Data Mining Techniques - Assignment 1 (Group 1)

Arda Ergin (2658918)[1], Julian Ramondo (2785746)[1], and Mika Rosin (2817059)[1]

Vrije Universiteit Amsterdam, Amsterdam, Netherlands

1 Introduction

Mood impacts all aspects of our lives, yet it is experienced internally. In day to day life it is defined as a conscious state of mind [3]. It is long lasting and is not directed towards a specific target [18]. Mood disorders are among the most prevalent mental health conditions, the estimate being 9.6%, to put it into perspective, it is one in 20 persons [19]. This makes it vital to develop interventions on a moment-to-moment basis.

Emotional experience can be broken into two parts: valence and arousal. The valence, a degree to which an emotion is positive or negative [23], of emotions experienced by the patient varies based on the disorder. Arousal ranges from quiet to active, and is thought to have a V-shaped relationship with valence [10]. For example, patients diagnosed with depression will show negative valence and low arousal. However, despite the high prevalence, patients diagnosed with mood disorders often do not receive adequate help. This could be due to barriers associated with mental health treatment, such as a shortage of healthcare professionals [9]. Early interventions can alleviate the symptoms and decrease the risks, and smartphone applications can offer a great solution for monitoring and mood prediction.

Early Warning signals (EWSs) can be used as indicators of systems instability right before a critical transition in psychopathology [6][21]. For example, EWSs were found to predict a sudden mood transition in patients with bipolar disorder [5]. Despite their potential, EWSs still remain an under-researched area. Mobile phones became so entailed in people's lives that it is hard to imagine our day without them. On average people check their phones 58 times per day [8]. Given this, it is a rather unsurprising that researchers will implement phones in their research to replace paper-and-pencil method of collecting mood data [11][14]. Phones can be used to collect EWSs for mood disorders in an efficient manner without having patients come into the hospital. Previous research showed a successful integration of phones to measure mood states [12][20] and physiological factors (e.g. sleep [16]). In this regard, our research attempts predict mood in such a manner to be able to develop interventions.

2 Methods

2.1 Exploring the Data

Before further exploring the data, we first cleaned the dataset in terms of incorrect observations. We have identified a total of 4 observations with negative values that would not be possible for that type of variables. 3 observations of the variable `appCat.builtin`, and 1 observation of the variable `appCat.entertainment` had negative values. Since these observations are not possible, we have removed them. We did not identify duplicate rows in the dataset.

After the removal of the 4 incorrect observations, the dataset contains a total of 376.912 observations from 27 participants. This data has been collected between the dates 2014-02-17 and 2014-06-09, amounting to 112 days. In the dataset, we have identified five different kind of variables: (1) psychological momentary ecological assessment variables of mood, valence, and arousal; (2) physical activity score; (3) screen time; (4) binary variables of call and SMS; (5) app variables regarding the usage of specific app categories, recorded in seconds.

Time Even though 112 days initially seemed to be a substantial time frame, upon further investigation, we have found that the number of observations in the first and last weeks of data collection were extremely scarce (see Figure 1). 24 out of 27 participants have 17-02-2014 as their first observation date, which shows that this scarcity is not caused by an 'early start' by several participants. Until about 13-03, the data presents some systematic unreliability. For instance, for most participants, until about mid-March, the only observations were phone calls received, or mood was logged but no other data was recorded. One possible explanation for this could be that the application was still being developed during this period. In comparison, even though also scarce, the data towards the end of the time period appeared to be more reliable.

Regarding the varying sparsity of data especially at the beginning of the time frame, even though it initially made sense to trim between certain dates, we have opted to trim the time frame individually for every individual. This is important since global trimming based on dates resulted in significant and unnecessary data loss. For each individual, we have manually identified the date were they *meaningfully* started recording data, as well as the date were they stopped *meaningfully* recording. Two of the group members independently conducted this analysis to establish reliability, subsequently comparing and discussing the results.

Psychological Variables We have assumed that the variables of *mood*, *arousal*, and *valence* were collected through **momentary assessment**. This assumption was based on the fact that their times of measurement were very precise (e.g., 12:00:00), as well as the regularity in the times of the day they were measured (see the bottom row in Figure 2). We have identified these set times as 9:00, 12:00, 15:00, 18:00, 21:00. And considering the identical frequency of observations in

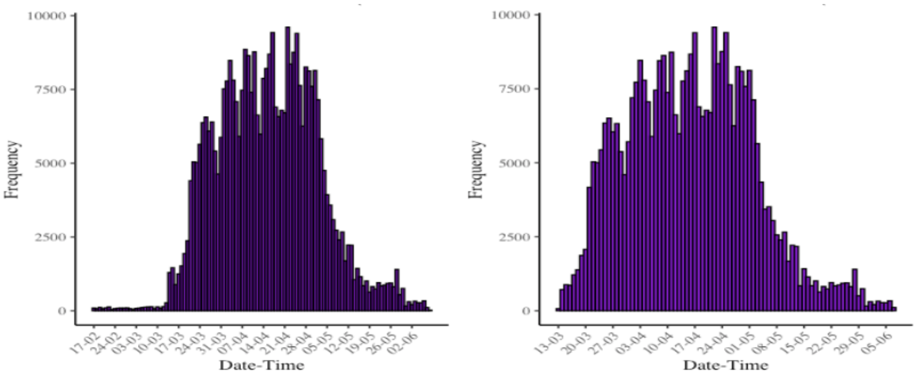

Fig. 1. The distribution of observations from 2014-02-17 to 2014-06-09 (left), and the time-based distribution resulting from restricting time intervals for each subject to meaningful sections that contained sufficient data.

mood, arousal, and valence, we also assumed that these variables were measured together. We have additionally observed that low responses to `mood` or `valence` were very rare, and the distribution was centered around 7 and 1 respectively. In comparison, `arousal` was more uniformly distributed.

Upon investigating individual statistics, and plotting the mood over time for each participant (e.g., see 3), we have observed that, overall, participants were quite similar in their fluctuations, and their overall mood. However, we have noticed two participants, 7 and 33, who have relatively more fluctuations, and could be considered outliers in terms of their standard deviation of mood. However, considering the fact that individuals who have mood disorders have higher mood fluctuations, and if we assume that this sample is a representative sample, we should be observing about $1-3$ individuals considering the population prevalence of mood disorders [19]. Hence, removing these participants would make the sample less representative. More importantly, these individuals would be the one that benefits from 'mood-interventions' the most, e.g., through an app. Hence, we opted to keep them in the dataset.

We have also explored the behaviour of mood based on different times: average mood per day of the week and week of the year. For instance, we could immediately notice graphically that participants had a higher mood on Saturday compared to the rest of the days, and the rest of the days did not significantly differ among each other in terms of average mood (see Figure 5).

Recorded Variables The rest of the variables are recorded automatically by phone. When investigated, they show the usual distribution behaviour of count data. How we handled them is further discussed below.

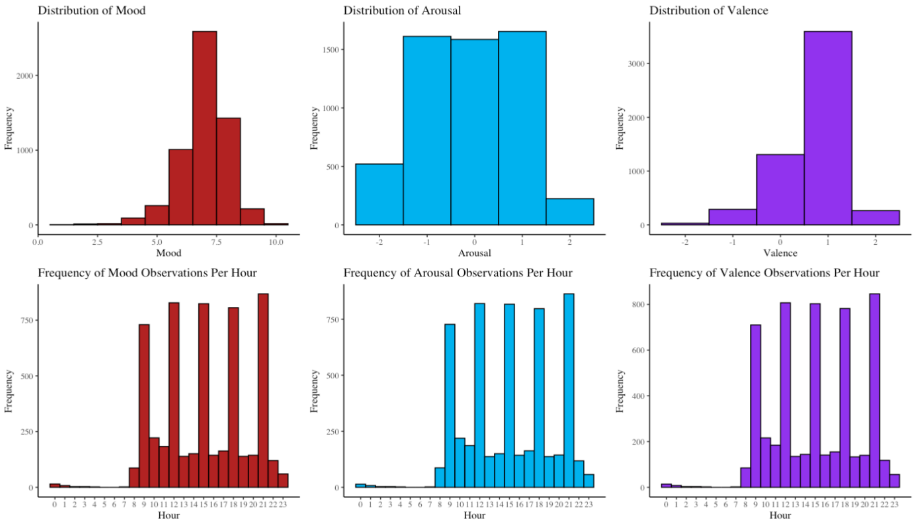

Fig. 2. Distribution of Observations for Mood, Arousal, and Valence.

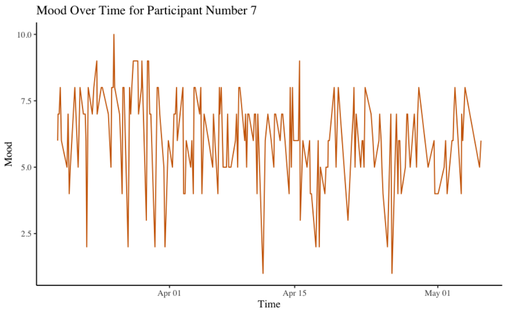

Fig. 3. Mood over time for participant 7. Compared to the other participants it shows much more fluctuations.

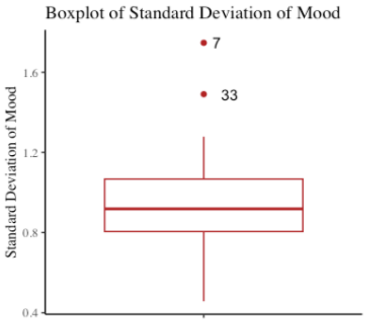

Fig. 4. Boxplot of Standard Deviation of Mood. Two outliers.

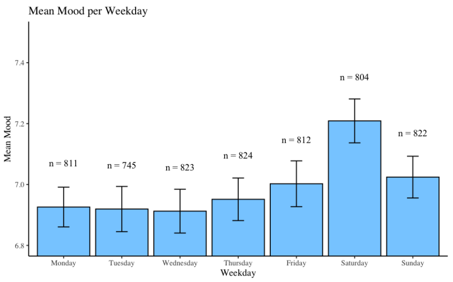

Fig. 5. Average mood per day of the week, with 95% CI and N annotated.

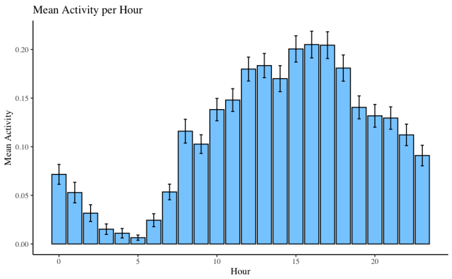

Fig. 6. Average 'activity' over the hours of a day, with 95% CI.

2.2 Data Engineering

Feature Engineering 1 In our study, we decided to first aggregate mood measurement data at fixed times ("beeps") throughout the day, as commonly done in ecological momentary assessment studies (see Figure 2) . For example, a measurement at 11:24 was grouped into the 12:00 beep period.

From this aggregation, we have further derived our 'night-time inactivity' variable. Various previous research indicates that mood is heavily affected by sleep [22], and, hence, we have added this as an appropriate feature. For this, we used a combination of inactivity (see Figure 6 for the distribution of average activity over hours), and the difference between observation points. We also combined all `appCat` variables to assess total phone use, influencing next day's mood, and excluded data between 21:00 - 06:00 to better align with individual sleep patterns rather than using a fixed `date` cutoff.

Additionally, we included time-based features like the day of the week and week of the year, and moving averages over various periods. For deep learning models, we encoded these time-based features using sine and cosine to preserve their cyclical nature, ensuring continuity at the start and end of cycles.

Dealing with Extreme Values and Bad Distributions Our analysis of the box plots revealed numerous outliers in the count variables, which we identified as valid data points — e.g., an individual playing a game on their phone for three hours. To manage these outliers, we applied a logarithmic transformation. For some of the variables, the logarithmic transformation resulted in a normal distribution with zero-inflation. For these variables, we have created a new variable with binary-coding the zeros, as information for the classification and regression models. For the other variables that did not produce a meaningful distribution, we have converted them to categorical variables appropriately. For instance, converting `sms` to three categories of 0, 1, and > 1, produced meaningful results. No major influential outliers were detected in the psychological variables.

Imputation In evaluating imputation methods for missing data (mostly present on mood, arousal, valence, and activity), we focused on the KNN Imputer and Iterative Imputer MICE (Multiple Imputation by Chained Equations). The KNN Imputer substitutes missing values with the mean of the nearest neighbors, effective in simpler, smaller datasets but limited by computational demand and sensitivity to outliers as dataset complexity increases. In contrast, MICE employs multiple regressions, imputing each missing feature based on the others in a round-robin approach, making it suitable for complex datasets with varied data types. Despite being computationally intensive, MICE's ability to handle intricate data patterns makes it preferable for preserving data integrity in complex scenarios. In our tests MICE outperformed KNN. Given these advantages and the complex interdependencies in our dataset, we selected MICE over KNN.

Even though we have carried out the analysis based on a day-aggregated data format, we have carried out the imputation process with a beep-aggregated data format (i.e., five observations per day). We believe that this format allowed for more variability and individual patterns throughout the day, which we believe to enrich the data. To avoid data leakage, instead of running a MICE on the test set, we have simply imputed the mean values we had for our training set.

Additionally, we have also considered the meaningfulness of missing responses in this dataset — that, if an individual misses a 'beep', it might be because of certain factors. Hence, for the classification, we have also tested with categorizing missing responses to mood as a separate category. However, based on the performance of the model, we have decided to not take this approach.

2.3 Network Analysis and Feature Selection

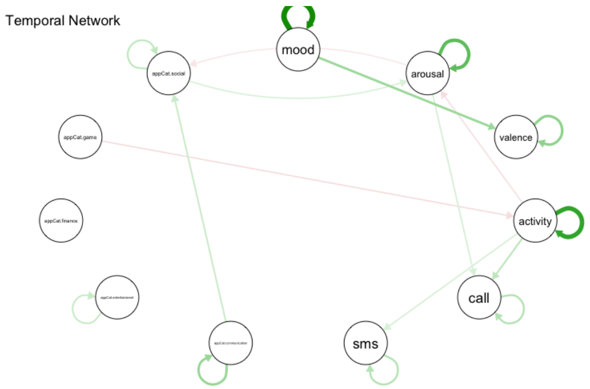

Fig. 7. Temporal Network of our data, only significant nodes are displayed.

In order to understand, and visualize, the relationship between the variables, we have conducted a network analysis using Dynamical Structural Equation Modeling (DSEM), which has been frequently used in other similar psychological studies [4]. This approach also allows us to account for the multi-level nature of our data. In this regard, we have used the `mlVAR` package in R, which provides tools for estimating multilevel vector autoregression models within the DSEM framework. This way, we could explore both cross-sectional and temporal interdependencies among our variables. From this analysis, we have produced the temporal network on Figure 7.

We have taken the results of this analysis into account for feature selection. First, we could see the significant variables and their interactions, and this informed us about the variables that are a good predictor. Secondly, when building the DSEM model, we could notice the variables that are highly correlated with other variables (e.g., screen, activity), and take them out. Additionally, we have realized that that the `appCat.other` mostly reflected the app where mood, valence, and arousal was entered, based on its distribution.

2.4 Train-Test Split

For the Train-Test Split, we considered two methods: a temporal 20/80 split at the end of the data or a split between participants. Given the varying start and end times among participants and potential data leakage issues, we chose to split the data based on participants. This approach aligns better with the research's goals, particularly in contexts where the aim is to develop an intervention application that predicts mood dips across a broader population. This participant-based split enhances the model's generalizability to new individuals, which is vital for effective interventions.

2.5 Feature Engineering 2

In optimizing our dataset for time-series analysis, we explored various window sizes N for mood, and aggregation methods to determine the best approach for predicting next-day mood. We tested different lengths of 'historical' mood to find the ideal window that maximizes predictive accuracy.

Methods of Data Aggregation:

1. **Direct Inclusion of Historical Data:** We appended columns to the actual values in which we entered the information about the past N days, maintaining day-specific details of the last day.
2. **Aggregated Historical Summary:** We also tried summarizing the past N days without keeping the information of the last day, which streamlined the data but potentially diluted the impact of recent days.
3. **Weighted Mean Aggregation:** This approach used weighted averages that prioritized recent days over earlier ones, hypothesizing that recent behaviors have a stronger influence on the next day's mood.

Despite experimenting with these diverse methods, we did not find a statistically significant difference in their predictive performances. Consequently, we chose the direct inclusion method, which adds historical data to the day's data. We selected this method as it retains the highest amount of information. After experimentation and fine-tuning, we determined that using a window size of $N = 3$ was optimal. This is possibly because the older history is not relevant anymore and only dilutes the relevant information.

To ensure the integrity and relevance of our data, we implemented a function that specifically handled the selection and preparation of sequences based on the chosen window size of $N = 3$ days. This function was designed to retain only those data segments that had at least $N + 1$ consecutive days, ensuring not only the availability of sufficient historical data to calculate the necessary features but also the presence of a subsequent day to serve as the prediction target.

The function systematically filtered out any sequences that did not meet this criterion, including those with gaps that might disrupt the continuity necessary for accurate mood prediction. Before deploying the classification algorithm, we also checked for correlations among the variables to identify any significant relationships that could influence our model's accuracy. By doing so, it guaranteed that each dataset entry used in training and testing the model was both complete and consecutive, providing a robust basis for generating reliable predictive insights into next-day mood fluctuations. This methodical approach helped in maintaining a high-quality dataset, optimized for deriving meaningful and actionable predictions from the model.

2.6 Classification with Machine Learning Approach

Our classification approach centered on predicting the *mood_of_next_day*, which we discretely labeled into three categories: low (1.0-6.999), medium (7.0-7.999), and high (8.0-10.0). This categorization enabled the effective application of various machine learning algorithms. We employed two primary predictive models: the **K-Nearest Neighbors (KNN) Classifier**, chosen for its ability to handle non-linear feature relationships, with hyperparameter tuning of *n_neighbors* from 1 to 30 to achieve optimal balance; and **Ensemble Methods**, incorporating a **Random Forest Classifier (RFC)** with tuned *n_estimators* from 1 to 200, and a **Voting Classifier** that merged Logistic Regression, Random Forest, and SVC, all standardized with `StandardScaler`. This ensemble utilized a soft voting mechanism, enhancing accuracy by leveraging the strengths of diverse algorithms.

In our study, models were carefully trained using data split into training and testing sets, which was crucial for ensuring they could generalize well across different sets of data. We evaluated model performance rigorously, employing metrics such as accuracy, precision, and recall. This meticulous process allowed us to finely tune the hyperparameters, enhancing each model's performance.

The application of these models illuminated the substantial value of historical data for predicting mood, with both the K-Nearest Neighbors (KNN) and various ensemble models, especially the Voting Classifier, proving particularly effective. These methods showed great promise for real-time mood monitoring applications.

Furthermore, to assess the stability and reliability of our predictions, we implemented k-fold cross-validation—specifically partitioning the data by participants. This method involves dividing the data into k subsets, or "folds," and systematically using one fold for testing while the others are used for training, rotating until each fold has served as the test set. This approach revealed that our models were robust and performed consistently across different subsets of data.

All classification models, data splitting functions, and k-fold cross-validation processes were implemented using the `scikit-learn` library [**?**], which provided robust tools for our analytical needs.

2.7 Classification with DL approach

For the classification with some temporal based algorithms, a long short term memory (LSTM) [7] based recurrent neural network was deployed. This model is appropriate for our data since it can train on arbitrarily long sequences (windows) of the data. It usually performs better than other recurrent neural networks like Elman-Networks for example, since the LSTM has a more complex internal cell structure. The LSTM-cell is build around a cell state that encodes the previously seen data which is used to give the output for the current input. For each time step the cell state is combined with the last cell state, the last prediction and the current input. How exactly the individual elements are weighted is learned by the model itself by so-called gates. The LSTM-based model was optimized using Bayesian optimization [15], which is a form of hyper-parameter tuning. The values being tuned can be seen in table 1. The basic idea of the model architecture is to have a number of dense layers before the LSTM-cell, those layers serve the purpose of encoding the raw input into a format more suitable for the LSTM-cell. Then, after the LSTM combined the temporal steps, the following layers should transform the final LSTM output into the correct classes. The data was transformed using windowing into sequences of 7 consecutive days.

Parameters	Search Range	Type	Best Value Classification	Best Value Regression
Batch size	[1, 1000]	Int	917	428
Learning Rate	[0, 0.02]	Float	0.0001333	0.0079
LSTM Units	[32, 512]	Int	250	81
Encoding Layers	[1, 4]	Int	4	3
Classification Layers	[1, 4]	Int	1	1
Dropout Rate	[0, 0.6]	Float	0.592	0.25

Table 1. Optimized hyper-parameters with search range and best found values for the LSTM based classification and regression models after 50 search steps.

2.8 Regression with Machine Learning approach

Building upon our classification approach, we extended our analysis to regression techniques to predict the continuous variable *mood_of_next_day*. Our regression analysis employed several models known for their efficacy in handling complex data relationships: the **Gradient Boosting Regressor**, which constructs an additive model in a forward stage-wise fashion for managing non-linear relationships; the **Random Forest Regressor**, which reduces variance through averaging multiple deep decision trees trained on different parts of the training set; and **Linear Regression**, valued for its high interpretability by assessing linear relationships between features and the target. These models were integrated into a **Voting Regressor** that averages their predictions to provide a balanced outcome.

We standardized features to normalize their scales and split the dataset into training and test subsets for model training and evaluation. The performance of these models was quantitatively assessed using **Mean Absolute Error (MAE)**, which indicates the average magnitude of errors, and **R-squared**

(R2) Score, which measures the proportion of variance in the mood scores that is predictable from the features. This approach allowed us to directly compare the performance of each model alongside the voting regressor.

Regression modeling provides a granular view of mood predictions compared to classification, offering finer interpretations of mood variations. Unlike classification, which categorizes mood into predefined states, regression allows for continuous outcomes, enhancing our understanding of predictive dynamics and improving the precision of mood assessments.

Also the regression models, data normalization processes, and model evaluations were facilitated using the `scikit-learn` library [**?**].

2.9 Regression with DL approach

For regression with temporal models, a similar approach to the classification was used. The only difference to the classification LSTM based model, is that the final layer is a dense layer with a single node, returning the prediction mood score directly. As a loss function the MSE was used. Hyper-parameter optimization was done using Bayesian optimization, the optimized values were the same as in the optimization for the classification model. The optimized values and corresponding best found values can be found in table 1.

2.10 Evaluation

As evaluation metrics we can use both the mean squared error (MSE) or the mean absolute error (MAE). Both of them have distinct characteristics, making their usage dependent on what should be achieved or investigated. The formulas can be found in equations 1 and 2.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \tag{1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - p_i| \tag{2}$$

As we can see the MAE is the average absolute error from the predictions made. This is useful since the unit is the same as in the original data. This metric is also more robust to outliers since they are "averaged out". This can also be a problem if the data is skewed, since the values further away from the median might get lost in the MAE, which can lead to overfitting of the data.

The MSE on the other hand considers the mean of the squared differences, therefore points that are predicted with a larger error gain more weight in the metric. This can be an issue if there a certain outliers in the data, but can also be beneficial if the distribution is skewed, because the model will be less likely to only predict the median or mean, since the values with greater distance to them would have a higher importance with an increased error. Thus, the MSE will in general lead to less big errors.

Conclusively it can be said that if less large errors are wanted the MSE is more suitable, and if the data contains some outliers that should be ignored the MAE could come in handy.

There are situation in which MAE and MSE will give the same results. One simple example would be if the errors are always one. A small proof can be seen in equation 3.

$$\begin{aligned} MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \overset{!}{=} \frac{1}{N}\sum_{i=1}^{N}(\pm 1)^2 \\ = \frac{1}{N}\sum_{i=1}^{N} 1 = \frac{1}{N}\sum_{i=1}^{N}|\pm 1| \overset{!}{=} \frac{1}{N}\sum_{i=1}^{N}|y_i - p_i| \end{aligned} \tag{3}$$

The equation holds under the condition that the error is one, since then $(\pm 1)^2 = 1 = |\pm 1| = |y_i - p_i|$, where $i \in \{1, ..., N\}$ and y_i and p_i being the actual and the predicted value.

Winning Classification Approaches

We will present the winning solution of the ICR - Identifying Age-Related Conditions Challenge which was a classification competition on Kaggle that started at the 11th of May in 2023 and ended the same year on August the 11th [1]. The aim was to perform a binary classification to identify if participants had at least one of three medical conditions, or if they had None. The data used for prediction contains 56 (anonymized) medical characteristics of participants. For evaluation a balanced logarithmic loss was used, this ensures that all classes are equally important in the evaluation. The given balanced loss can be seen in equation 4:

$$LogLoss = \frac{-\frac{1}{N_0}\sum_{i=1}^{N_0} y_{0i}\log p_{0i} - \frac{1}{N_1}\sum_{i=1}^{N_1} y_{1i}\log p_{1i}}{2} \tag{4}$$

Where $N_c, c \in \{0, 1\}$ is the number of observations in class c and y_{ci} is 1 if observation i belongs to class 1, otherwise it is 0. p_{ci} is the predicted probability of observation i belonging to class c.

The winner of this competition was a user with the name ROOM722 [2]. The main technique used was a Deep neural network, namely a Temporal Fusion Transformer (TFT). These types of models are capable of multi horizon forecasting, i.e. predicting multiple steps ahead for time series with a single prediction. The distinctiveness of the TFT is that it integrates static covariates encoders that enable usage of static variables for prediction, a gating mechanism (inspired by the LSTM gates) that allows sample dependent variable selection and a ”temporal self-attention decoder to learn any long-term dependencies present within the dataset” [13]. By analyzing the variable selection mechanism the authors claim that is is also possible to derive patterns in the data that lead to certain classifications, i.e. the model allows for global explainability.

What is a little unconventional is that the winner trained models on K-Fold-Crossvalidation (CV) subsets of the data, 10 to 30 times per fold, selecting the two best models for each fold. The selected models of all folds were then used to predict the classification probabilities. The probabilities are re-weighted and averaged to handle class imbalances. Therefore the competitioner used an ensemble approach with models trained on the CV datasets. Another interesting addition was to convert the problem into a multi-label prediction problem. Initially a baseline DNN was trained. The resulting differences of predicted class probabilities to the actual class probabilities were then used to give each instance a hardness label. This was a binary label being 1 or "hard" if ($y_{1i} = 1$ and $p_{1i} < 0.2$) or ($y_{0i} = 1$ and $p_{1i} > 0.8$), otherwise the "not hard" label (0) was assigned.

What makes the approach stand out is 1. its use of a fairly advanced type of Neural Network that allowed for accurate predictions automatically weighing the input features, reducing the need for feature selection, 2. The conversion of the problem into multi-class problem, which might help with accuracy since the classification can additionally focus on samples that were harder to learn (with the initial model) and 3. The use of K-Fold-Cross validation to train multiple models to use in an ensemble approach. This typically reduces over-fitting, as we have seen it in the lecture.

3 Association Rules

To group raw features into higher level features, for example by adding a pizza feature to individual pizza items, the k-nearest neighbors (KNN) algorithm might be used. Product categorization involves representing each product by a feature set and using the similarity in this feature space to classify products. The process could start with selecting expressive features that define the products, such as specifications and user ratings. KNN operates by comparing a new or uncategorized product against all others in a pre-prepared dataset of categorized products. It classifies the product based on the majority category among its closest neighbors, determined by the chosen distance metric. KNN is simple to implement and intuitive to understand, having enough flexibility to be tailored to specific categories by chosing relevant features. However, KNN is computationally intensive, especially for large datasets, as it calculates distances to all training points for each classification. Its performance is highly sensitive to the choice of features; irrelevant or excessive features can distort 'closeness' and degrade categorization quality. The number of neighbors, k, also impacts performance: a small k may be noise-sensitive, while a large k could obscure category distinctions. Additionally, KNN struggles in high-dimensional spaces due to the curse of dimensionality, where distances become less meaningful. Despite these challenges, KNN can be effective in scenarios with well-separated categories and manageable data dimensions.

Results

The training train and validation loss of classification and regression models can be seen in figure 8and 9. We omitted the confusion matrix for the LSTM-based model since it always predicted the same class. Accuracy, precision and recall all had the same value of 0.5045. The values are the same since they are averaged over all classes: the class representing medium mood has a recall, accuracy and precision score of one, while all others have scores of 0. Therefore the scores is the (weighted) average of these metrics.

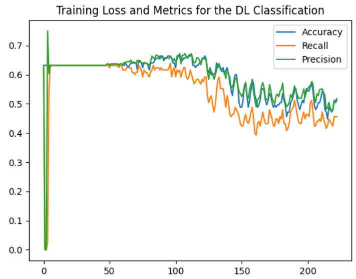

Fig. 8. Validation accuracy and metrics for the classification model.

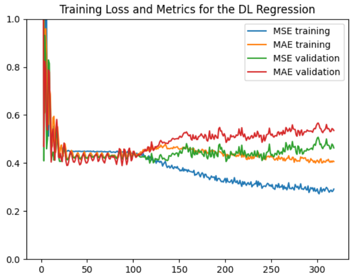

Fig. 9. Training and validation loss of the LSTM based regression model.

The results of the regression models can be seen in were an MSE score of 0.5758 and an MAE score of 0.5626.

ML approach results

The K-Nearest Neighbors (KNN) classifier achieved an accuracy of 59.75% with the best performance at 29 neighbors, while the Random Forest Classifier demonstrated a slightly better accuracy of 61.44% with 65 estimators. The ensemble Voting Classifier improved overall accuracy slightly to 60.59%. In regression, the Voting Regressor showed promising results with an R-squared (R2) score of 0.66, indicating it could explain about 66% of the variance in mood scores, with a Mean Absolute Error (MAE) of 0.32.

Additionally, as an exploratory analysis, we have tried to predict ’the mood of next beep’. Instead of day-to-day prediction, moment-to-moment prediction might be very useful. For instance, previous research has done such beep-to-beep prediction and intervention for suicidal ideation [17]. We have followed the same process, but the model performed significantly worse, only achieving about 55% accuracy for classification.

Discussion

Deep Learning models

he LSTM-based model tended to predict a single value, likely due to optimizing for the validation loss on data from four distinct subjects not included in the training set. This homogeneity in predictions resulted from the model's strategy to minimize validation loss, predominantly predicting the majority class value, which was ”medium mood” for over 60% of the validation data. Consequently, the accuracy on the test set was about 50%, slightly better than random guessing in a three-class mood scale problem. The regression model performed better. Although the model did not generalize perfectly to the test data, evidenced by an MSE of 0.5758 and an MAE of 0.5626, it performed better in the regression task. This improved performance in regression could be attributed to the ordinal nature of the mood scale used, which allows the regression model to approximate the actual mood values more closely than the classification approach, which cannot utilize the ordinal relationship between mood classes.

Impact of MSE vs MAE

Switching from using MSE to MAE as the loss function in training an LSTM-based model resulted in a lower MAE (from 0.5626 to 0.5511) but a significantly higher MSE (from 0.5758 to 0.6404). This outcome supports the previously discussed behavior, where MAE, focusing on minimizing average errors, may not penalize outliers heavily, thereby increasing the MSE. This suggests the presence of outliers in the dataset which the MAE-based model might not handle as effectively as the MSE-based model. A test on two data points of the test data with mood values of 5 and 8 data showed slightly better performance by the MSE-based model for both mood scores, but this was not enough to conclusively determine overall outlier prediction accuracy. Nonetheless, this finding is supported by the overall lower MSE score of the model trained with an MSE loss.

3.1 ML discussion

The results obtained with the ML approach suggest that while the classification models provide reasonable predictive capabilities, ensemble methods slightly enhance performance, highlighting the potential benefits of combining different modeling approaches for psychological data. The regression results further reinforce the effectiveness of ensemble methods in capturing the nuances of mood variability, suggesting a robust framework for mood prediction in psychological studies.

References

1. ICR - Identifying Age-Related Conditions, https://kaggle.com/competitions/icr-identify-age-related-conditions
2. ICR_adv_model, https://kaggle.com/code/room722/icr-adv-model
3. Definition of MOOD (Apr 2024), https://www.merriam-webster.com/dictionary/mood
4. Borsboom, D., Deserno, M.K., Rhemtulla, M., Epskamp, S., Fried, E.I., McNally, R.J., Robinaugh, D.J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.M., Wysocki, A.C., van Borkulo, C.D., van Bork, R., Waldorp, L.J.: Network analysis of multivariate data in psychological science. Nature Reviews Methods Primers **1**(1), 1–18 (Aug 2021). https://doi.org/10.1038/s43586-021-00055-w, https://www.nature.com/articles/s43586-021-00055-w, publisher: Nature Publishing Group
5. Bos, F.: Ecological momentary assessment as a clinical tool in psychiatry: promises, pitfalls, and possibilities. University of Groningen, [Groningen] (2021). https://doi.org/10.33612/diss.177817937
6. Dablander, F., Pichler, A., Cika, A., Bacilieri, A.: Anticipating critical transitions in psychological systems using early warning signals: Theoretical and practical considerations. Psychological Methods **28**(4), 765–790 (2023). https://doi.org/10.1037/met0000450, place: US Publisher: American Psychological Association
7. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://direct.mit.edu/neco/article/9/8/1735-1780/6109
8. Howarth, J.: Time Spent Using Smartphones (2024 Statistics). Exploding Topics (Dec 2023), https://explodingtopics.com/blog/smartphone-usage-stats
9. Kline, A.C., Klein, A.B., Bowling, A.R., Feeny, N.C.: Exposure Therapy Beliefs and Utilization for Treatment of PTSD: A Survey of Licensed Mental Health Providers. Behavior Therapy **52**(4), 1019–1030 (Jul 2021). https://doi.org/10.1016/j.beth.2021.01.002, https://www.sciencedirect.com/science/article/pii/S0005789421000022
10. Kuppens, P., Tuerlinckx, F., Russell, J.A., Barrett, L.F.: The relation between valence and arousal in subjective experience. Psychological Bulletin **139**(4), 917–940 (2013). https://doi.org/10.1037/a0030811, place: US Publisher: American Psychological Association
11. La Porte, L.M., Kim, J.J., Adams, M.G., Zagorsky, B.M., Gibbons, R., Silver, R.K.: Feasibility of perinatal mood screening and text messaging on patients' personal smartphones. Archives of Women's Mental Health **23**(2), 181–188 (Apr 2020). https://doi.org/10.1007/s00737-019-00981-5, https://doi.org/10.1007/s00737-019-00981-5
12. Lieberman, D.Z., Kelly, T.F., Douglas, L., Goodwin, F.K.: A randomized comparison of online and paper mood charts for people with bipolar disorder. Journal of Affective Disorders **124**(1), 85–89 (Jul 2010). https://doi.org/10.1016/j.jad.2009.10.019, https://www.sciencedirect.com/science/article/pii/S0165032709004790
13. Lim, B., Arik, S.O., Loeff, N., Pfister, T.: Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting (Sep 2020), http://arxiv.org/abs/1912.09363, arXiv:1912.09363 [cs, stat]

14. Liu, Y.S., Hankey, J., Lou, N.M., Chokka, P., Harley, J.M.: Usability and Emotions of Mental Health Assessment Tools: Comparing Mobile App and Paper-and-Pencil Modalities. Journal of Technology in Human Services **39**(2), 193–211 (Apr 2021). https://doi.org/10.1080/15228835.2021.1902457, https://doi.org/10.1080/15228835.2021.1902457, publisher: Routledge _eprint: https://doi.org/10.1080/15228835.2021.1902457
15. Nogueira, F.: Bayesian Optimization: Open source constrained global optimization tool for Python (2014), https://github.com/bayesian-optimization/BayesianOptimization
16. Ortiz, A., Maslej, M.M., Husain, M.I., Daskalakis, Z.J., Mulsant, B.H.: Apps and gaps in bipolar disorder: A systematic review on electronic monitoring for episode prediction. Journal of Affective Disorders **295**, 1190–1200 (Dec 2021). https://doi.org/10.1016/j.jad.2021.08.140, https://www.sciencedirect.com/science/article/pii/S0165032721009459
17. Rath, D., de Beurs, D., Hallensleben, N., Spangenberg, L., Glaesmer, H., Forkmann, T.: Modelling suicide ideation from beep to beep: Application of network analysis to ecological momentary assessment data. Internet Interventions **18**, 100292 (Dec 2019). https://doi.org/10.1016/j.invent.2019.100292, https://www.sciencedirect.com/science/article/pii/S2214782919300727
18. Sedikides, C.: Mood as a determinant of attentional focus. Cognition and Emotion **6**(2), 129–148 (Mar 1992). https://doi.org/10.1080/02699939208411063, https://doi.org/10.1080/02699939208411063, publisher: Routledge _eprint: https://doi.org/10.1080/02699939208411063
19. Sekhon, S., Gupta, V.: Mood Disorder. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2024), http://www.ncbi.nlm.nih.gov/books/NBK558911/
20. Thompson, W.K., Gershon, A., O'Hara, R., Bernert, R.A., Depp, C.A.: The prediction of study-emergent suicidal ideation in bipolar disorder: a pilot study using ecological momentary assessment data. Bipolar Disorders **16**(7), 669–677 (2014). https://doi.org/10.1111/bdi.12218, https://onlinelibrary.wiley.com/doi/abs/10.1111/bdi.12218, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bdi.12218
21. Wichers, M., Groot, P.C., Psychosystems, EWS Group, E.G.: Critical Slowing Down as a Personalized Early Warning Signal for Depression. Psychotherapy and Psychosomatics **85**(2), 114–116 (Jan 2016). https://doi.org/10.1159/000441458, https://doi.org/10.1159/000441458
22. Wong, M.L., Lau, E.Y.Y., Wan, J.H.Y., Cheung, S.F., Hui, C.H., Mok, D.S.Y.: The interplay between sleep and mood in predicting academic functioning, physical health and psychological health: A longitudinal study. Journal of Psychosomatic Research **74**(4), 271–277 (Apr 2013). https://doi.org/10.1016/j.jpsychores.2012.08.014, https://www.sciencedirect.com/science/article/pii/S002239991200219X
23. Yik, M., Mues, C., Sze, I.N.L., Kuppens, P., Tuerlinckx, F., De Roover, K., Kwok, F.H.C., Schwartz, S.H., Abu-Hilal, M., Adebayo, D.F., Aguilar, P., Al-Bahrani, M., Anderson, M.H., Andrade, L., Bratko, D., Bushina, E., Choi, J.W., Cieciuch, J., Dru, V., Evers, U., Fischer, R., Florez, I.A., Gararsdóttir, R.B., Gari, A., Graf, S., Halama, P., Halberstadt, J., Halim, M.S., Heilman, R.M., Hřebíčková, M., Karl, J.A., Knežević, G., Kohút, M., Kolnes, M., Lazarević, L.B., Lebedeva, N., Lee, J., Lee, Y.H., Liu, C., Mannerström, R., Marušić, I., Nansubuga, F., Ojedokun, O., Park, J., Platt, T., Proyer, R.T., Realo, A., Rolland, J.P., Ruch, W., Ruiz, D., Sortheix, F.M., Stahlmann, A.G., Stojanov, A., Strus, W., Tamir, M., Torres,

C., Trujillo, A., Truong, T.K.H., Utsugi, A., Vecchione, M., Wang, L., Russell, J.A.: On the relationship between valence and arousal in samples across the globe. Emotion **23**(2), 332–344 (2023). https://doi.org/10.1037/emo0001095, place: US Publisher: American Psychological Association