Collection and Preprocessing

Functions were called in order to execute the intended tasks and load the necessary modules and libraries. Data visualisation and analysis are done with libraries like matplotlib, numpy, and pandas. Moreover, the Linear Regression, PCA, and Standard Scaler. They are all accomplished using modules from the scikit-learn toolkit. Data from the CSV file is imported by the code, which then gets it ready for additional scikit-learn analysis. The "Date" column is eliminated since it is not required for the analysis and features are established. (see appendix 1)

2.2 Principal Component Analysis

Large datasets are increasingly common across various fields. To effectively interpret them, methods that reduce their dimensionality while retaining key information are crucial. PCA is widely used for this purpose. It simplifies datasets by transforming original variables into a new set of uncorrelated components that capture the maximum variance [1].

Steps in PCA:

- Standardisation: Guaranteed that every variable made an equal contribution to the analysis.
- Calculate the Covariance Matrix: To see how variables relate to one another.
- Computation of Eigenvalue and Eigenvector: Resolute in determining the primary components.
- Choosing the Number of Components: A minimum of 90% of the cumulative variance may be explained by the number of principal components that were chosen. By examining the explained variance ratio plot, or scree plot, this threshold was established [2].

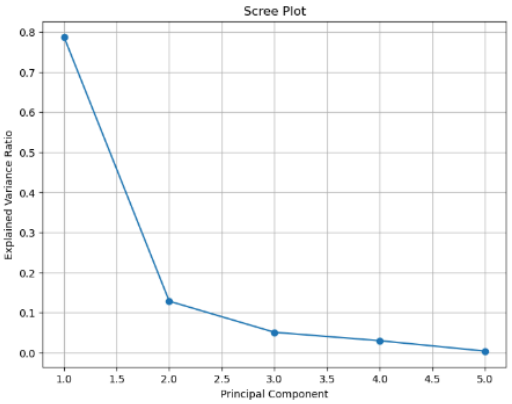
FIGURE 1: SOURCE CODE AND SCREE PLOT

```
# Standardize numerical features
scaler = StandardScaler()
encoded_data[numerical_cols] = scaler.fit_transform(encoded_data[numerical_cols])

# Split the data into features (X) and target variable (y)
X = encoded_data.drop('Gas consumption [kWh/day]', axis=1)
y = encoded_data['Gas consumption [kWh/day]']

# Perform PCA
pca = PCA()
X_pca = pca.fit_transform(X)
explained_variance_ratio = pca.explained_variance_ratio_
```

```
# Scree plot of the possible five principal components
plt.figure(figsize=(8, 6))
plt.plot(range(1, 6), pca.explained_variance_ratio_[:5], marker='o', linestyle='-')
plt.xlabel('Principal Component')
plt.ylabel('Explained Variance Ratio')
plt.title('Scree Plot of the Possible Five Principal Components')
plt.show()
```



2.3 Linear Regression Model

Next, we constructed a linear regression model with the major components found by PCA to forecast gas use. In addition, a regression plane is drawn to show the relationship between the first two principal components and the target value.

Steps in linear regression:

- Data Splitting: To assess the model's performance on unobserved data, the dataset was split into training and testing sets.
- Model Training: Using the training data, a linear regression model was fitted with the goal variable being gas consumption and the principle components acting as predictors.
- Model Evaluation: To determine the accuracy and goodness of fit of the model, its performance was assessed using Mean Squared Error (MSE) and R-squared (R^2) metrics on the test set [3]. (see appendix 2)

3 Results and Discussions

3.1 Principal Component Analysis Results

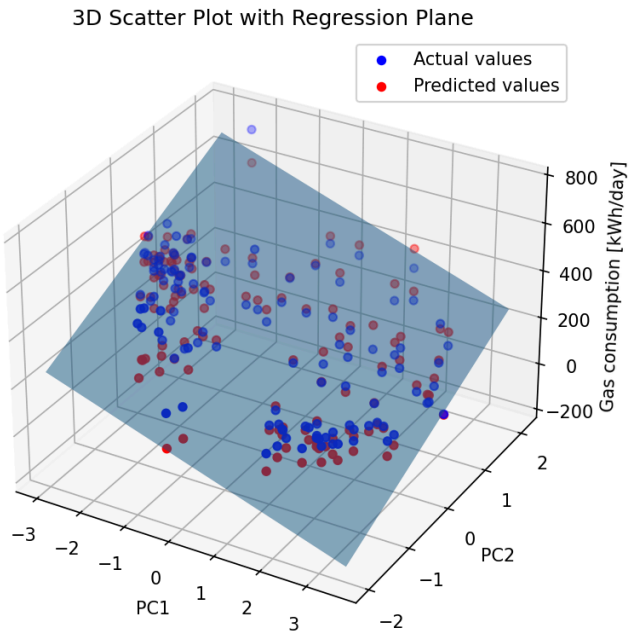
The PCA was used to extract the most important information from the dataset while reducing its dimensionality. The following conclusions were drawn from the analysis:

Explained Variance: The first two principal components explained approximately 91.58% of the total variance in the dataset, with the first component alone explaining 78.75%.

Optimal Number of Components: Based on the cumulative explained variance, and also by considering the scree plot (figure 1), we determined that two principal components were sufficient to capture over 90% of the variance in the data. (see appendix 3)

The 3D scatter plot of the first two principal components provides a visual representation of the data in reduced dimensions. Every point on the three-dimensional scatter plot denotes a dataset sample. The scores a point receives on PC1, PC2, and the target variable determine its location. The relationship between the principal components (PC1 and PC2) and the target variable can be seen more clearly with the use of the regression plane. The target variable will exhibit a distinct upward or downward trend on the plane if it grows linearly with the primary components [3]:

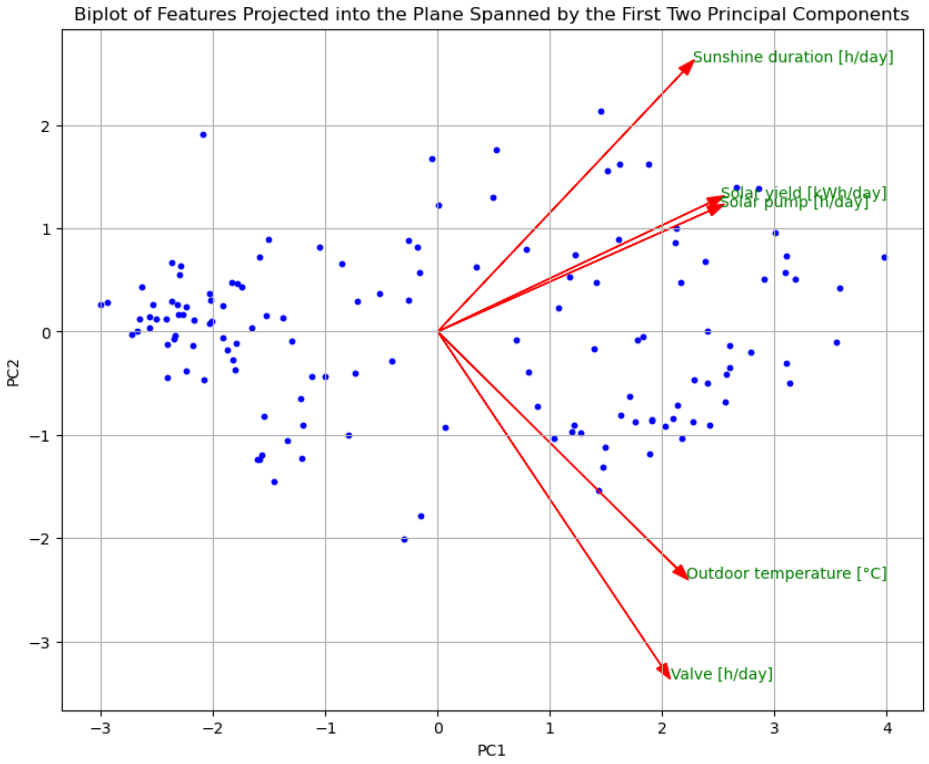
FIGURE 2: 3D SCATTER PLOT WITH REGRESSION PLANE



To provide a deeper understanding of the PCA results, we included a biplot. The biplot combines the principal components' scores and loadings in a single plot, showing how the original variables relate to the principal components and how they influence the data distribution. The points on the plot represent the samples or observations in the dataset, projected onto the new principal component space. Each sample's principal component representation is indicated by the position of each point and points that are close to each other in the biplot are similar to each other in the original feature space. The loadings, which display the magnitude and direction of each original variable in the main component space, are represented by the arrows in the biplot. An arrow's direction points in the direction that the original variable contributed the most.

The degree to which the original variable contributed to the major components is indicated by the length of an arrow. Positive correlation exists between arrows heading in the same direction and negative correlation between arrows pointing in opposite directions [4].

FIGURE 3: BIPLOT

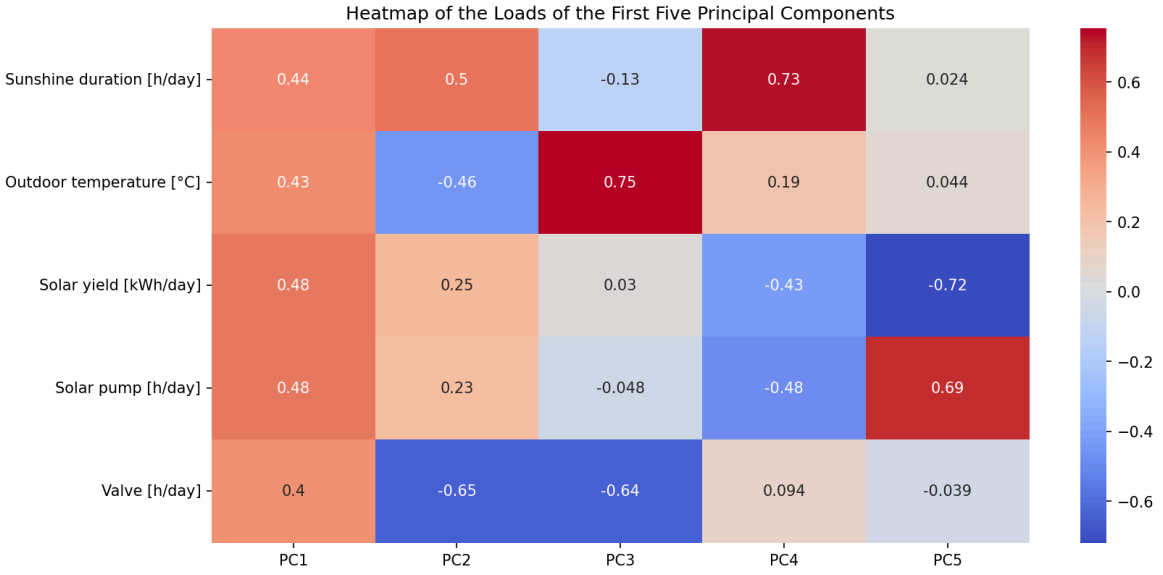


3.2 Essential Variables and Feature Importance

To illustrate the contribution of each variable to the principal components, we created a heatmap of PCA loadings. This heatmap provides a clear visual representation of how each variable correlates with the principal components, helping us identify the most

influential variables [3]. It shows the loadings (coefficients) of each feature for the principal components. The colour intensity represents the magnitude of the loadings: darker colours indicate higher contributions. With the sum of loadings in PC1 and PC2 we realise that all variables play crucial roles in gas consumption.

FIGURE 4: HEATMAP



However, taking a look at Biplot (figure 3) we can conclude that for PC1 solar yield (as well as solar pump) is an essential variable, and for PC2, regardless of the direction, valve explains most of the variety in gas consumption (A negative loading simply means that the feature is inversely related to the principal component, but it does not reduce the importance or contribution of the feature to that principal component. A feature's contribution to a principal component is determined by the magnitude of its loading, regardless of whether the loading is positive or negative).

3.3 Linear Regression Performance

The linear regression model on the original data shows strong performance, with a high R-squared value of 0.91 on the test data, indicating that 91% of the variance in gas consumption can be explained by the model. The model also exhibits a low mean squared error (0.10 on the test data), suggesting good predictive accuracy. The training data shows slightly better performance (R^2 of 0.92 and MSE of 0.07), indicating that the model fits the training data very well.

The model on the PCA-transformed data (using the first two principal components) also demonstrates strong performance, albeit slightly less effective than the model on the

original data. The R-squared value of 0.87 on the test data indicates that 87% of the variance in gas consumption is explained by the first two principal components. The mean squared error is slightly higher (0.15 on the test data) compared to the original data model, reflecting a minor reduction in predictive accuracy. The training data results are consistent, with an R^2 of 0.89 and an MSE of 0.10, indicating good model fitting. (see appendix 4)

3.4 Assumptions and Limitations

In this project, some key assumptions were made to facilitate the analysis and modelling process. It was assumed that the historical gas consumption data provided was accurate and representative of typical usage patterns. Furthermore, it was assumed that linear relationships between the features and gas consumption were sufficient for capturing the underlying patterns, justifying the use of linear regression models. Lastly, PCA was assumed to effectively reduce dimensionality and retain the most significant variance within the dataset.

The study also faced some limitations that may impact the generalizability and accuracy of the results. The dataset's scope was limited to specific temporal and geographic contexts, which may not reflect broader or more diverse conditions. The linear regression model, while simple and interpretable, may not capture complex, non-linear relationships inherent in gas consumption data. Additionally, PCA, despite reducing dimensionality, may lose some nuanced information by focusing only on the components with the highest variance. External factors influencing gas consumption, such as sudden weather changes or policy shifts, were not explicitly accounted for in the model. Lastly, the assumptions made during data preprocessing and modelling, such as the linearity assumption, may not fully hold in real-world scenarios, potentially affecting the model's predictive performance.

4 Conclusion

The objective of this project was to assess and improve the solar-thermal water heating system of an apartment building, which provides space heating via a hot water tank. Our goal was to maximise the efficiency of the system by examining the effects of several observed factors on gas consumption. After dimensionality-reducing our dataset, we found two principal components that accounted for more than 90% of the variance. We developed a linear regression model using the principal components, achieving a high R-squared value of 0.87, indicating a strong fit. Our analysis also revealed that all five original variables are essential in predicting gas consumption, with valve, solar yield and solar pump being particularly influential in the first two principal components.

Ultimately, our findings suggest that optimising the use of solar energy and the operational efficiency of the solar pump, yield and valve can significantly reduce gas consumption. These insights offer valuable recommendations for improving the heating system's performance, potentially leading to lower energy costs and greater savings for the apartment complex owners.

Bibliography

[1] Jolliffe, I.T., & Cadima, J. (2016). *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374. https://doi.org/10.1098/rsta.2015.0202

[2] Kelta, Z (2023, February): *Principal Component Analysis in R Tutorial* https://www.datacamp.com/tutorial/pca-analysis-r?dc_referrer=https%3A%2F%2Fwww.google.com%2F

[3] Dr. De Vries, A. (2024): *Lecture notes on Machine Learning*

[4] Wicklin, R (2019, November 6): *What are biplots?* https://blogs.sas.com/content/iml/2019/11/06/what-are-biplots.html

Appendix

1) Source code for data processing

```
# Convert Date column to datetime and extract useful features
data['Date'] = pd.to_datetime(data['Date'])
data = data.drop(columns=['Date'])

# Explore the data and handle missing values
print(data.head())
print(data.describe())
data = data.dropna()

# Identify and handle non-numeric columns (if any)
categorical_cols = data.select_dtypes(include=['object']).columns
numerical_cols = data.select_dtypes(exclude=['object']).columns

# Encode categorical data (if any)
encoders = {}
encoded_data = data.copy()
for col in categorical_cols:
    encoder = LabelEncoder()
    encoded_data[col] = encoder.fit_transform(data[col])
    encoders[col] = encoder

# Standardize numerical features
scaler = StandardScaler()
encoded_data[numerical_cols] = scaler.fit_transform(encoded_data[numerical_cols])
```

2) Linear regression source code

```
# Select the first two principal components
X_pca_two_components = X_pca[:, :2]

# Split the data into training and testing sets using the first two principal components
X_train, X_test, y_train, y_test = train_test_split(X_pca_two_components, y, test_size=0.2, random_state=42)

# Fit a linear regression model
lr = LinearRegression()
lr.fit(X_train, y_train)

# Evaluate the model's performance
y_pred = lr.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
model_score = lr.score(X_test, y_test)
```

3) Explained variance ratio by components

```
Explained Variance Ratio by Components:
[0.78751883 0.12825661 0.05066179 0.02978784 0.00377493]
```

```
Optimal number of components to explain at least 90% variance: 2
Raw PCA Components:
[[ 0.43731676  0.42623536  0.48404288  0.48365388  0.39857105]
 [ 0.50345736 -0.45889674  0.2491928   0.23345292 -0.64757019]
 [-0.13053856  0.75388262  0.03021625 -0.04801942 -0.64140611]
 [ 0.73324962  0.19364101 -0.43059337 -0.48020337  0.09403311]
 [ 0.02427787  0.04360455 -0.71922303  0.69186525 -0.03936925]]
```

4) Conclusion of the code

```
Test Data - Mean Squared Error (First Two Components): 0.15
Test Data - R-squared (First Two Components): 0.87
Test Data - Model Score (R², First Two Components): 0.87
Coefficients (First Two Components): [-0.44631241  0.36994004]
Training Data - Mean Squared Error (First Two Components): 0.10
Training Data - R-squared (First Two Components): 0.89
```

```
Original Data - Test Data - Mean Squared Error: 0.10, R-squared: 0.91, Model Score (R²): 0.91
Original Data - Training Data - Mean Squared Error: 0.07
Original Data - Training Data - R-squared: 0.92
```