Choosing Features for Stress Prediction with Machine Learning

Katja Bengeri
University of Ljubljana
Ljubljana, Slovenia
kb96968@student.uni-lj.si

Junoš Lukan*
Jožef Stefan Institute
Department of Intelligent Systems
Ljubljana, Slovenia
junos.lukan@ijs.si

Mitja Luštrek*
Jožef Stefan Institute
Department of Intelligent Systems
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Abstract

Feature selection is a crucial step in building effective machine learning models, as it directly impacts model accuracy and interpretability. Driven by the aim of improving stress prediction models, this article evaluates multiple approaches for identifying the most relevant features. The study explores filter-based methods that assess feature importance through correlation analysis, alongside wrapper methods that iteratively optimize feature subsets. Additionally, techniques such as Boruta are analysed for their effectiveness in identifying all important features, while strategies for handling highly correlated variables are also considered. By conducting a comprehensive analysis of these approaches, we assess the role of feature selection in developing stress prediction models.

Keywords

Feature selection, Correlation matrix, Balanced accuracy score

1 Introduction

Machine learning models are increasingly being applied to predict stress, which is critical in various domains such as healthcare, workplace management, and wearable technology. However, one of the major challenges in developing reliable predictive models is identifying the most relevant features from extensive datasets, comprising physiological and behavioural information.

Feature selection plays a key role in addressing this challenge. By selecting only the most informative features, we can reduce noise, prevent overfitting, and enhance model accuracy. As we showed in previous work [8], even simple feature selection techniques can increase the F_1 score of predictive models. This paper builds upon this finding and explores several feature selection techniques, ranging from simple correlation-based methods to more sophisticated wrapper approaches.

The aim of this work is to assess how feature selection can enhance stress prediction models. By comparing different methods, we aim to identify the optimal strategies for feature selection in stress prediction which would lead to more reliable and more easily interpretable machine learning models.

2 Data collection

The data used in this work comes from the STRAW project [1], results of which have been previously presented at Information Society [6, 8]. The dataset includes the data of 56 participants, recruited from academic institutions in Belgium (29 participants) and Slovenia (26 participants). They answered questionnaires named Ecological Momentary Assessments (EMAs) roughly every 90 minutes, with smartphone sensor and usage data continuously collected by an Android application [7], while also wearing an Empatica E4 wristband recording physiological data. In 15 days of their participation, each participant responded to more than 96 EMA sessions, on average, which resulted in around 2200 labels.

3 Target and feature extraction

To fully leverage the potential of the data, we computed a comprehensive set of features. While some sensors only reported relatively rare events, such as phone calls, others had a high sampling frequency, such blood volume pulse which sampled data at 32 Hz. On the other hand, labels were only available every 90 min. Therefore, we preprocessed the data in several steps.

3.1 Target variable

While participants responded to various questionnaires, for this study, we selected their responses to Stress Appraisal Measurement [9] as the target variable. It was used to report stress levels on a scale from 0 to 4, so using it as is the prediction task can be approached as a regression problem.

However, many stress detection studies tend towards a discrete approach, treating stress predominantly as a classification task, often only working with a binary target variable. To convert this into a classification problem, we discretized the target variable into two distinct categories: "no stress", which included all responses with a value of 0, while all others were coded as "stress". With that, we ensured a balanced distribution of the target variable values.

3.2 Features

3.2.1 Data preprocessing. In our work, features were calculated on 30-minute intervals preceding each questionnaire session. From the wide variety of smartphone data and physiological measures, a total of 352 features were extracted and grouped into 22 categories, listed in Table 1. Using physiological data from Empatica wristband, we first calculated specialized physiological features on smaller windows (from 4 s to 120 s, depending on the sensor; see [4] for more details), which were then aggregated over 30 min windows by calculating simple statistical features: mean, median, standard deviation, minimum, and maximum. All of the categorical features were converted into a set of binary features using the one hot encoding technique and the missing values were replaced with the mode.

First, some preliminary data cleaning was performed by excluding one of the feature in pairs exhibiting a correlation coefficient of $|r| \geq 0.95$. Despite this, some of the remaining features still exhibited quite strong correlations as shown in Fig. 1. An interesting observation used in the later stages of feature selection was that high correlation, $|r| \geq 0.8$, was mostly observed

*Also with Jožef Stefan International Postgraduate School.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.991

Table 1: Feature categories with the number of features included in each category in parentheses

1. Empatica electrodermal activity (99)
2. Empatica inter-beat interval (50)
3. Empatica temperature (33)
4. Empatica accelerometer (23)
5. Empatica data yield (1)
6. Phone applications foreground (47)
7. Phone location (18)
8. Phone Bluetooth connections (18)
9. Phone calls (10)
10. Phone activity recognition (7)
11. Phone Wi-Fi connections (7)
12. Phone screen events (7)
13. Phone light (6)
14. Phone battery (5)
15. Phone speech (4)
16. Phone interactions (2)
17. Phone messages (2)
18. Phone data yield (1)
19. Baseline psychological features (7)
20. Language (2)
21. Gender (2)
22. Age (1)

between features of the same category. As an example, correlations between features related to phone application use are shown in Fig. 2.

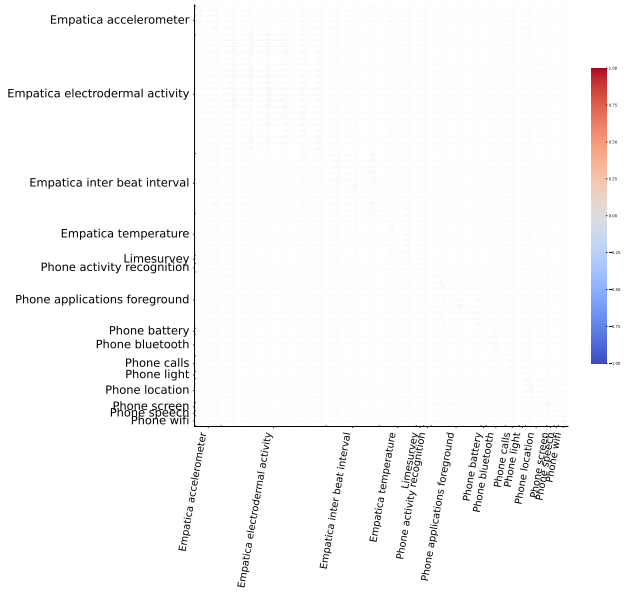

Figure 1: Correlation matrix of the initial feature set. Only feature categories with more than two features are labelled.

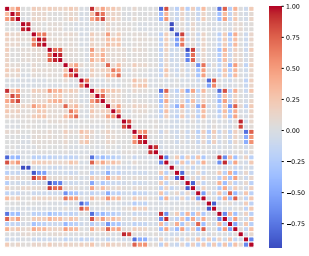

Figure 2: Correlation matrix of the feature set in the Phone applications foreground category.

4 Prediction models

4.1 Model performance and validation

To evaluate the performance of the models we used balanced accuracy score which is defined as the average of recall obtained on each class. When adjusted for random chance, it is calculated as

$$\textit{Balanced accuracy} = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1,$$

in the binary case, where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives and FP is the number of false positives. This definition is equivalent to Youden's $\mathcal{J}$ [11], which assigns a 0 to a random classifier (indeed, a dummy classifier achieved a score of 0.0208 in our case), while a perfect classifier would achieve a score of 1.

To evaluate the stress detection models described in the following sections, we considered several ways of data partitioning. Since the variations in the results depending on the data split were significant, in order to achieve more consistent accuracy, we employed shuffled 5-fold cross-validation.

We also considered a leave-one-subject-out cross-validation technique. However, this method yielded poor results: using all available features, balanced accuracy was 0.05, while with the 5-fold cross validation it was 0.45. This suggested that the participants were quite different from each other, making it challenging to generalize predictions for a subject the model had not encountered.

4.2 Baseline model

Our initial approach for building a prediction model was to use all available features. This served as a baseline, which we aimed to improve through feature selection.

We evaluated various predictive models, as shown in Table 2, all as implemented in `scikit-learn` [10]. Among these, the Random Forest model yielded the best performance.

In this work, we aimed to find the best model for predicting stress and improve it using the optimal feature subset. Consequently, we used the Random Forest as the benchmark for comparing feature selection algorithms.

Table 2: Performance of different models for the classification problem. The mean over five folds, its standard error, and the maximum are shown.

Model	Mean	Max	SEM
Logistic Regression	0.077	0.151	0.025
Support Vector Machines	0.090	0.158	0.022
Gaussian Naive Bayes	0.061	0.122	0.020
Stochastic Gradient Descent	0.027	0.054	0.007
Random Forest	0.475	0.558	0.026
XGBoost	0.441	0.473	0.013

In Table 2, SEM represents the Standard Error of the Mean. It measures how far the sample mean of the data is likely to be from the true population mean.

4.3 Correlation-Based Feature Reduction

We began the feature selection process by eliminating highly correlated features. For each highly correlated pair, we removed the feature with the lower rank when sorted by mutual information, setting the correlation threshold at $|r| \geq 0.8$ to maintain a manageable number of features. This reduction left us with approximately 180 features out of the original 352 for model training and evaluation.

While selecting the optimal set of features for stress prediction, we aimed to retain all 22 different categories from Table 1, as

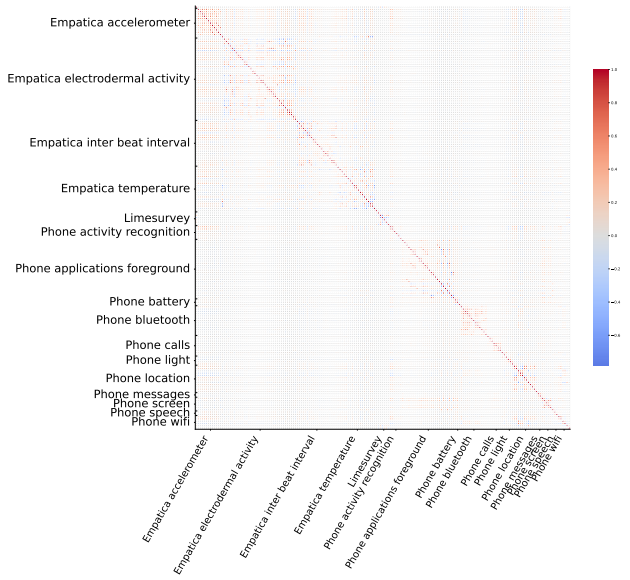

Figure 3: Correlation matrix of the feature set after correlation-based feature reduction. Only feature categories with more than two features are labelled.

each could provide unique information. Comparing Figs. 1 and 3, we were left with about half the number of features which were still moderately correlated.

4.4 Feature Selection using the mutual information scoring function

Before applying more complex feature selection algorithms, it was necessary to reduce computational complexity by further reducing our set of 180 features obtained through correlation-based reduction. Therefore, we used the `SelectKBest` method and the mutual information scoring function to retain the top 100 features. This resulted in features derived from 19 to 20 categories, as categories *language*, *gender*, and, in some cases, *Empatica accelerometer* were not deemed important for predicting stress.

Going forward, we will refer to the elimination of features within highly correlated pairs and the selection of the top 100 features using the mutual information scoring function as the preprocessing step.

4.5 Recursive Feature Elimination with Cross-Validation (RFECV)

One of the previously mentioned complex feature selection methods we employed was Recursive Feature Elimination with Cross-Validation (RFECV) [3]. The feature set we got after the preprocessing step was passed to the RFECV algorithm for thorough evaluation.

RFECV operates by initially fitting a model to the dataset and evaluating its performance through cross-validation. After the initial fit, RFECV ranks feature importance and iteratively removes the least important features based on the models feature importances attributes, which in the case of Random Forest are impurity-based feature importances. This process continues until there is no significant improvement in the model's performance. To ensure a reasonable duration for the feature selection process, we set the cross-validation in RFECV to 3 folds. The number of features selected varied across folds, ranging from 50 to 93 features.

4.6 Sequential Forward Selection

Another feature selection method we employed was Sequential Feature Selector (SFS), a wrapper-based technique [2]. SFS and RFECV differ in their approaches. SFS constructs models for each feature subset at every step, while RFECV builds a single model and evaluates feature importance scores. Consequently, SFS is more computationally expensive, as it must evaluate numerous feature combinations before identifying the optimal subset.

In the absence of specified parameters for number of features to select (`n_features_to_select`) and tolerance (`tol`), the method defaults to selecting half of the available features. The default configuration was used in our analysis, leading the SFS to select the top 50 features.

4.7 Boruta method

The final feature selection technique we employed was the Boruta method [5]. With the assistance of "shadow features", which are original features that have been randomly shuffled, the method identifies a subset of features that are relevant to the classification task at hand. In our case, shadow features were introduced into the feature subset obtained after the preprocessing step.

The updated dataset was trained using the Random Forest model for 100 iterations. In each iteration, all original features ranked higher in importance than the highest-ranked shadow feature were marked as relevant.

Ultimately, a binomial distribution is used to evaluate which features have enough confidence to be kept in the final selection. The number of features selected varied across folds, ranging from 47 to 55 features.

5 Results

In Table 3, the final scores for a Random Forest model built on various feature subsets, as derived from the methods described above, are presented. The data was split using shuffled 5-fold cross-validation, to ensure that the results were not overly dependent on a data split.

Table 3: Adjusted balanced accuracy scores of a Random Forest model, trained on the different feature sets. Last column represents a number of features selected.

Feature set	Mean	Max	SEM	N
All available features	0.464	0.498	0.011	352
Correlation-based reduction	0.483	0.507	0.007	~180
Correlation-based, 100 best	0.486	0.498	0.006	100
Preprocessing, RFECV	0.471	0.511	0.012	50 to 93
Preprocessing, SFS	0.483	0.520	0.017	50
Preprocessing, Boruta	0.481	0.545	0.020	47 to 55
RFECV only	0.465	0.494	0.020	16 to 89
SFS only	0.426	0.468	0.017	30
Boruta only	0.456	0.509	0.015	~75

From Table 3, we can see that the most significant improvement in accuracy came after removing the highly correlated features, with the average adjusted balanced accuracy score rising from 0.46 to 0.48. Best mean accuracy was achieved after the preprocessing step, with only a minor improvement from 0.483 to 0.486.

After eliminating highly correlated features, wrapper methods did not significantly improve the accuracy on average (rows 3 to 6 in Table 3). The Boruta method performed best among the three, with the highest overall maximum accuracy in a single fold. These results led us to investigate whether the wrapper feature selection method alone could manage correlated features without their prior removal and to evaluate the impact of the correlation threshold.

We employed the RFECV, SFS, and Boruta method on the entire feature set of 352 features without applying the preprocessing step. For SFS, only 30 features were selected due to its computational complexity. As shown in the last three rows of Table 3, none of the methods alone were able to improve the result achieved with correlation removal. Highly correlated features were left in the final feature set: for example, we identified three pairs of features with a correlation coefficient exceeding $|r| \geq 0.8$ using SFS alone. Poor results could be attributed either to the importance of the correlation removal step or to the feature subset being too small in the case of the SFS.

5.1 Selecting the best correlation threshold

As previously mentioned, the biggest improvement in score came from removing the feature inside the highly correlated pair. Therefore, we have also experimented with different correlation cut-off values to determine the best threshold.

The highest score was achieved with a correlation threshold of $|r| \geq 0.75$ (Table 4). Considering the impact of cross-validation splits and the relatively minor variance in scores, it appears that our initial threshold of $|r| \geq 0.8$ was also quite effective.

Table 4: Adjusted balanced accuracy scores of a Random Forest model trained on a feature subset excluding features above the correlation threshold. The number of features left after correlation-based feature selection differed over validation folds and its range is shown in the final column.

Threshold	Mean	Max	SEM	N
0.55	0.462	0.506	0.018	28 to 33
0.60	0.467	0.493	0.009	39 to 41
0.65	0.474	0.498	0.008	47 to 50
0.70	0.460	0.501	0.017	61 to 65
0.75	0.498	0.526	0.012	74 to 80
0.80	0.470	0.543	0.022	101 to 107

6 Conclusions

This paper examined different feature selection algorithms to find the most effective subset for stress prediction. The model using the feature subset after correlation removal achieved the highest adjusted balanced accuracy score of 0.483.

Alternative feature selection approaches, including the wrapper methods SFS and RFECV, as well as the Boruta method, applied to the preprocessed feature subset, did not lead to further optimization of the feature subset in terms of model performance. Additionally, applying these methods to the entire set of features did not achieve accuracy levels as high as those obtained after the correlation-based reduction. In the case of SFS, this may be attributed to its selection of only 30 features.

Therefore, our results underscore the critical role of the correlation-based reduction step. In contrast, when this step was omitted wrapper methods alone were unable to effectively perform correlation-based feature reduction. We can therefore conclude that simply relying on feature selection methods, however sophisticated, is not as effective as also considering relationships between features.

It should be noted that the improvements in balanced accuracy are low in all cases. This indicates that results cannot be easily generalized and correlation-based feature selection should not be seen as sufficient in general. Instead, we can speculate that no single feature selection method is the best one and that several should be considered. We should also note that the Pearson correlation coefficient that we used in this work only considers linear relationships between features. Other methods can select features even if they are related in a different way.

References

[1] Larissa Bolliger, Junoš Lukan, Mitja Luštrek, Dirk De Bacquer, and Els Clays. 2020. Protocol of the STRess at Work (STRAW) project: how to disentangle day-to-day occupational stress among academics based on EMA, physiological data, and smartphone sensor and usage data. *International Journal of Environmental Research and Public Health*, 17, 23, (Nov. 2020), 8835. DOI: 10.3390/ijerph17238835.

[2] Francesc J. Ferri, Pavel Pudil, Mohamad Hatef, and Josef V. Kittler. 1994. Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, 16, 403–413. DOI: 10.1016/b978-0-444-81892-8.50040-7.

[3] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using Support Vector Machines. *Machine Learning*, 46, 1/3, 389–422. DOI: 10.1023/a:1012487302797.

[4] Vito Janko, Matjaž Boštic, Junoš Lukan, and Gašper Slapničar. 2021. Library for feature calculation in the context-recognition domain. In *Proceedings of the 24th International Multiconference Information Society – IS 2021*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 4–8, 2021). Vol. A, 23–26.

[5] Miron B. Kursa and Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, 36, 11, 1–13. DOI: 10.18637/jss.v036.i11.

[6] Junoš Lukan, Larissa Bolliger, Els Clays, Primož Šiško, and Mitja Luštrek. 2022. Assessing sources of variability of hierarchical data in a repeated-measures diary study of stress. In *Proceedings of the 25th International Multiconference Information Society – IS 2022*. Pervasive Health and Smart Sensing (Ljubljana, Slovenia, Oct. 10–14, 2022). Vol. A, 31–34.

[7] Junoš Lukan, Marko Katrašnik, Larissa Bolliger, Els Clays, and Mitja Luštrek. 2020. STRAW application for collecting context data and ecological momentary assessment. In *Proceedings of the 23rd International Multiconference Information Society – IS 2020*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 5–9, 2020). Vol. A, 63–67.

[8] Marcel Franse Martinšek, Junoš Lukan, Larissa Bolliger, Els Clays, Primož Šiško, and Mitja Luštrek. 2023. Social interaction prediction from smartphone sensor data. In *Proceedings of the 26th International Multiconference Information Society – IS 2023*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 9–13, 2023). Vol. A, 11–14.

[9] Edward J. Peacock and Paul T. P. Wong. 1990. The stress appraisal measure (SAM). A multidimensional approach to cognitive appraisal. *Stress Medicine*, 6, 3, (July 1990), 227–236. DOI: 10.1002/smi.2460060308.

[10] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[11] Charles Sanders Peirce. 1884. The numerical measure of the success of predictions. *Science*, ns-4, 93, 453–454. DOI: 10.1126/science.ns-4.93.453.b.