Introduction of prediction-based feature selection methods in computational psychology: An application in the psychology of sleep

Theodoros Efthymiadis
Dep. of Informatics and Telecommunications
NCSR "Demokritos"
Athens, Greece
efthimiadisthodoris@gmail.com

Tamara Rigo
Dep. of Mathematics
University of Trento
Trento, Italy
tamara.rigo@studenti.unitn.it

Kyriaki Bei
Dep. of Informatics and Telecommunications
NCSR "Demokritos"
Athens, Greece
sandybei058@gmail.com

***Abstract*—The current paradigm in psychological research involves the use of basic computational techniques and a focus on identifying causal mechanisms through application of well-established statistical and modelling tools. These traditional approaches mainly focus on identifying important predictor variables based on the model's 'goodness of fit'. However, the modern trends in computational psychology are indicative that the adoption of a prediction performance-based approach derived from the field of Machine Learning will be highly beneficial for the domain. The aim of this study is to analyze existing traditional and emerging feature selection methods and provide a framework for comparison between the two. The proposed methodology is applied on a specific classification task in the field of the psychology of sleep. The comparison of the different feature selection methods is grounded on a baseline classification task using a logistic regression model, given that it's potentially the most familiar machine learning algorithm to the psychology research community. Our methodology involves the employment of a k-fold cross validation scheme using the accuracy evaluation metric, as well as the application of a series of statistical tests. We conclude that, regarding the specific data set, the application of our methodology did not result to statistically significant differences in the achieved model accuracy. Consequently, we cannot directly argue about the need to change the current paradigm of computational psychology and, thus, further studies are required.**

***Index Terms*—computational psychology, feature selection, sleep quality**

I. Introduction

The motivation behind this study was mainly derived from the review of Tal Yarkoni and Jacob Westfall (2017) about the potential benefits that could be provided through the shift from the current dominant paradigm used in psychological research towards a more predictive approach [1]. The authors pointed out the tendency of psychology studies to focus too much on identifying the causal mechanisms behind behaviors, instead of striving for model predictive performance and accuracy. They argued that enriching the traditional modelling approach, which mainly focuses on 'goodness of fit', with the adoption of machine learning prediction - based techniques and concepts, such as Cross Validation, would provide invaluable advantages: higher efficiency and reproducibility of researchers' results, better models' performance and, ultimately, a better understanding of behavioral patterns. In that regard, it is also important to examine if adopting prediction performance-based feature selection methods (wrapper methods), as well as tree-based feature selection methods (embedded methods) will potentially lead to the selection of more informative feature sets than traditional correlation-based feature selection methods (filter methods), allowing for the better understanding of the underlying causal mechanisms. This is a very broad question that has to be answered through iterative work of scientists in the field of psychology over a long period of time.

The actual limits of the computational psychology are also underlined by Sharp and Eldar (2019) [2]. Most statistical approaches in psychological studies consist in the probabilities generation of how unlikely one's data are, assuming a point-null hypothesis is true [2]. This method is fallacious, issue already noted by Mehls in the 90's, when he elaborated the concept of crud factor [3] — *the idea that everything correlates with everything else in psychological science* (Orben and Lakens, 2020). Sharp and Eldar suggest instead the computational approach of quantitative Bayesian model to compare posterior probabilities, i.e. the probability that each theory is to be true given the collected data.

The study of H.B.F. David et al (2019) represents a good example of the advantages of machine learning algorithms application in psychology. The author used a combination of 7 different heuristic search algorithms with 7 different feature subset evaluation methods on a dataset of 30 attributes for the classification of prisoners according to psychological and behavioural factors. Each derived feature subset was used to train a fixed Support Vector Machine Model and reported a series of evaluation metrics. It was found that the usage of correlation-based feature subset evaluation technique and radial basis function classifier, combined with the 'wolf search algorithm' method, could lead to 97.8 % precision, 97.5 % recall and low error estimates [4].

In order to contribute to this effort, this study aims to investigate if the current feature selection techniques mainly

used in psychological studies are effective, especially when compared to more advanced methods. To test that we will provide a case study of evaluation of different feature selection methods in the field of the psychology of sleep. Many studies in this area mainly involve the basic use of linear and logistic models for regression or classification tasks, without utilizing more advanced algorithms. For instance, these methods were applied by Marzabadi and Amiri [5] during their study of the association of chronotype with social anxiety, as well as by Chan et al. [6] in their work on the correlation between delayed school start time and various positive traits, such as better sleep quality. Feature selection is also frequently performed by simply looking at the correlations between the predictors and the response and choosing the predictors with the strongest association. In this approach, correlations are obtained through linear or logistic regression coefficient [7], [8], Spearman coefficients [9] and Chi -Square statistics [10] [11]. An exception can be found in the work of Gomes et al. (2011), whose aim was to test whether sleep variables are significant predictors of academic performance, when other potential predictors such as class attendance are considered. To test that the authors used three different techniques sequentially. At first, they identified through ANOVA analysis which sleep variables were significantly associated with end-of-semester marks. [12] Then, from a total of 30 variables initially considered, a preliminary univariate analysis selected 19 of them significantly related to z scores ($p < .05$). [12] At the end, after having removed other variables from the analysis to avoid multicollinearity with similar ones, stepwise multiple regression analysis identified 5 significant predictors of end-of-semester marks: previous academic achievement, class attendance, frequency of sufficient sleep, night outings and sleep quality [12].

II. Theoretical Foundations

We decided that the comparison of different feature selection methods in the field of the psychology of sleep would be conducted based on each method's performance in a specific classification task. Therefore in order to start defining a it we needed:

- A dataset.
- Different feature selection methods: each of one producing a different feature subset.
- A baseline machine learning algorithm: the model choice fell on the Logistic Regression, as it is widely used in psychology experiments [13].
- An evaluation metric: we decided to obtain through the usage of k-fold Cross Validation a sample for each feature subset the model was trained on, each sample containing k values of prediction accuracy.

We will first briefly introduce in a pure theoretical way the methods we used , and then deepen them by describing step by step our workflow.

A. Cross Validation

Cross Validation refers to a fundamental family of resampling methods. Resampling involves repeatedly drawing samples from a training set and fitting a model on each sample, in order to obtain additional information that would not be available from fitting the model on the original training set only once [14]. Although its explicit use is not so common in contemporary psychological studies, Cross Validation has been known in the field since the late 40s [1], as it was observed that *if the combining weights of a set of predictors have been determined from the statistics of one sample, the effectiveness of the predictor-composite must be determined on a separate, independent sample. This is the case whether the combining weights are multiple-regression beta weights or item-analysis weights of one or zero*. (Charles Mosier, 1951)

K-fold Cross Validation is currently the most commonly used method: this approach starts with randomly splitting the set of observations into k groups or folds, typically $k = 10$. $k-1$ folds are treated as training set to fit the model, while one held-out fold is treated as validation set , on which the estimate of the test error is computed. The procedure is repeated k times, each time a different group of observations is treated as validation set [14]. The final k-fold Cross Validation estimate is computed by averaging all the resulting k estimates of the test error. In particular for our work we opted for Stratified Cross Validation, a variation of k-fold in which the distribution of samples between classes is kept constant in each fold.

B. Statistical Tests

Our work involves the use of the following tests, applied on the accuracy samples:

- **Shapiro Wilk test, Levene's test and Bartlett's test**
  Prelimany tests necessary for checking the assumption for ANOVA analysis:
  - Each group sample is drawn from a normally distributed population: determined by Shapiro Wilk test.
  - All samples have the same variance (homoskedasticity): determined by Bartlett's test or Levene's test.
  - All samples are independent: determined by the way the samples were created.
- **ANOVA**
  One-way analysis of variance is a statistical method for testing the null hypothesis according to which the means of three or more k groups would be equal, against the alternative hypothesis according to which at least one mean would be different. In statistical terms:
  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
  $H_1 : not\ all\ means\ are\ equal$
  To determine if there is a statistically significant difference between groups, ANOVA analysis relies on F statistic, whose formula is:
  $F = \frac{SSB/(k-1)}{SSW/(n-k)}$, where n is the number of valid observations, SSB (Sum of Squared Between) is the sum of the variances between the means of the various groups

and SSW (Sum of Squared Within) is the sum of the variances within the means of the various groups.

- **Tukey's Honest Significant Difference test**
  The Tukey HSD test is supposed to be performed only after a significant ANOVA test in order to find out which specific groups' means are different, since ANOVA can only say if these differences are present or not. The Tukey test does that by comparing all possible pairs of means. In particular HSD statistic is calculated for each pair using this formula: $HSD = \frac{M_i - M_j}{\frac{\sqrt{MS_w}}{n_h}}$
  ,where $M_i - M_j$ is the difference between the pair of means, MS_w is the Mean Square Within and n is the dimension of the group or sample.
- **Kruskal-Wallis H test and Wilcoxon Test**
  To be applied instead of ANOVA analysis in case the latter's assumption are not met, since Kruskal-Wallis is a nonparametric (distribution free) test. Null hypothesis assumes that the samples are from identical populations, while the alternative hypothesis assumes that at least one of the groups comes from a different population than the others. H statistic follows an approximation of Chi-Square distribution and is calculated using this formula: $H = (\frac{12}{n(n-1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j}) - 3(n-1)$
  ,where n is the sum of sample sizes for all samples, k the number of samples/groups, T_j the sum of ranks in the j-th sample and n_j the size of the jth sample.
  After having performed a significant Kruskal-Wallis test, Wilcoxon test can be performed as post hoc testing procedure to ascertain which pairs of groups differ significantly from one another.

C. Feature Selection

Data can often involve a large number of features in their representation. However, only a few of them may be relevant to a target variable. In order to reduce data dimensionality, deal with high data dependency, speed up the the learning process and improve its performance, has given rise to the need to further develop feature selection techniques. Feature selection can be defined as the process of obtaining a subset of the original dataset's feature set that provides efficient understanding of the data by removing features that withhold irrelevant or redundant information [15]. Feature selection techniques can be classified into 3 main categories with respect to the learning approach they follow. These categories are:

1) **Filter methods**: The filter approach evaluates the importance of features by ranking them with the use of different measures. In order to select the ones that are statistically significant, a threshold has to be set for their score. Thus, the feature subset that results from this approach is comprised of the features whose scores are higher than the threshold value. In addition, these methods can be devided into univariate and multivariate filter methods. Univariate methods rank each feature individually, while multivariate ones rank entire feature sets. A disadvantage of univariate methods is that they do not take into consideration the relationships among features as they evaluate the importance of features based on a certain criterion independently. This could result to a feature set that includes redundant features. Moreover, filter methods do not include the use of a learning algorithm, which render them computationally fast when compared to other approaches that integrate their use. Common filter methods include include Pearson Correlation Coefficiant, Spearman's Rank Correlation Coefficient, ANOVA, Chi-Square and Mutual Information.
2) **Wrapper methods**: Wrapper methods evaluate different feature subsets based on the performance of a machine learning algorithm trained on the data. The optimal feature set gets selected after the repeated evaluation of the model after its training on different feature subsets. Although the use of a learning algorithm for the selection of the most efficient feature subset is more demanding computationally than the filter methods, it has been empirically proven that it can have better results [16]. Wrapper methods can either add or remove features based on the model's prediction error and based on the approach that they follow they can be divided into 3 types of methods: Forward Selection, Backward Elimination and Step-wise Selection. All of the above methods can use any machine learning algorithm in order to obtain the best feature subset. However, it is common to select models with less computational requirements, such as Logistic Regression, Linear Regression and Naïve Bayes.
3) **Embedded methods**: Embedded methods, like wrappers, use the evaluation of a learning algorithm for the selection of the best features but they take on a different approach. Embedded methods perform the process of feature selection during the construction of the model. The best feature subset gets obtained by keeping only the features that are important for the prediction of the target variable. Although they can be computationally faster compared to wrappers, the obtained feature set is dependent on the assumptions of the used algorithm as their selection takes place during its construction. For this reason, practices that can help overcome this problem , such as using an independent validation set for the model's evaluation, have to be applied. Embedded methods include tree-based methods, such as Decision Trees, Random Forest and XGBoost, and methods that perform regularisation by adding a penalty to the parameters of the model, such as Lasso Regression, Ridge Regression and Elastic Net.

In our methodology, we have selected methods from all three of the aforementioned method categories for the evaluation of our dataset's features' importance. Moreover, in order to have a better comparison, we decided to choose both simpler methods typically used in psychologist studies and more complex algorithms that are still rarely used in this

field. The subset of the features was derived for each method based on a different evaluation criterion. The methods that were chosen are the following:

- **Chi-Square**: Chi-Square, also denoted as χ^2, is a non-parametric statistical test that determines whether two categorical variables are independent. The formula of the test is: $x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
  ,where c is the degrees of freedom, O_i is the observed value and E_i is the expected value. The null hypothesis (H_0) and alternative hypothesis (H_1) can be expressed in statistical terms as follows:
  H_0: X_1 is independent of X_2
  H_1: X_1 is not independent of X_2
  ,where X_1 and X_2 are two variables of the population. Feature selection Chi-Square statistic indicates which features are highly dependent on the response and which must therefore be selected. It can be applied on a categorical response combined with another categorical variable or multiple numerical variables.
- **Logistic Regression coefficients**: After performing a multivariate logistic regression model, it is possible to calculate adjusted odd ratios for associations between the predictors and the outcome and select the predictors with the strongest association.
- **Forward Feature Selection**: Forward Selection is an iterative greedy procedure that finds the best feature to add at each iteration using the performance of a model as an evaluation criterion. At the first step of the procedure we start with no features. At each step, the feature that maximizes the model's cross-validation performance score gets selected. The procedure terminates when there is no feature that can better the model's performance.
- **Random Forest**: Random Forest is a classifier which is a combination of a number of decision-trees and can be used for feature selection. In a decision tree, each node represents a condition on an individual feature [17]. In order to optimize the performance of Random Forest, we have to determine which is the best feature to split at each step of a tree. Measures like Gini Impurity and Information Gain can be used for feaure ranking.
- **Lasso Logistic Regression**: Lasso Logistic Regression (also referred to as L1-Regularisation) is a Logistic Regression model that has an additional penalty term in its loss function. Loss function is the measure that evaluates how well a prediction model performs. Adding a penalty term to that function causes the shrinkage of the model's coefficients' magnitude, thus performing feature selection. Unlike other regularisation techniques, like Ridge Regression, Lasso allows coefficients of variables to be set equal to zero. The final subset of important features that result from this method include those variables whose coefficient do not equal to zero.

As already mentioned, the first two techniques are frequently used in computational psychology, while the remaining more advanced algorithms are not so common.

III. Methodology

The end goal of the analysis is the comparison of the different feature selection methods and their suitability for the identification of important predictor variables. The comparison is conducted on the basis of a baseline classification task and the evaluation of the cross - validation accuracy achieved by the same machine learning algorithm when using the feature subsets that were generated through the different feature selection methods. A graphical representation of the workflow is illustrated in "Fig. 1" below:

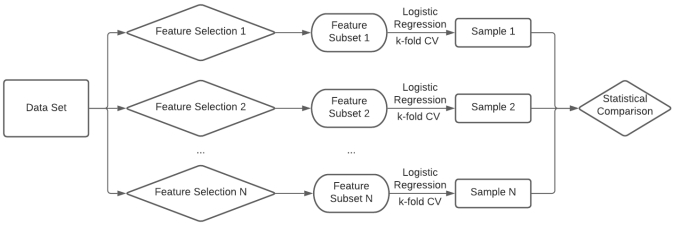

Fig. 1. Designed workflow of the project

Specifically, the feature subsets were used to train a non-penalized logistic regression model using a 10-fold cross validation scheme. However, instead of only reporting the mean 10-fold accuracy value for each model, all 10 accuracy values were used to create a size 10 sample for each different feature selection method. The samples were compared through a pipeline of statistical tests that is described in detail in the following paragraphs.

A. The Data

The data we used were collected by Ray Norbury and Simon Evans (2018) during their research about the associations between sleep quality, trait anxiety and university start time [18]. The sample was composed by 546 students from two universities of South East England, ranged in age 18 – 55 years and 84 % self-identified as female [18].

The 21 attributes are the following:

- Case Number (ID)
- University
- Age
- Sex as a dichotomous (male/female) variable (1 = male, 2 = female)
- Year of Study
- Department Name
- MEQ: subjects scores on the Reduced Morningness-Eveningness Questionnaire to determine their chronotype
- Trait Anxiety: subjects scores on the Trait Anxiety Index
- Start time for university activities
- Start time code: the above variable codified into a categorical variable
- Average Weekly Sleep Duration
- Average Sleep Duration on working days
- Average Sleep Duration on free days
- Daytime Dozing: assessed using the Epworth Sleepiness Scale

- Daytime Dozing Groups: created by dividing subjects in 5 groups based on their response on the Epworth Sleepiness Scale
- Caffeine and coffee consumption codified as dichotomous (yes/no) variables (1 = yes, 2 = no)
- Tobacco use and alcohol consumption codified as dichotomous (yes/no) variables (1 = no, 2 = yes)
- PSQI component 1: subjective sleep quality of the participants classified in 4 categories according to the Pittsburgh Sleep Quality Index (0 = very good sleep, 1 = fairly good, 2 = fairly bad and 3=very bad)
- PSQI 2 Groups: the above variable dichotomized as good/poor sleepers (1 = good sleepers, 2 = bad sleepers)

For our features selection purposes not all of the original variables are useful, therefore we removed the Case Number (ID) variable and those which are also present in a more convenient codified version (Daytime Dozing, Start time code and PSQI component 1), choosing PSQI 2 Groupsas response.

B. Data Preprocessing

Data processing steps were necessary in order to prepare the data for the training of the Logistic Regression model. Initially, the dataset was explored for the existence of noisy data instances. The analysis that was performed showed that there are no missing values or inconsistent data points.

The next step of data processing was to identify the data types of the features that would be used for the training of the model and to handle them accordingly. The data types that were identified were categorical, ordinal and numerical. Numerical features (Average Weekly Sleep Duration, Average Sleep Duration on working days, Average Sleep Duration on free days) were rescaled to take values within the range [0, 1]. Dichotomous categorical features (Sex, Caffeine, coffee, tobacco use and alcohol consumption) were transformed to take a value between 0 and 1 instead of 1 and 2, so that they had ranges similar to the majority of the features. Categorical features that had more than two categories (University, Department Name, Start time code, Daytime Dozing Groups) were encoded by the use of one-hot encoding method where binary columns get created for each category. Lastly, we did not perform any transformations for the ordinal features of the dataset (Age, Year of Study, Trait Anxiety, MEQ, PSQI Component 1).

C. Feature Selection Methods

After performing the necessary processing steps, the dataset was split in training set (70% of the data set) and test set (30% of the data set). The existence of an independent validation set was a prerequisite for reliable error estimates of the wrapper and embedded employed feature selection methods. The same approach was adopted for the selected filter methods as well in order to ensure a fair comparison of results. Five different methods were explored. In the following paragraphs, the application of each method will be briefly presented along with the corresponding feature subset.

To determine the statistical significance of the difference between features and sleep quality using Chi-Square statistical test, we used p-value as an evaluation criterion. The significance level p-value was chosen to be equal to 0.05. Using this threshold, the features that were statistically significant (features that had a p-value higher than 0.05) were removed from the original feature set. The variables that were selected to be the most important by this method were Year of Study, MEQ, Trait Anxiety, University and Start Time Code.

In the case of the Logistic Regression coefficients method, p-value was also used as an evaluation criterion: variables were selected only if their adjusted odd ratios had at least a significance level of 0.05 . In this way Age, Year of Study, MEQ, Trait Anxiety, Average Weekly Sleep Duration, Average sleep free days, coffee dichotomous, University, Department Name, Start time code and Daytime Dozing Groups were selected.

For the implementation of Forward Feature Selection, we selected to evaluate the features by using a Logistic Regression model. At each iteration of this method, each feature was added to the previous feature set and this subset was then evaluated based on its accuracy scores that resulted from a 10-fold cross validation. The best features that were selected from this wrapper method were Trait Anxiety, Average Sleep Working Days, Department Name, Start time Code, Daytime Dozing Groups.

For Random Forest method, we selected to keep the features that had an importance score of at least 0.03. The features that were selected were Age, Year of Study, MEQ, Trait Anxiety, Average Weekly Sleep Duration, Average Sleep Working Days, Average Sleep Free Days, Alcohol Consumption, University and Daytime Dozing Groups.

Lasso Logistic Regression gave that the significant features are Trait Anxiety, Year of Study, Average sleep free days, MEQ, Department Name, University, Daytime Dozing Groups, Start Time Code, Sex, Department Name, Daytime Dozing Groups, Cigarettes, Coffee, Caffeine and Alcohol Consumption affect most sleep quality.

D. Statistical Comparison

As mentioned before, a Logistic Regression model was developed and trained on each one of the five feature subsets derived from the selected feature selection techniques. The model was selected to be non-penalized, as adding a penalization factor to its loss function would potentially perform further feature selection.

The five logistic regression models were evaluated using a 10-fold cross validation scheme, while the model accuracy was the evaluation metric of choice. This process led to the creation of five samples, each having ten values of accuracy that represented the accuracy of the five models and, thus, the five corresponding feature selection methods. Consequently, the comparison of the five feature selection methods will be performed through the comparison of the five samples through employment of appropriate statistical tests. The significance level for all tests was set to a=0.05. A graphical representation

of the statistical comparison workflow is illustrated in "Fig. 2" below:

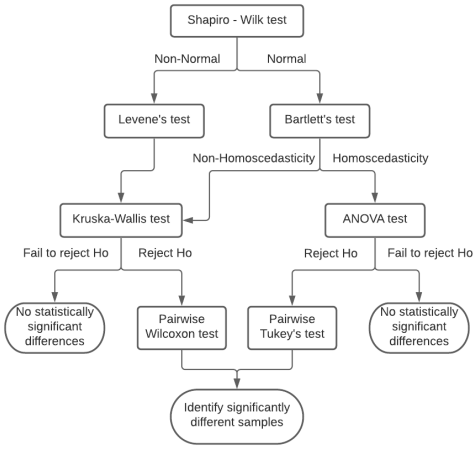

Fig. 2. Workflow for statistical comparison of the k-fold samples

The preferred method for the comparison of the different samples was the one-way Analysis of Variance (one-way ANOVA). However, as already explained, ANOVA is a parametric test with very strong assumptions that have to be evaluated. In case the assumptions of ANOVA do not hold, the non-parametric test of Kruska-Wallis can be used instead. In order to check for the assumption of normality, the Shapiro-Wilk test was used. Depending on the result of the first test, a second test was used to assess the second assumption of ANOVA, i.e. homoscedasticity. There are two common options here, either Bartlett's test, which performs better on normally distributed samples, or Levene's test, which performs better for non-normally distributed samples. It is useful to remember that while both ANOVA and Kruskal-Wallis H test are able to identify the existence of significant differences between the mean accuracy of the compared samples, they are unable to pinpoint the specific samples that manifest these differences in term of accuracy. In order to identify the exact feature selection methods that lead to statistically different results, a post-hoc test is required. After receiving a statistically significant result from one-way ANOVA, the pairwise Tukey's Honest Significant Difference test is employed. Likewise, after receiving a statistically significant result from Kruskal-Wallis test, the pairwise Wilcoxon test is used. In both cases, it's essential that the post-hoc tests are only utilized when the initial test has provided a statistically significant result.

IV. RESULTS

The first part of the analysis was the application of the different feature selection methods and the generation of the corresponding feature subsets.

When using the chi-square test to identify the features that have a statistically significant correlation with the response variable, our results were similar to those of Norbury and Simon Evans [18]. Namely, the most important predictor variables for predicting sleep quality were the Chronotype (measured with MEQ), the year of study, the anxiety levels, the university of the student and the late class starting time (indicated by Start time code 8).

Moreover, when using the stepwise feature selection methods, the most prominent features are average sleep during working days, the department of the student, the class starting time and the daytime dozing. Overall, this was the best set of features to train a non-penalized logistic regression after performing a forward search of the feature space.

On the flipside, the other two feature selection methods that involve the training of a logistic regression model gave significantly different results. For instance, the penalized logistic regression (using lasso) selected more than ten important features. This could be explained by the fact that there are two local accuracy maxima that were identified during the stepwise selection. It's quite possible that the lasso logistic regression converged to the one with the highest number of features. Furthermore, in the logistic regression coefficients method, where the logistic regression coefficients were evaluated using t-tests, even more features were evaluated as important, as most features tend to have some sort of statistically significant correlation with the response variable.

Finally, the tree-based feature selection that capitalizes on the random forest algorithm generated ten features that overlap to a great extend with the subsets that were generated by the other methods. Overall, the most frequent features that are selected by all methods are year of study, chronotype (MEQ), trait anxiety, university and department of the student and daytime dozing. It becomes evident that the various feature selection methods create different feature subsets and that it's not trivial to decide on the best feature subset. Consequently, a robust evaluation process is deemed necessary. The proposed approach in this study is the statistical comparison of a baseline machine learning model that is trained repeatedly on the different feature subsets.

As described earlier, all generated feature subsets were used to train a non-penalized logistic regression model, which was evaluated using a 10-fold cross validation scheme, while the model accuracy was used as an evaluation metric. In order to compare the samples that are generated through the 10-fold cross validation of the five models with one-way ANOVA, the normality assumption was checked using the Shapiro-Wilk test and a significance of a=0.05. The results of the test are illustrated in table I:

TABLE I
SHAPIRO - WILK TEST RESULTS

Feature Selection Method	p***-value***
Forward Selection	0.431
Random Forest	0.146
Chi-Square	0.398
Logistic Regression Coefficients	0.020
Lasso Logistic Regression	0.446

According to the test results, the normality assumption holds

for the four out of the five samples, excluding the method of applying statistical tests on logistic regression coefficients. It's worth noting that Forward Selection, Penalized Logistic Regression and Chi-square correlation resulted in models with more normal-like accuracy distributions, which indicates the selection of more robust feature subsets. In order to check if a larger sample size would resolve the normality issue, the whole process was repeated using a 20-fold cross validation. However, the results of Shapiro-Wilk test continued to indicate divergence from normality. In that regard, a more conservative approach was followed and the alternative of the one-way ANOVA was rejected at this point.

The homoscedasticity of the samples was assessed through the application of Levene's test, which performs better in non-normally distributed samples. The test's results could not reject the null hypothesis with a p-value of 0.06. Consequently, there is evidence of homoscedasticity, but the statistical significance is limited.

Finally, the mean accuracy values of the different 10-fold samples were compared using the Kruskal-Wallis test. The null hypothesis could not be rejected with a p-value of 0.25, indicating that there is no statistically significant difference between the average accuracy of the models that were trained using the different feature selection methods.

V. Conclusions - Discussion

The main motivation behind this study was to investigate if providing psychology experts with a robust end to end methodology to identify causal mechanics in their studies would lead to concrete benefits. Specifically, it is argued that the causal mechanics of interest can be explored through the application and comparison of different feature selection methods. The proposed process involves a mixture of filter, wrapper and embedded feature selection methods as well as a k-fold cross validation scheme and a pipeline of statistical tests.

The above methodology was applied in the domain of sleep quality using the data set created by Norbury and Simon Evans [18]. Although the different feature selection methods provided diversified feature subsets, the employed statistical tests were not able to distinguish between the different feature selection methods at a statistically significant confidence level. When interpreting this result from the point view of machine learning, it is quite unlikely that the same non-penalized logistic regression model would provide so consistent results in terms of accuracy, when trained in so diverse feature subsets. Consequently, it is assumed that the inability of the method to distinguish between the different accuracy samples is associated with the applicability of the employed statistical tests in the specific case.

What is certain is that further studies and meta-analyses are needed to monitor the current situation of computational psychology. We worked on a very specific field of psychology, using prevailing categorical data. As possible future work, we therefore suggest the application of a similar pipeline with the same statistical tests, but applied in a different domain of psychology that uses mostly numerical data. In this way it would be possible to compare the results of the same workflow but applied on different data and consequently on different feature selection techniques. As an alternative, other research teams could experiment with different comparison pipelines to identify the most suitable feature selection method. Such efforts could include the employment of different statistical tests than the ones that were used in this study, as well as the modification of the significance level in order to create a sensitive pipeline, capable of capturing the differencies under consideration. Finally, while our analysis used the logistic regression as a baseline classification task to compare the feature selection methods, the study of H.B.F. David et al (2019) [4] opted for the SVM as the machine learning algorithm of choice. There is the possibility to experiment with different algorithms, as well as potential hyperparameter tuning approaches.

References

[1] Y. Tal and W. Jacob, "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning", Perspectives on Psychological Science, vol. 12, no. 6, pp. 1100-1122, 2017. [Online]. Available: doi:10.1177/1745691617693393

[2] P. B. Sharp and E. Eldar, "Computational Models of Anxiety: Nascent Efforts and Future Directions", Current Directions in Psychological Science, vol. 28, no. 2, 2019. [Online]. Available: doi:10.1177/0963721418818441

[3] P.E. Meehl, "Why Summaries of Research on Psychological Theories are Often Uninterpretable", Psychological Reports, vol. 66, no. 1, pp. 195–244, 1990. [Online]. Available: doi:10.2466/pr0.1990.66.1.195

[4] H.B.F. David, A. Suruliandi and S.P. Raja, "Preventing crimes ahead of time by predicting crime propensity in released prisoners using data mining techniques", International Journal of Applied Decision Sciences, vol. 12, no. 3, pp.307–336, 2019. [Online]. Available: doi:10.1504/IJADS.2019.100433

[5] E. Azad-Marzabadi and S. Amiri, "Morningness–eveningness and emotion dysregulation incremental validity in predicting social anxiety dimensions", International Journal of General Medicine, vol. 10, pp. 275—279, 2017. [Online]. Available: doi:10.2147/IJGM.S144376

[6] C. S. Chan, C. Y. S. Poon, J. C. Y. Leung, K. N. T. Lau and E. Y. Y. Lau, "Delayed school start time is associated with better sleep, daytime functioning, and life satisfaction in residential high-school students". Journal of Adolescence, vol. 66, pp. 49–54, 2018. [Online]. Available: doi:10.1016/j.adolescence.2018.05.002

[7] A. S. Angelika, I. Bihlmaier, M. Hautzinger, M. D. Gulewitsch and B. Schwerdtle, "Nightmares and Associations with Sleep Quality and Self-Efficacy among University Students", 2015.

[8] K. Sullivan, C. Ordiah, "Association of mildly insufficient sleep with symptoms of anxiety and depression", Neurology, Psychiatry and Brain Research, vol. 30, pp. 1-4, 2018. [Online]. Available: doi: 10.1016/j.npbr.2018.03.001

[9] N. Choueiry, T. Salamoun, H. Jabbour, N. El Osta, A. Hajj and L. Rabbaa Khabbaz, "Insomnia and Relationship with Anxiety in University Students: A Cross-Sectional Designed Study". PLoS ONE 11(2): e0149643, 2016. [Online]. Available: doi:10.1371/journal.pone.0149643

[10] H. Chia-Yueh, G. Susan Shur-Fen, S. Chi-Yung, C. Yen-Nan and L. Ming-Been, "Associations Between Chronotypes, Psychopathology, and Personality Among Incoming College Students", Chronobiology International, vol. 29, no. 4, pp. 491-501, 2012. [Online]. Available: doi: 10.3109/07420528.2012.668995,

[11] M. Najafi Kalyani, N. Jamshidi, J. Salami and E. Pourjam, "Investigation of the Relationship between Psychological Variables and Sleep Quality in Students of Medical Sciences". Depress Res Treat. 2017;2017:7143547, Epub 2017 Sep 28. PMID: 29093971; PMCID: PMC5637842, 2017. [Online]. Available: doi: 10.1155/2017/7143547

[12] A. A. Gomes, J. Tavares and M. H. P de Azevedo, “Sleep and Academic Performance in Undergraduates: A Multi-measure, Multi-predictor Approach”,Chronobiology International, vol. 28, no. 9, pp. 786–801, 2011. [Online]. Available: doi:10.3109/07420528.2011.606518

[13] Francis L. Huang (2019) “Alternatives to logistic regression models in experimental studies”, The Journal of Experimental Education, 2011. [Online]. Available: doi:10.1080/00220973.2019.1699769

[14] G. James, “An Introduction to Statistical Learning: with Applications in R”, Springer Texts in Statistics, Springer Science+Business Media New York, 2013. [Online]. Available: doi:10.1007/978-1-4614-7138-7

[15] G. Chandrashekar and F. Sahin, “A survey on feature selection methods”, 2013. [Online]. Available: doi:10.1016/j.compeleceng.2013.11.024

[16] A. Jović, K. Brkić and N. Bogunović, “A review of feature selection methods with applications”, 2015. [Online]. Available: doi:10.1109/MIPRO.2015.7160458

[17] L. Breiman, “Random Forests”, Psychiatry Research, 2001. [Online]. Available: doi:

[18] R. Norbury and S. Evans, “Time to think: Subjective sleep quality, trait anxiety and university start time.”, Psychiatry Research , 2018. [Online]. Available: doi:10.1016/j.psychres.2018.11.054