Course 02445 at DTU 21/06/2024 Jonathan Tybirk, s216136

Task 1: paper evaluation

Paper 1: Automatic movie ratings prediction using machine learning
Link to PDF: https://www.csc.kth.se/~miksa/papers/AutomaticMovieRatingsPrediction_MIPRO.pdf

The paper assesses how well different regression models can predict a given user's IMDb rating of a movie based on 3 different methods:

1. Content-based methods, where the user has an individualized model trained on their previous ratings of movies. The actors, screenwriters, directors, and genre of the movies are features of the model, and the rating is the target variable.
2. Collaborative methods, where the ratings of 'similar' users are used to predict the given user's rating.
3. Hybrid methods, specifically the SVD-kNN where the ratings of 'similar' users on movies with similar features are used to predict the given user's rating on the given movie.

Models were evaluated with leave-one-out cross validation, and from this, the root mean square error as well as mean standard deviation were calculated.
The model performances as well as that of a baseline model that always predicts the mean were displayed in Table 1. This is almost fully appropriate, but unfortunately it also seems that the same dataset was previously used to tune the hyperparameters of some of the models with k-fold cross-validation. This is not fully appropriate, as the hyperparameter combinations of the chosen models were specifically chosen *because* they scored well on the test data, leading to overoptimistic evaluation. Ideally, two-layer cross validation would have been performed to prevent this data leakage. This is what we were taught in our machine learning course, but there is rarely a mention of separate validation and test sets in the papers I could find.

Paper 2: Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms
Link to PDF: https://www.warse.org/IJETER/static/pdf/file/ijeter117852020.pdf

This paper evaluates many different models (all classifiers expect one regressor) seemingly on their ability to predict whether individuals are confirmed to have COVID, as well as if they died or recovered from it. It is not entirely clear to me from the paper, and it not stated in either the abstract or introduction.

The hyperparameters of the models are not disclosed either - for example, the accuracy and r^2 of a kNN-model is displayed in Table 2, but the value of k is never stated. A simple hold-out method is used to split between a train and test set, which might be appropriate for large enough data sets, as the 30% in the test set can be expected to be representative of the whole.
The models are not compared to a baseline, which is especially troubling, as one might expect big class imbalances in a dataset like this where many people are tested for a virus. Simply testing the models on accuracy might therefore also be inappropriate, as simply guessing the majority class might give a high accuracy but have bad recall, which is important when trying to find out who is sick.
There is also no mean standard deviation calculated, so it is difficult to assess the uncertainty of the results.

The paper should be a lot clearer about its methods and describe its datasets and what the target variables are a lot more clearly. A baseline should be trained, and other metrics like recall or the F1 be calculated as well. At the least, a standard deviation should be calculated, and ideally also significance tests like pair-wise t-tests, ANOVA, or Kruskal-Wallis.

Task 2: model evaluation

Introduction and pre-study considerations

I am interested in whether it is possible to predict the frustration level of an individual purely based on their heart rate data. I will therefore not take into consideration the round, phase, or cohort or role of the individual despite these features also being given in the dataset. I will thus purely train models on the HR features with the target variable 'Frustrated' and grouped by the 'Individual' feature.

As the frustration levels are discrete, I briefly considered training classifier models, but this would lose the quality of a guessed class being closer or farther from the true class. A classifier model would for example consider '7' and '2' equally bad guesses if the right frustration level was '8'. This seems wrong to me, as I believe a closer guess should score better. I will thus use regression models with mean square error as the metric to minimize.

Since I do not know whether the self-reported frustration level data has interval consistency and can be treated as interval data or should be treated as purely ordinal, I will compare the performance of a linear and a non-linear model. I assume that a non-linear model might perform better on non-interval-consistent data. Specifically, I will use a multi-linear Ridge regressor and a random forest regressor. I will also compare them against a baseline model that always predicts the mean of the training set to test whether they actually benefit from utilizing the data features.

Exploratory analysis of data

The heart rate features are summary statistics of heart rate tracking that 14 individuals underwent before, during, and after solving puzzles. The data is a subset of the data from the 2024 paper *EmoPairCompete - Physiological Signals Dataset for Emotion and Frustration Assessment under Team and Competitive Behaviors*. As I am interested in predicting frustration levels purely based on the heart rate of individuals, I am using the following features:

Input Features						**Target variable**	**Grouping**
'HR_Mean'	'HR_Median'	'HR_std'	'HR_Min'	'HR_Max'	'HR_AUC'	'Frustrated'	'Individual'

Table 1: features used for model training and evaluation. A full description and implementation can be found in the assignment GitHub.

Grouping the datapoints by the individual they were measured on is important, as the cross folds must be split so that the models are not tested on data from individuals whose data they were also trained on. This will prevent data leakage and test the models' ability to generalize and predict frustration levels from heart rate data on new individuals.

I will first look for relationships in the data. Linear relationships can be gauged with a correlation matrix, and nonlinear ones by plotting each feature against the others.

As the dataset consists of multiple datapoints per individual, the plotted correlation matrix is the average of all in-group feature correlation matrices. This is because I am interested in the correlation between heart rate data and frustration level for a given individual.

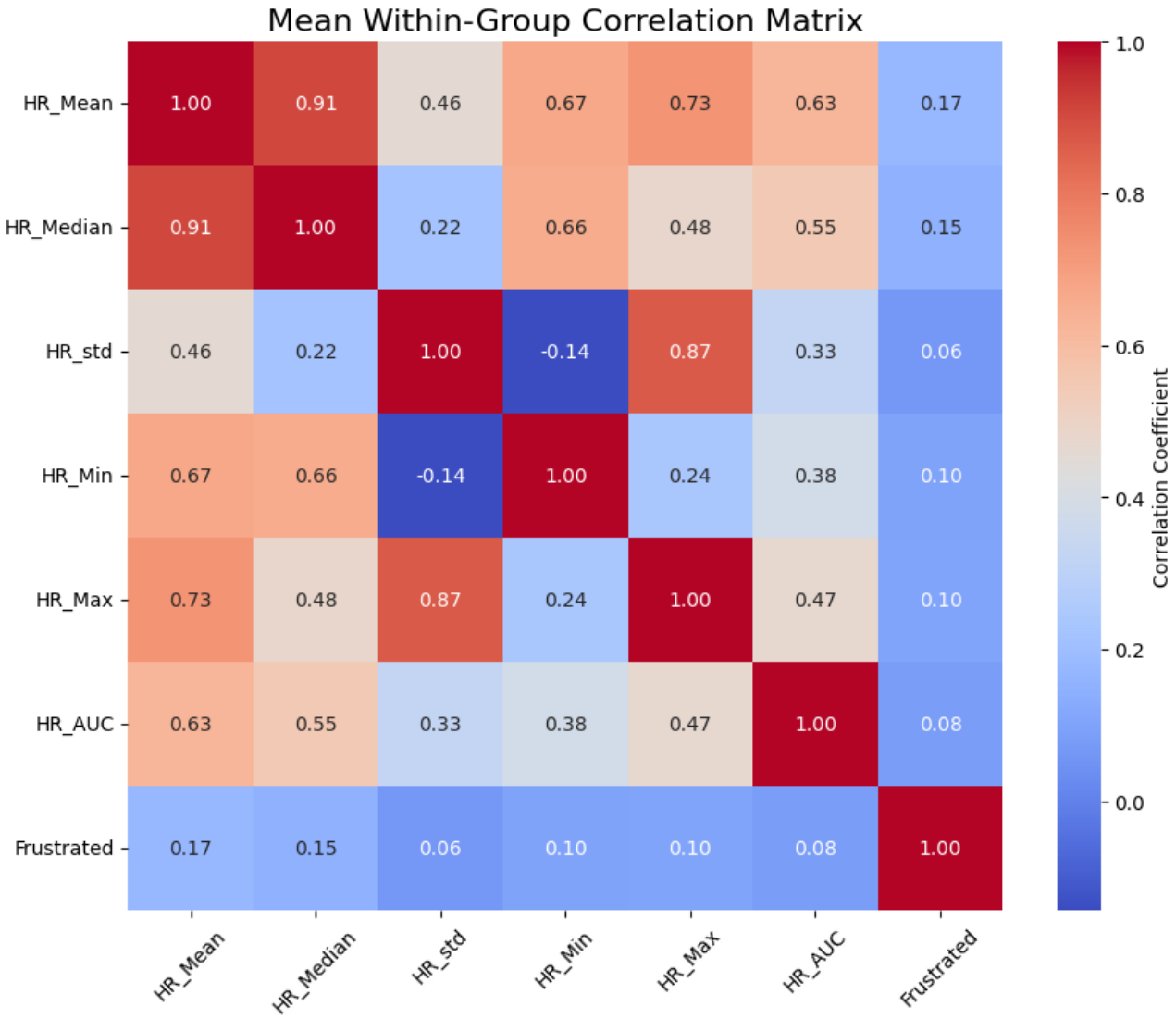

Figure 1: Mean within-group correlation matrix

There are strong correlations between the HR features, which is expected, as they are all summary statistics on the same data: it is for example natural that an individual with a high mean and standard deviation would also have a high maximum measurement. Mean and median heart rate both have some correlation with the frustration level, but the correlation between any feature and frustration is generally weak. However, there could be nonlinear relationships, especially if the frustration levels are non-interval-consistent. This can be gauged with by plotting each feature against the others in a seaborn pair plot

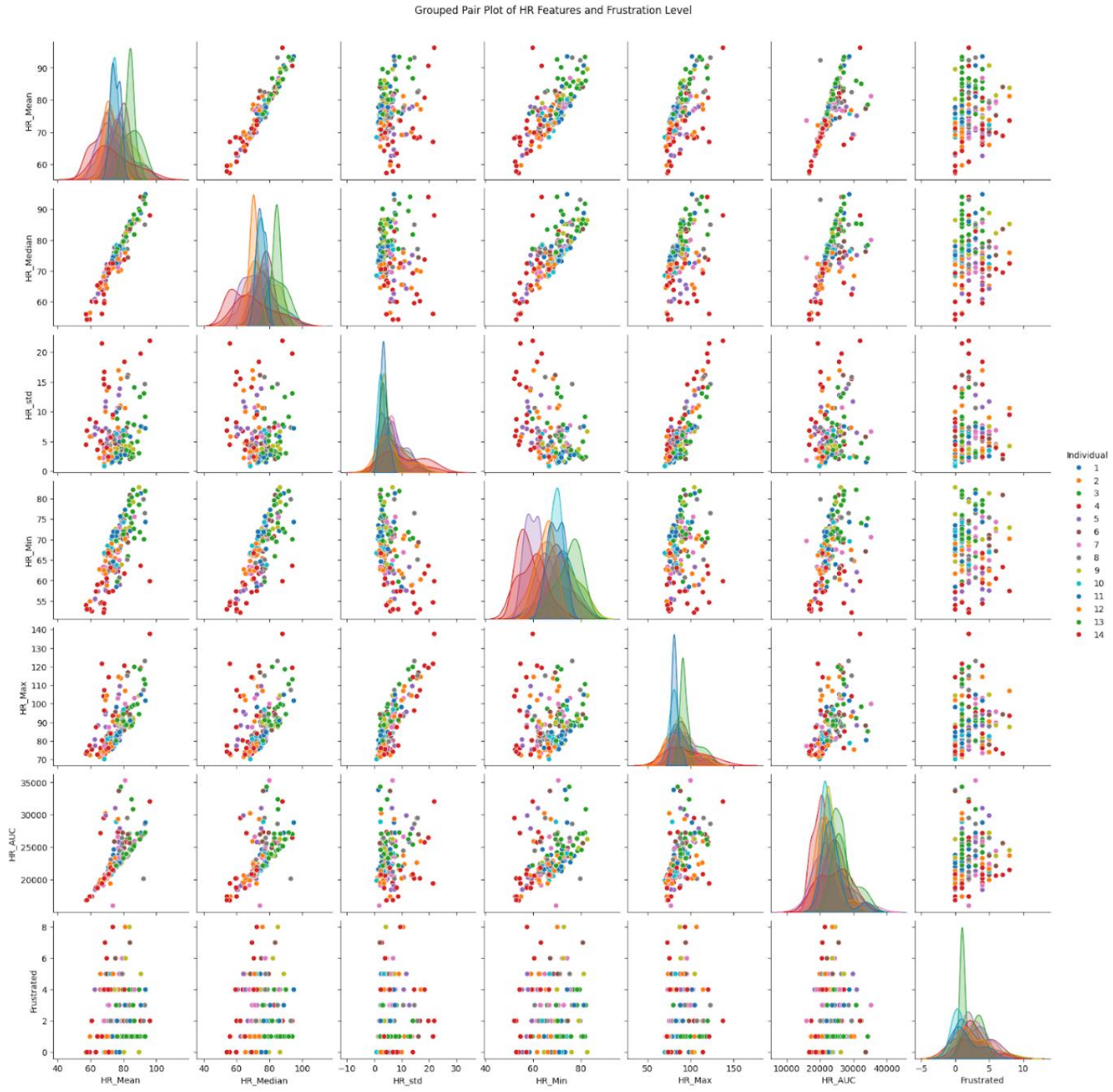

Figure 2: Pair plots of features and target variable. The datapoints are color-coded by the individual they were measured on.

There do not seem to be significant nonlinear relationships between the input features and the frustration level either. With low correlation and relatively few datapoints (14 groups of 12 datapoints each, total n=168), it is uncertain whether the weak relationship will overcome noise and allow either the random forest regressor or the Ridge regressor to beat the baseline.

Methods for model training and evaluation

The six HR input features were standardized to eliminate influence of scale on the models.
I wanted to optimize the hyperparameters of each model, and used grid search on the following hyperparameters for this purpose:

LR Parameter	**Options**	
Regularization (alpha)	0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0	
RF Parameter	**Option 1**	**Option 2**
Number of Estimators	50	100
Maximum Depth	10	20
Minimum Samples Split	2	5
Minimum Samples per Leaf	1	2
Maximum Features	Square root	Log2

Table 2: tuneable hyperparameters for grid search. All other hyperparameters are the default for the model_selection.GridSearchCV method in scikit-learn version 1.1.3 (see the assignment GitHub for full implementation).

To first optimize the hyperparameters of the two models and then compare them without data leakage, I used two-layer cross-validation. This approach meant that the models were not tested for their final generalization error on the validation data used for hyperparameter tuning, which could otherwise lead to an over-optimistic generalization error estimate.

I chose to do 5-fold cross-validation for both inner and outer folds because it seemed like a good middle-ground between the holdout method (where the test/validation sets might not approximate the full dataset well for a smaller dataset like this one) and the Leave-One-Out method, where variance is high, and which is more computationally expensive.
As the datapoints include multiple heart rate tracking from the same individuals, I made sure to group the data by the ‘Individual’ feature so that no data leakage would occur by training and validation/test sets containing the same individuals.

Results and discussion

Metric	Baseline	Ridge Regression	Random Forest
Mean MSE	3.7032 ± 1.8041	3.7826 ± 1.6543	4.0490 ± 1.4149
STD MSE	1.4562 ± 1.6560	1.3323 ± 1.5152	1.1395 ± 1.2959

Table 3: model mean square error and root mean with 95% confidence intervals

Comparison	t-statistic	p-value
MSE (Ridge vs RF)	-0.7423	0.4992
MSE (Ridge vs Baseline)	0.9618	0.3906
MSE (RF vs Baseline)	0.8111	0.4628

Table 4: pair-wise t-tests with null hypothesis that model performances are equal.

The confidence intervals for the means and standard deviations in **table 3** were calculated with the t-distribution and the chi squared-distribution respectively. This was done as the p-values of Shapiro-Wilk tests on the baseline, ridge regression, and random forest model MSEs were calculated to 0.7453, 0.9388, and 0.9492 respectively. The mean square errors can thus be assumed to be normally distributed. However, the independence assumption is less definitive, as the features within groups are strongly correlated. This might present a problem for the credibility of the tests. These concerns also apply to the pair-wise t-test comparisons presented in **table 4**.

Assuming the tests are credible, the models do not perform significantly differently, and neither the random forest nor the Ridge regressor outperforms the baseline model. Interestingly, the optimal hyperparameter for the Ridge regressor was a regularization strength of 100, the maximum value considered by the grid search. As the regularization strength makes the bias term larger, this is an interesting indicator that the correlation between the heart rate features and frustration level so weak that high bias beats high variance for minimizing generalization error.

As two-layer cross validation was executed to prevent data leakage, the mean MSE should meaningfully convey the robustness of the models. The random forest regressor was the most consistent across folds (shown by lower standard deviation of MSE, but similar standard deviations for the baseline and Ridge regression models as well as wide confidence intervals indicate that the consistency of the 3 models are not significantly different.

With these findings, it cannot definitively be said whether frustration levels can be predicted from heart rate measurements. I could not show that it is possible, but with more data it might be possible to train more capable models.

Reference and source of data

As mentioned in the section on exploratory analysis, the data used to train and evaluate models for this assignment is a subset of the data collected previously collected from the paper *EmoPairCompete - physiological signals dataset for emotion and frustration assessment under team and competitive behaviors* by Sneha Das, Nicklas Leander Lund, Carlos Ramos Gonzáalez, and Line H Clemmensen. It was published in ICLR 2024 Workshop on Learning from Time Series For Health in 2024.

A thorough description of the data was provided to me with the assignment description, titled' data_description.pdf' in the assignment GitHub repository.

Link to project GitHub

The full code and data for the project can be found on the project repository:
https://github.com/jonathantybirk/Individual-Assignment-02445-course-DTU
The code is seeded to make reproduction of results easy.