Report for Machine Learning

University of Amsterdam

(6012B0809Y)

Assignment III

Exploring and Forecasting Stress Levels in Dutch Regions through Machine Learning Approaches.

December 2023

Introduction

In a quest to understand and mitigate the complexities surrounding stress levels[1] among the citizens of the Netherlands, this project embarks on a comprehensive exploration of the myriad factors influencing mental well-being. Stress, a pervasive element of modern life, manifests in diverse ways, and our endeavor seeks to unravel its nuanced connections with various demographic, social, and environmental variables.

Our first objective is to identify and analyze the diverse factors contributing to stress levels among Dutch citizens. By meticulously examining demographic, social, and environmental indicators, we aim to uncover the unique mosaic of stress influencers in different regions of the Netherlands.

A cornerstone of our project involves the construction of regression models to quantitatively measure stress levels across various regions. These models, renowned for their efficacy in handling multicollinearity and feature selection, will be instrumental in distilling the intricate relationship between stress and its contributing factors.

As we delve into the data, our focus extends to the determination of the most influential features impacting stress levels. This feature importance analysis is crucial in identifying the key drivers of stress, providing actionable insights for policymakers and public health practitioners.

In our pursuit of a holistic understanding, we turn our attention to unsupervised learning techniques. By uncovering clusters and outliers within the dataset, we aim to identify regions with distinct stress profiles. Visualization tools will play a pivotal role in presenting these patterns, offering a clear and intuitive depiction of stress dynamics in the Netherlands.

Data Pre-Processing

Working with substantial and well-defined datasets is crucial for model efficacy. However, the challenge arises when large datasets harbor values that do not contribute meaningfully to the model. Consequently, preprocessing becomes imperative to enhance the model's performance.

Upon inspecting our dataset, it became evident that numerous missing values needed addressing. To rectify this, we employed an algorithm that systematically tallied and cleared NaN values in each column. Notably, columns featuring more than 1 and 300 NaN values were identical, signifying their lack of utility for the model.

In addition to NaN values, certain columns contained '.' characters, which we treated as missing values. Consequently, the following columns were eliminated: ['ind_wbi', 'p_stadsv', 'pst_mvp', 'pst_dekp', 'a_inkont', 'g_ink_po', 'g_ink_pi', 'p_ink_li', 'p_ink_hi', 'p_n_act', 'p_hh_li', 'p_hh_hi', 'p_hh_lkk', 'p_hh_osm', 'g_wodief', 'g_vernoo', 'g_gewsek'].

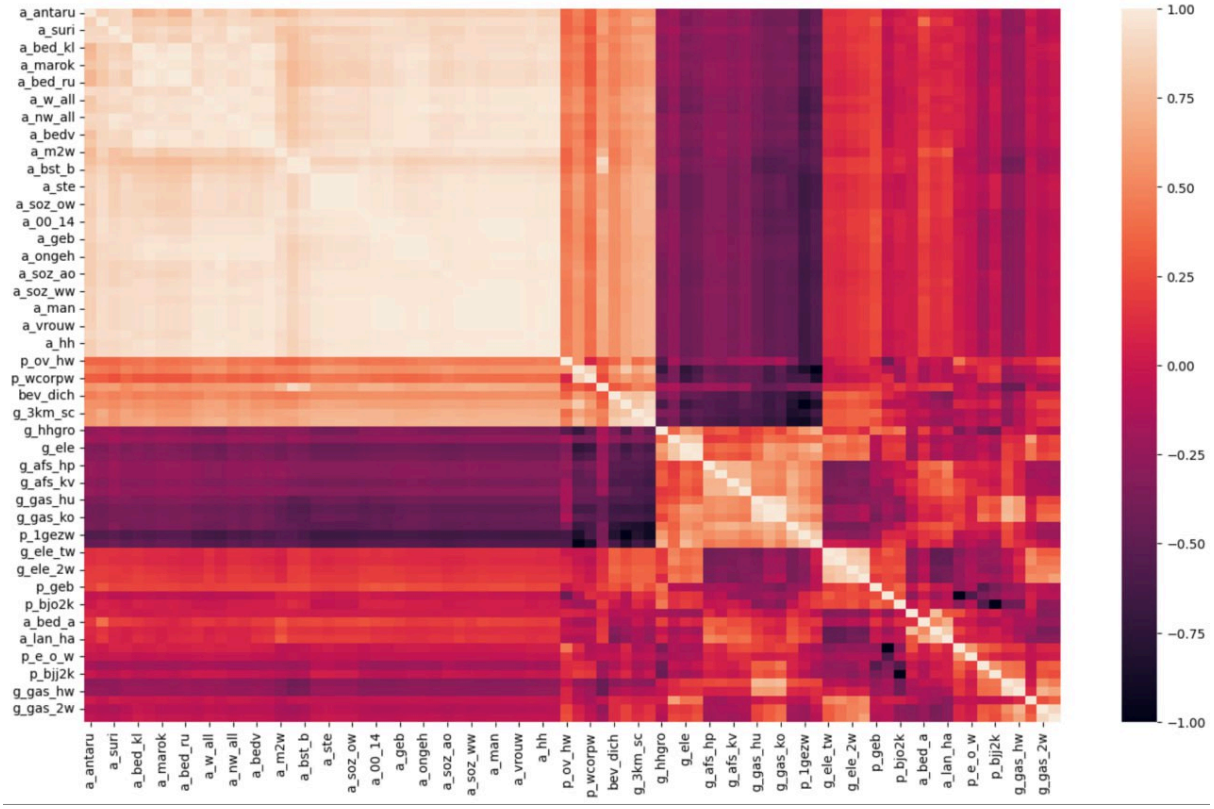
---

[1] A data set was taken from National Institute for Health and Environment: https://statline.rivm.nl/#/RIVM/nl/dataset/50077NED/table?ts=1702247247296

Furthermore, due to the merging of two datasets based on common values in 'gwb_code' and 'Regio' columns, we had to discard them to eliminate redundant data. Similarly, columns such as 'ID', 'recs,' etc., were identified as non-contributory to our model, as they primarily served to identify regions in the dataset. Consequently, we omitted these columns, streamlining the dataset for model efficiency.

During dataset examination, it came to our attention that certain columns ostensibly containing numerical values were, in fact, formatted as strings in Python. To ensure accurate data interpretation, we converted these string entries to float values. Notably, these regions, totaling 108, were designated as 'regions to predict.'

Post-cleanup, it became apparent that some regions lacked stress data. Intriguingly, we decided to leverage our model to predict stress values for these data-deficient regions, adding an additional layer of insight to our analysis.



(Fig. 1)

The correlation matrix reveals strong correlations, approaching values close to -1 and 1 among several features *(Fig 1)*. To refine our model, it's essential to implement feature selection. This process helps improve predictive accuracy by focusing on the most influential variables, reducing redundancy, and ensuring the model's efficiency in capturing patterns associated with stress levels.

Regression model

To identify the optimal model for our dataset, we experimented with various models, including Lasso Regression, Ordinary Linear Regression, Elastic Net, Ridge Regression, and Random Forest Regressor.

The performance scores for each model are as follows:

Method name/score	R^2	MSE	MAE
Ordinary Regression	`0.42002`	`12.71511`	`2.76800`
Lasso Regression	`0.41402`	`12.84672`	`2.75775`
Ridge Regression	`0.40863`	`12.96484`	`2.82206`
Elastic Net Regression	`0.38892`	`13.39688`	`2.82947`
Random Forest Regressor	`0.30926`	`15.14327`	`3.06269`

Highlighting a significant observation, the most favorable outcomes across all models were achieved through feature selection using the f-regression score. Recognizing the potential impact of scaling, we explored standard scaling and observed improved scores, with the standart scaler proving most effective.

Ultimately, the Ordinary Linear Regression model emerged as the top performer, attaining a score of 0.42 with the incorporation of five features. These features, pivotal for the model's success, are:

Feature	Coefficient
`'g_hhgro'`	`-1.85729584`
`'p_1gezw'`	`0.46097515`
`p_mgezw`	`-0.46097515`
`ste_mvs`	`-1.55427963`
`ste_oad`	`0.77665945`

The analysis reveals that the `'g_hhgro'` feature (Average household size) makes a substantial negative contribution to the model. This suggests that households with more members tend to have lower stress levels. On the contrary, the `'ste_oad'` feature (Environmental address density) indicates a positive association with stress levels. This feature reflects the concentration of human activities, suggesting that higher environmental address density corresponds to increased stress levels, likely due to heightened daily activities in the vicinity.

However, it's crucial to note that while these features contribute significantly, the overall score is not exceptionally high. Stress levels are influenced by a multitude of factors not captured in the current dataset. A larger dataset with more samples could potentially improve the model's predictive power by encompassing a more comprehensive range of influencing factors. Despite the limitations, the insights provided by 'g_hhgro' and 'ste_oad' contribute valuable information toward understanding the dynamics of stress levels in relation to household size and environmental address density.

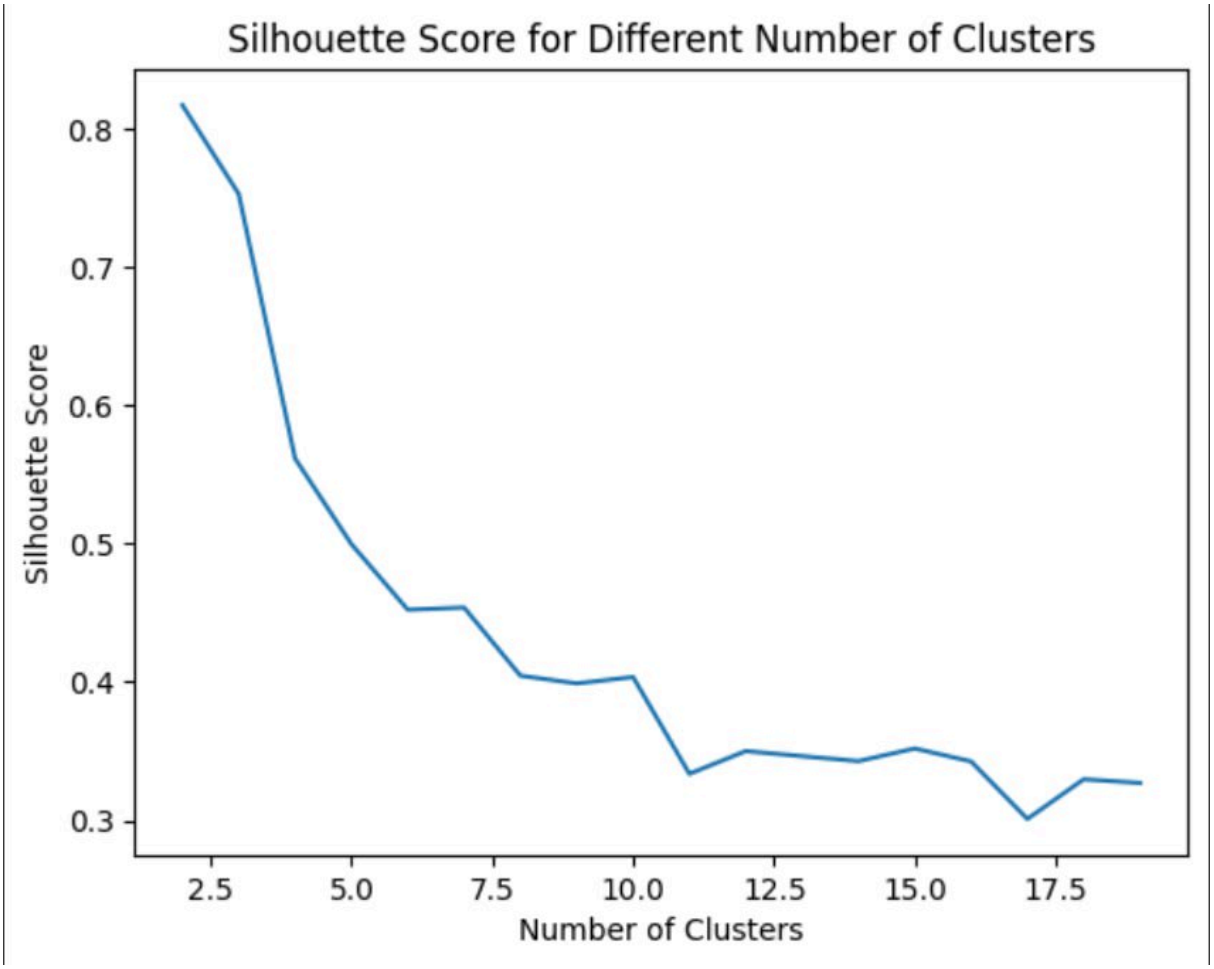
Predicting values

For the prediction phase, we utilized the Ordinary Linear Regression model, which emerged as the most effective for our dataset. The predictions for the 108 values we aimed to forecast are stored in a variable named 'regions_predicted.' These predictions provide insights into the anticipated stress levels for the regions under consideration, offering valuable information for further analysis and decision-making in the context of stress management and public well-being.

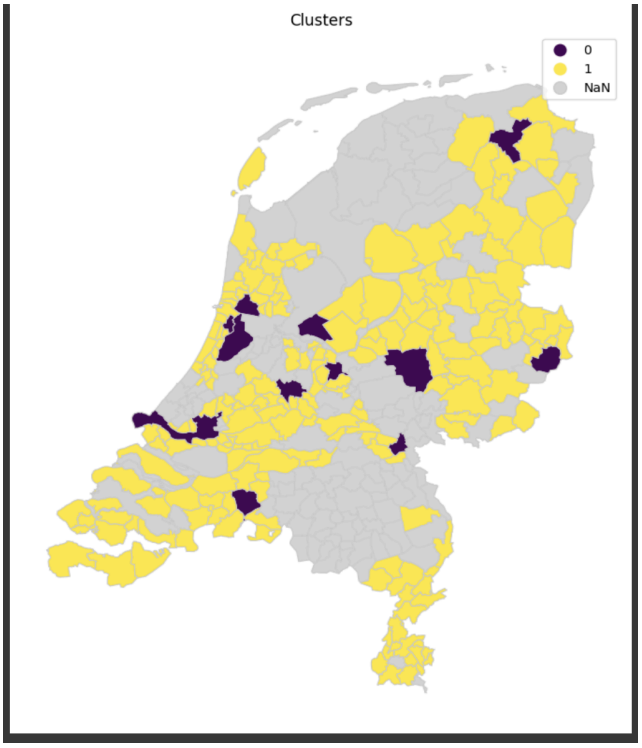
Clustering and outlier detection

In exploring the potential clustering of regions, we employed K-means clustering, evaluating its performance using the silhouette. Remarkably, the best score was attained when the dataset was partitioned into two clusters. Subsequently, we implemented outlier detection using SVM and Isolation Forest models[2] (Fig. 2).

[2] JSON File can be obtained by link https://gadm.org/download_country.html

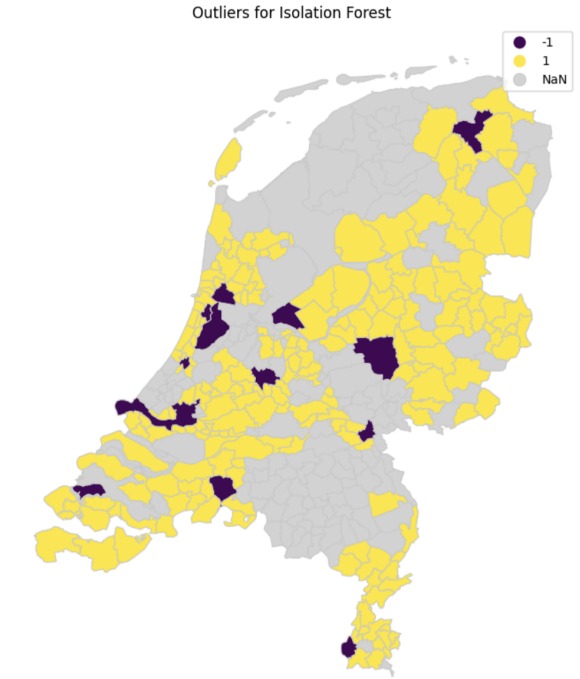


(Fig. 2)

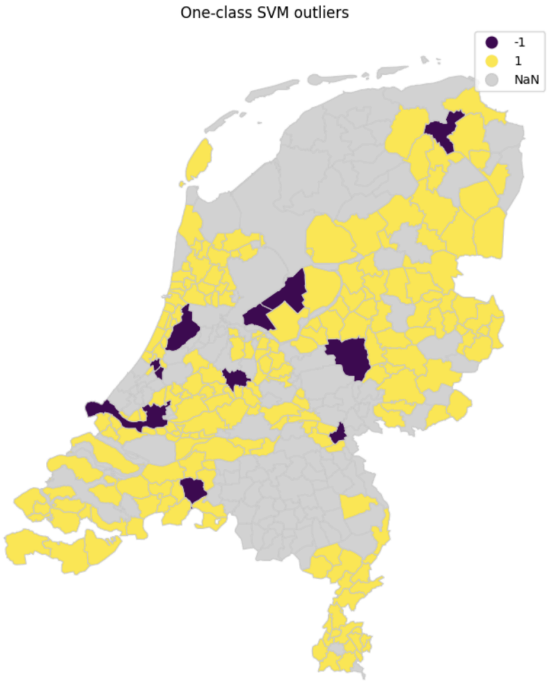


(Fig. 3)

Interestingly, the cities identified as outliers are similar with those found in the second clusters. To ascertain whether these outliers truly represent distinct clusters or share characteristics with the broader dataset, further investigation is warranted. This observation prompts a deeper exploration to unravel the nuances of regional patterns and identify potential factors contributing to the formation of clusters and outliers within the dataset. In the visualizations you can see municipalities marked as different clusters and as outliers.


Outliers for Isolation Forest
-1
1
NaN

(Fig. 4)


One-class SVM outliers
-1
1
NaN

(Fig. 5)

Conclusion

Throughout this investigation, we thoroughly examined the performance of diverse models, including Lasso Regression, Ordinary Regression, Elastic Net, Ridge Regression, and Random Forest Regressor, within our dataset. Our comprehensive analysis revealed that Ordinary Linear Regression consistently yielded the highest scores.

To optimize our results, we employed various data preparation techniques, including feature selection and standard scaling. The predictive aspect of our study involved generating stress level predictions for regions with missing data, utilizing the best-performing Ordinary Linear Regression model.

For the classification of regions, we turned to K-means clustering, leveraging the support of SVM (Fig. 5) and Isolation Forest (Fig. 4) models for outlier detection. This multifaceted approach provided a holistic understanding of the dataset, uncovering patterns, predicting missing values, and identifying potential clusters and outliers within the regional context.

.