# Predicting Relationship Status from Emotional, Cognitive, and Behavioral Measures

**Samuel Heinrich**
Matrikelnummer 6972178

## Abstract

Understanding the psychological differences between individuals in and outside of romantic relationships is a long-standing research question. This study aims to evaluate whether relationship status can be accurately predicted based on psychological and cognitive factors. To address this, we analyze data from the *MPI Leipzig Mind-Brain-Body* dataset ($N = 227$), using self-report questionnaires and cognitive assessments to train a Random Forest classifier. Our model achieves a weighted F1-score of 0.72, surpassing a baseline classifier (weighted F1 = 0.52), indicating that relationship status is, to some extent, predictable from psychological data. The feature importance analysis highlights six key predictors, primarily linked to emotion regulation and cognitive performance, with perceived social support from family and significant others, along with anxiety and neuroticism, emerging as key differentiators. These findings suggest that emotional and cognitive factors play a role in romantic involvement, though ethical and methodological considerations, such as privacy concerns and dataset limitations, must be addressed.

## 1 Introduction

Close relationships play a pivotal role in psychological well-being, influencing mental health, emotional support, and social integration [1]. Yet, researchers have traditionally focused on correlational links between emotional regulation, cognitive functioning, behavioral traits, and relationship outcomes [2], rather than on the feasibility of using these variables to *predict* whether someone is in a relationship. This introduces a vital **knowledge gap**: while we know these psychological constructs are important, it is less clear whether they can effectively classify individuals' relationship status at a given moment [3]. To address this problem, we pose the key **research question**:

> *Can an individual's relationship status (in a relationship vs. single) be accurately predicted using emotional, cognitive, and behavioral measures, and do individuals in a relationship exhibit significantly different values on core features compared to those who are single?*

As a **plan of attack**, we leverage a subset of data from the publicly available *MPI Leipzig Mind-Brain-Body* database [4], focusing on self-report questionnaires (e.g., emotion regulation, anxiety inventories) and cognitive tests (e.g., memory, attention). After rigorous data cleaning and normalization, we train a Random Forest classifier to handle potential imbalances in relationship status [5]. We then validate our model using the weighted F1-score — a performance metric well-suited for imbalanced classes [6] — and compare the results to a baseline dummy classifier. Finally, we test our hypothesis by examining whether the most important features (as identified by our model) differ significantly between participants in a relationship and those who are single.

Project Report for *Data Literacy* 2024/25

Our **key contributions** can be summarized as follows:

1. We introduce a practical predictive framework for classifying relationship status from emotional, cognitive, and behavioral assessments.
2. We demonstrate moderate predictive success (weighted F1-score of 0.72) using Random Forest model, while highlighting ethical and methodological considerations (e.g., privacy, measurement errors) [7].
3. We reveal differences in psychological traits between individuals in a relationship and those who are single using hypothesis tests and a game-theoretic approach [8].

## 2 Methods

### 2.1 Data Preprocessing

All relevant files were merged into a single dataset, unifying different data sources that used various Likert scales (e.g. 1-5, 1-7, reversed scales) and differing score reporting methods—some tests reported mean scores of the items for a trait, while others provided summed scores for traits. To standardize the data, all values were converted to mean-based scores. To mitigate the impact of extreme values, particularly those stemming from diverse Likert scales, each feature was transformed using the RobustScaler from scikit-learn [10]. Yes/No items were label-encoded as 0 or 1 to ensure uniform processing across features. To maintain data consistency, participants exceeding an outlier threshold of 0.2 (i.e., more than 20% of responses falling outside a valid range) were removed [9]. Additionally, participants who missed more than two tests were dropped to ensure adequate coverage of key variables. Inconsistent entries, such as ages stored as text, were corrected before analysis.

### 2.2 Modeling Approach

A Random Forest classifier was chosen for its ability to handle mixed data types and for its outputs of significance of characteristics, which aid in interpretability [11]. A five-fold cross-validation grid search was performed to find the optimal hyperparameters [12]. The data set was divided into 80% for training and 20% for testing, with stratification to preserve the proportion of single participants vs. in-relationship. Because the distribution of classes was imbalanced, the weighted F1-score was adopted as the principal metric to balance precision and recall for both classes [13]. A dummy classifier that always predicted the majority class served as a baseline for comparison.

### 2.3 Hypothesis Testing and SHAP values

Two types of statistical tests were applied to assess differences in key features between the relationship and single groups [14]. Initially, an independent-samples t-test was performed for each feature assumed to be normally distributed. When normality assumptions were violated, a Mann–Whitney U test was used instead. Additionally, Shapley Additive Explanations (SHAP) were inspected to illustrate how specific features influenced individual predictions and to confirm their relative importance in the Random Forest model [15].

## 3 Results

### 3.1 Grid Search

The hyperparameters of the Random Forest model were selected based on the weighted F1-score criterion. The optimal configuration balances computational efficiency and predictive performance (Table 1). Additionally, balanced class weights were used to account for class imbalances in the dataset, ensuring fair representation of all classes.

| Hyperparameter | Value |
|---|---|
| max_depth | 3 |
| max_features | None |
| min_samples_leaf | 1 |
| min_samples_split | 20 |
| n_estimators | 150 |

Table 1: Hyperparameters selected by GridSearch using CV for the RF-model.

### 3.2 Model Performance

The Random Forest model demonstrated strong predictive capabilities, achieving a weighted average F1-score of 0.72, which substantially outperforms the Dummy Classifier's score of 0.52 (Table 2). A more detailed examination of class-specific performance reveals a considerable disparity in F1-scores between class 0 and class 1. The model achieves an F1-score of 0.82 for class 1, indicating high predictive accuracy in identifying individuals in a relationship. In contrast, the performance for class 0 is noticeably lower, with an F1-score of 0.50. This discrepancy can be caused by the uneven distribution of training data, where class 0 is underrepresented, although we accounted for class imbalance by adding class weights in the model training.

| Model | Class 0 F1-Score | Class 1 F1-Score | Weighted Avg. F1-Score |
|---|---|---|---|
| Random Forest | 0.50 | 0.82 | **0.72** |
| Dummy Classifier | 0.31 | 0.62 | 0.52 |

Table 2: class individual F1-scores and weighted average F1-score of the Random Forest model and a Dummy Classifier.

Beyond establishing predictive capability, our sub-hypothesis explores how individual characteristics contribute to this classification and whether significant differences exist between individuals in a relationship and those who are not. By identifying the most influential features, we can determine which psychological or behavioral traits are most relevant for distinguishing between these two groups. Significant others' support (MSPSS) emerged as the most important feature (Figure 1), reflecting the critical role of perceived social support from significant individuals. This is logical given that it includes romantic partners, directly distinguishing individuals in a relationship from those who are single. Self-control (TEIQue-SF) was second, highlighting the importance of emotional regulation, stress management, and impulse control in the results of relationships of individuals. Family support (MSPSS) was also highly predictive, showcasing the role of familial support in fostering emotional stability and relational well-being. Personality traits, particularly neuroticism (NEOFFI) and extraversion (NEOFFI), were pivotal; neuroticism reflects emotional instability and susceptibility to distress, whereas extraversion highlights sociability and an energetic, outgoing disposition—traits linked to relational behaviors. Trait anxiety (STAI) further underscored the role of emotional health, indicating a stable predisposition toward experiencing anxiety.

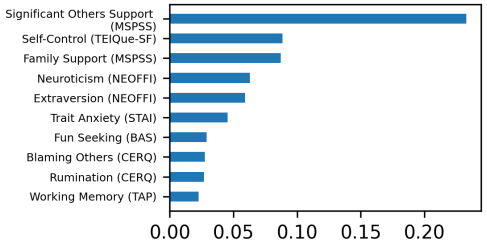


Figure 1: Feature importance scores of the 10 most influential features from the Random Forest model for predicting relationship status.

### 3.3 Hypothesis Tests and SHAP-values

Statistically significant differences ($p < 0.05$) were observed for all features except extraversion (NEOFFI), indicating that emotional support, self-control, and family relationships show significant differences in distribution between individuals in a relationship and those who are single, making them critical in distinguishing relationship status (Figure 2).

This evidence is also supported by the SHAP value analysis, which reveals the nuanced contributions of the most important features to the model predictions of the relationship status (Figure 2). Support from significant others (MSPSS) exerts the strongest positive influence, with higher perceived support (red points) consistently increasing the likelihood of being classified as in a relationship, while lower values (blue points) often result in predictions of being single. Self-control (TEIQue-SF) follows a similar trend, where greater emotional regulation and impulse control positively impact relationship status predictions. Family support (MSPSS) also demonstrates a positive influence, emphasizing its role in fostering relational well-being. Conversely, higher scores on neuroticism (NEOFFI) and trait anxiety (STAI) (red points) negatively affect predictions, as emotional instability and anxiety increase

the likelihood of being single. Lower values for these traits contribute positively to the prediction of being in a relationship. In contrast, extraversion (NEOFFI) has minimal impact, as high and low values result in SHAP values close to zero, suggesting that it plays a less decisive role in model decision making.

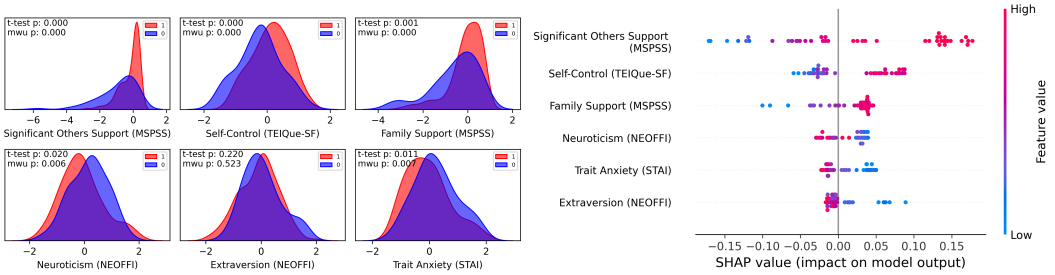


Figure 2: (Left) Feature distributions for the six most important features. Persons in a relationship in red (1) and not in a relationship blue (0). P-values from a t-test and a Mann-Whitney U test ($\alpha = 0.05$) in the upper left corner. (Right) SHAP values for predicting whether a person is in a relationship, highlighting the contributions of the features to the model's predictions. Positive SHAP values indicate a positive impact on the prediction, while negative values indicate a negative impact.

These findings support the hypothesis that people in a relationship exhibit significantly different values in core characteristics compared to those who are single, with stronger social support (MSPSS) and greater self-control (TEIQue-SF) distinguishing them. In contrast, single individuals show higher neuroticism (NEOFFI) and trait anxiety (STAI).

## 4 Discussion and Conclusion

This paper examined whether relationship status (in a relationship vs. single) can be predicted from emotional, cognitive, and behavioral data. The central question was whether these measures would yield a reliable classification of individuals' relationship status. By training a Random Forest model on self-report and cognitive assessments, a weighted F1-score of 0.72 was achieved, indicating that such data can indeed offer moderate predictive power. Emotional regulation and anxiety-related features contributed most to the model's success, aligning with existing evidence that emotional well-being is closely tied to relationship outcomes.

Despite these promising results, several limitations must be acknowledged. First, the cross-sectional design of the study restricts any inference of causality: whether being in a relationship fosters certain emotional or cognitive attributes, or whether these attributes make relationships more likely cannot be determined here. Second, reliance on self-report introduces the possibility of response biases. Third, measurements of subjective constructs (e.g., anxiety) can be prone to error, especially given the diversity of Likert scales and scoring methods. Additionally, the sample size was moderate, and all participants came from a specific dataset collected by the Max Planck Institute, limiting broader generalizability.

Ethical considerations also arise. Predicting a personal attribute like relationship status may be viewed as intrusive if used without explicit consent or appropriate safeguards. Moreover, the fact that these variables show moderate predictive accuracy does not guarantee they should be used for any high-stakes decisions in clinical or interpersonal settings. However, the approach highlights how easily accessible questionnaires and cognitive tests might shed light on underlying factors that distinguish individuals in a relationship from those who are single.

## References


1. Holt-Lunstad, J., Smith, T. B., Layton, J. B. (2010). Social relationships and mortality risk: A meta-analytic review. *PLoS Medicine*, 7(7), e1000316. https://doi.org/10.1371/journal.pmed.1000316
2. Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3), 281-291. https://doi.org/10.1017/S0048577201393198

3. Kahneman, D., Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6), 515–526. https://doi.org/10.1037/a0016755

4. Babayan, A., Erbey, M., Kumral, D., et al. (2019). A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6, 180308. https://doi.org/10.1038/sdata.2018.308

5. Chen, C., Liaw, A., Breiman, L. (2004). Using Random Forest to learn imbalanced data. *University of California, Berkeley*.

6. Saito, T., Rehmsmeier, M. (2015). The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

7. Vitak, J., Shilton, K., Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing (CSCW '16)* (pp. 941–953). Association for Computing Machinery. https://doi.org/10.1145/2818048.2820078

8. Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint* arXiv:1705.07874.

9. Osborne, J. W., Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation*, 9(1), 6. https://doi.org/10.7275/qf69-7k43

10. Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(20150202). https://doi.org/10.1098/rsta.2015.0202

11. Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

12. Bergstra, J., Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.

13. Japkowicz, N., Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

14. Steyerberg, E. W., Harrell, F. E. Jr., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., Habbema, J. D. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8), 774-781. https://doi.org/10.1016/s0895-4356(01)00341-9

15. Shapley, R. (1953). A value for n-person games. In *Contributions to the Theory of Games (AM-28)*, Vol. 2 (pp. 307–317). Princeton University Press.