02445 - Individual Assignment

Statistical Evaluation for Artificial Intelligence and Data

Peter Vestereng Larsen, s234839

June 21, 2023

1 Building and Evaluating Predictive Models to Predict Frustration from Heart-Rate Signal

1.1 Choice of Models

For predicting frustration from heart-rate signal, five different machine learning algorithms were chosen:

- Random Forest Classifier (RF)
- Naïve Bayes Classifier (NB)
- Linear Support Vector Classifier (SVC)
- Multi-layer Perceptron (MLP)
- Baseline Classifier (BASE)

A RF-classifier is an ensemble method, building several decision trees that are each constructed from a bootstrap sample of the training data. RF-classifiers are known to have relatively low variance, as each decision tree as diverse trees are combined and avereged. On the other hand, some bias may be introduced. The NB-classifier utilizes Bayes Theorem with the naïve assumption in order to estimate probabilities of class label. SVC-classifiers are known to be effective in high-dimensional spaces and when working with few samples. It works by fitting a hyperplane that splits data effectively into classes. Finally, the MLP-classifier is a relatively standard feed-forward neural network. All models, except BASE are provided by the SciKit-Learn library. Apart from increasing number of iterations allowed for the MLP-classifier, all models were initiated with their default settings. The BASE-classifier is created manually. SciKit-Learn, n.daSciKit-Learn, n.dbSciKit-Learn, n.dc[SciKit-Learn, n.dd]

The baseline classifier will always predict the majority class seen during training and will therefore serve as a benchmark for the remaining models.
The reasoning behind the choice of models is to see how different approaches to the classification task performs. Due to the innate difference of the inner workings, these models all have different advantages and disadvantages. Hopefully, this study will shed a light on which of these model is most appropriate for this kind of task.

1.2 Data Preprocessing

This study assumes that the reader is familiar with the EmoPairCompete dataset[Das et al., 2024]. Upon inspection of the dataset, part of the EmoPairCompete dataset, categorical variables is one-hot encoded in order for the data to be passed to the models.

Ultimately, data consisting of 5 numerical inputs and 25 boolean inputs are passed to the models. These 25 boolean inputs represents 5 different categorical values. In total 30 inputs are fed to the models.

It is noted that the dataset is fairly unbalanced with very few individuals reporting frustration of high levels (see fig 1).

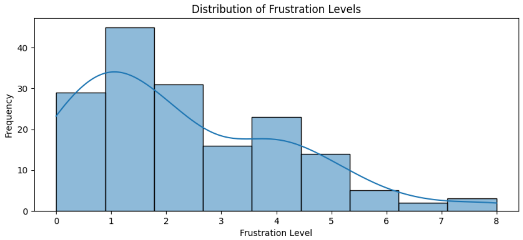

Figure 1: Frustration Levels in EmoPairCompete dataset

1.3 Methods

For evaluating model performance and generalization error, leave-one-group-out (LOGO) cross-validation (CV) is performed. Analyzing the dataset, it is evident that one could group data together in several different ways, however, the index of the individual is chosen as the grouping attribute. The models will train on all but one individual and test on the remaining. This is done, as one can suspect that much data may be leaked from training to test sets if observations of the same individual are found in both sets. It is however acknowledged that other issues may arise due to the inherent class imbalance of the dataset. The numerical attributes of the test data will be standardized within the fold and the testing data will be scaled according to the distribution of the training data. In order to ensure that enough data is available for eventual analysis, this cross-validation schema is repeated 100 times and metrics are stored. Each time, the data within the training and testing groups is separately shuffled using different seeds, so that observations will appear in a random order.

The models will output predicted classes, representing the discrete frustration scale. A metric used for evaluating model performance will be the mean squared error between the model prediction of frustration and actual label. This metric is appropriate for the task, as a wrong guess which is close to the actual label on the scale will be punished less than a wrong guess far from the actual label. The models will also be evaluated based on their weighted F1-score (WF1). The F1-score for a given class is calculated by[Leung, 2022]:

$$F1 = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{TP})}$$

Where TP and FP are true positives and false positives respectively. The weighted F1-score is then calculated as:

$$\text{WF1} = \Sigma_{i=1}^{n} \text{supp}(y_i) \cdot \text{F1}_i$$

Where n is the number of classes, supp(y_i) is the support of that class divided by the total number of instances and F1_i is the F1-score related to that class. The weighted F1-score is an appropriate metric for multi-class classification problem with an imbalanced dataset, much like the one, this study is treating. Metrics are calculated for each CV-fold and stored for subsequent analysis. It should be noted that the stored MSE is the average MSE of predictions accross each fold, while the WF1-score is calculated on each fold. The metrics are both calculated 'fold-wise' in order to facilitate fair comparisons between the two.

It should be noted that frustration was rated on a scale from 0-10. However, the dataset only includes frustration levels 0-8. Therefore, the models will only be able to output values found in the support, and levels 9-10 are ignored in this study.

Regarding the significance level used in this study, we are choosing a significance level of 0.05. Bonferroni correction is applied:

$$\alpha_{\text{bonf}} = \frac{\alpha}{(k(k-1))/2}$$

With k being number of models compared. This gives us a Bonferroni-corrected significance level of $0.005 = 5 \cdot 10^{-3}$. This brings the Family-Wise error rate down to 0.0489.

1.4 Results and Discussion

Subsequent to the training-testing phase, the resulting MSEs and WF1 are analyzed for each model. The distributions for MSEs are shown in figure 2. The distributions of the model MSEs do not all appear to be normally distributed. Histograms for NB, SVC, and BASE appear quite patchy while distributions for MLP and RF appear to be heavily skewed. In order to determine which statistical tests are appropriate for this scenario, the normality of the distribution of each model's MSE and WF1 is investigated. QQ-plots for the MSE of the models are shown in figure 3. The plots seem to reveal somewhat systematic deviations from the theoretical quantiles, suggesting that normality may not be assumed. This conclusion is supported by the Shapiro-Wilk test, where all p-values fall below the significance level of .005 on both metrics, see table 1. It is shown that log-transforming the metrics of the models does not result in normality (see table 6 in appendix). It is concluded that normality cannot be assumed.

In order to assess whether or not there is significant difference between the distributions for model performance, a non-parametric test must be utilized. Given the confines of this course, we assume independence between model performance. No statistical tools for testing difference between paired groups where normality can not be assumed have been introduced. Since the models operate differently, we will assume that this is enough to assume independence. It

Model	**MSE Shapiro-Wilk Test p-value**	**WF1 Shapiro-Wilk Test p-value**
Baseline	1.555e-22	4.542e-48
MLP	1.950e-17	1.581e-21
Naïve Bayes	1.623e-21	6.969e-34
Random Forest	7.056e-24	3.801e-22
SVC	1.817e-34	4.681e-34

Table 1: Shapiro-Wilk Test p-values

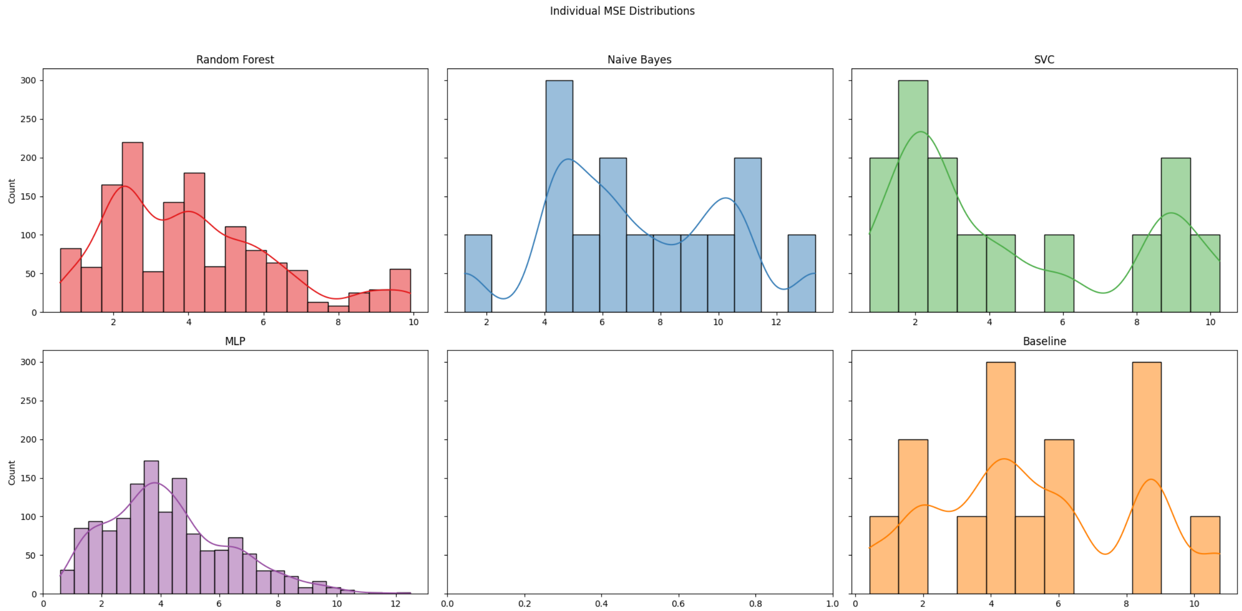

Figure 2: Distributions of MSE for RF, NB, SVC, MLP, and BASE

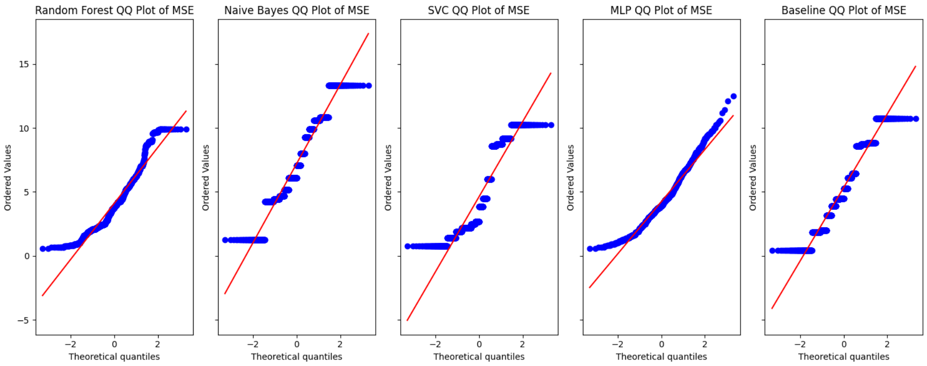

Figure 3: QQ-plots of MSEs

should, however, be noted that since the models are trained on the same data, the performance of them will most likely be more or less dependent on this data. However, given that independence is assumed, the Kruskal-Wallis H test is chosen. The results of the Kruskal-Wallis test implies significant difference between distributions of groups on both metrics with both p-values appearing below the significance level. The results are shown in table 2.

When investigating the difference of distribution between pairs of models, the Wilcoxon Rank-sum test is chosen for the post-hoc analysis. This test is non-parametric but assumes independence, and is therefore appropriate given the data. The Wilcoxon test is performed for each pair of models and the p-values are collected in a matrix. This matrix is displayed in figure 4. It is suggested by the analysis, that in regards to both metrics, most pairs of models do statistically significantly differ from one another in regards to distribution of metric performance. There are however

Metric	**H Test Statistic**	**H Test p-value**
MSE	925.979	3.916e-199
WF1	316.437	3.080e-67

Table 2: Kruskal-Wallis H Test Results

some exceptions. These exceptions are shown in table 3.

MSE Comparison	**P-value**	**WF1 Comparison**	**P-value**
RF - SVC	0.53	NB - BASE	0.82
SVC - MLP	0.17	RF - MLP	0.18
RF - MLP	0.0072	NB - SVC	0.10

Table 3: P-values for MSE and WF1 Distribution Comparisons

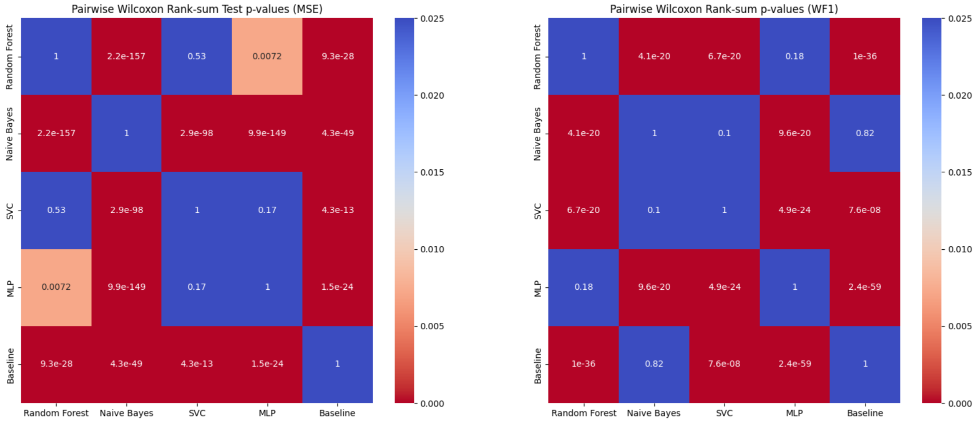

Figure 4: Matrices of Wilcoxon Rank-sum test p-values

It is thus not possible to reject the null hypothesis for these classifiers on the given metrics. Interestingly, the null hypothesis for WF1 score of the Naïve Bayes and Baseline Classifier cannot be rejected, meaning that their WF1-score distributions may be similar. When correcting by the Benjamini-Hochberg procedure with a q-value of 0.1, the null hypothesis for the MSE of RF - MLP is rejected, as is the null hypothesis for WF1-score of NB-SVC.

The rejection of many null hypotheses across both metrics suggests that some models are performing statistically significantly different from one another. This implies that some models may generally perform better or worse than their adversaries.

To further evaluate the performance of the models, bootstrapping is performed in order to find 95% confidence intervals of the performances. Utilizing bootstrapping, distributions for the means of the metrics are estimated. This is appropriate, since the Central Limit Theorem states that given a large enough sample size, the means of resamples will follow a normal distribution. It is deemed that when displaying confidence intervals, the Bonferroni corrected significance level presents itself as too conservative and the resulting confidence intervals do not give much information. This fact will and should however be kept in mind. No hypothesis tests will be performed based on these confidence intervals, and the process of Bonferroni correction followed by Benjamini-Hochberg procedure is therefore deemed unnecessary.

When observing the confidence intervals in figure 5, it would seem that mean model performance differ relatively more for the MSE-metric than for the WF1-score. There appears to be some evidence suggesting that the Random Forest and MLP Classifiers outperform the remaining classifiers on both metrics. When observing the confidence intervals for the WF1-score, the SVC, BASE, and NB are grouped very close to each other. It is again noted that confidence intervals are not Bonferroni corrected. Exact values for confidence intervals, means and standard deviations are found in table 4. Standard deviations are based on original data.

Due to the fact that the distributions for model metrics are not normal, mean performance may not be the appropriate measurement of general model performance. Therefore, the medians of the metrics are analyzed as well. These can be observed in table 5. Boxplots are shown in figure 6. These are created based on original data and not the bootstrapped data.

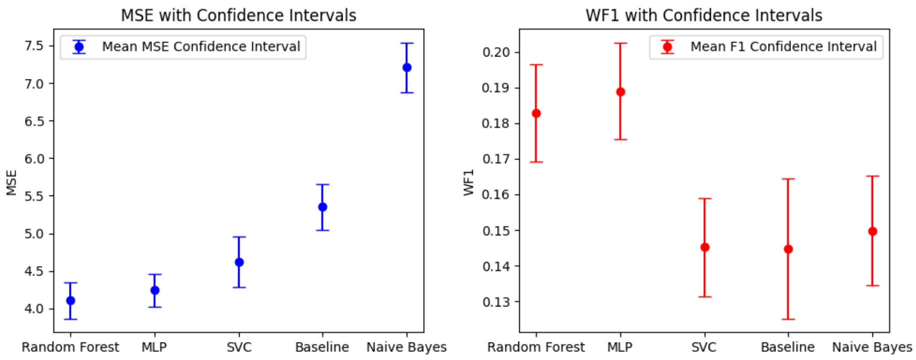

Figure 5: Bootstrap Confidence Intervals for MSE and WF1

Model	**MSE Lwr**	**MSE Upr**	**MSE Mean**	**MSE Std**	**WF1 Lwr**	**WF1 Upr**	**WF1 Mean**	**WF1 Std**
RF	3.9864	4.2250	4.1004	2.2524	0.1761	0.1898	0.1826	0.1301
MLP	4.1336	4.3504	4.2840	2.0678	0.1821	0.1958	0.1838	0.1295
SVC	4.4530	4.7890	4.7590	3.1687	0.1384	0.1524	0.1444	0.1329
BASE	5.2034	5.5058	5.2240	2.9440	0.1352	0.1548	0.1520	0.1883
NB	7.0436	7.3742	7.1517	3.1555	0.1423	0.1576	0.1493	0.1470

Table 4: Confidence Intervals, Means, and Standard Deviations of metrics

The medians for model performance appear to support the confidence intervals. It can be observed that the median MSE of the SVC is relatively low compared to other models, while it's WF1-score does not get very high. This suggests that the SVC excels at giving close estimates of frustration levels while rarely being totally accurate. This observation is further reinforced by the boxplot exhibiting a wider range in MSE performance and relatively low general WF1-scores.

Model	WF1 Median	MSE Median
Random Forest	0.1655	3.833
Naive Bayes	0.1042	6.583
SVC	0.1138	3.250
MLP	0.1574	3.917
Baseline	0.1000	4.875

Table 5: Medians of metrics

The MLP seems to perform reasonably well on both metrics, but seems prone to outliers of both good and bad performance.

The RF-Classifier generally scores relatively high on the WF1-score while performing reasonably well on the MSE metric.

It appears that the NB-classifier does not conclusively outperform the BASE-classifier, as there is no evidence to suggest, that the distribution of its WF1-metric is different from that of the BASE-classifier. When Observing the MSE medians of both classifiers, the BASE-classifier appears to outperform the NB, supported by a statistically significant Wilcoxon test.

Given that many pairs of models tested significantly different from each other in the Wilcoxon test, the medians for metric performance of the classifiers may suggest, that models such as RF, MLP and SVC generally perform better than the BASE and NB-classifiers. However, the RF - SVC and SVC - MLP tests showed no significant difference in MSE distribution. On the other hand, the RF-classifier seems to excel on the WF1 metric. It is not significantly different from the MLP, but there is some evidence to suggest that these two classifiers outperform the others on this metric.

Ultimately, there is no evidence to conclusively decide which model is the best. If one model were to be chosen, many factors would have to be considered in order to determine, which model performs optimally for the specific problem. Perhaps one metric is more favorable than another - for one task, false positives may be a serious issue, while in another situation, a false negative is a more grave error. A factor could also be the consistency of the model - perhaps one would prefer a stable model with a slightly worse mean MSE, or perhaps one would prefer a more inconsistent model, but one that achieves better median or mean metric performance. These are some of the factors that should be considered when choosing a model for a specific task

General trends

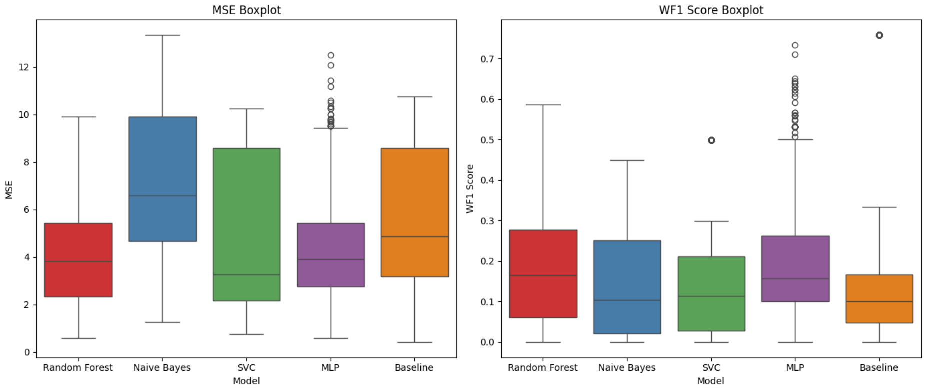

Figure 6: Boxplot of Model Metrics

1.5 Model Generalization, Robustness, and Consistency

When performing cross-validation, the metrics of the model represents the ability of the models to classify new, unseen data. These metrics are therefore estimates of model generalization. Analyzing the resulting metrics, is essentially analyzing how well the models perform on new data. As alluded to earlier, this is not necessarily an easy or simple task.

Regarding model robustness and consistency, these can viewed in different ways. In some sense, models are robust if they handle unseen data well, similar to how we analyze model generalization. Robustness and consistency could also be viewed as a term for variance in results. When observing the boxplots of model metrics in figure 6, it seems that the MSE of the RF and MLP classifiers do not vary as much as for the remaining classifiers accross multiple CV-folds and rounds as their respective 2nd and 3rd quantiles are relatively close to each other compared to those of the NB, SVC, and BASE. This is further supported by the standard deviation of their MSEs (table 4), as they are relatively low compared to other models. It is noted, though, that the Wilcoxon Rank-sum test did not find statistically significant difference in the distributions of RF - SVC MSE or MLP - SVC MSE, and one should therefore not draw any permanent conclusions. Even though the median of the SVC-MSEs is lower than that of the MLP, one might prefer the MLP due to it's general consistency on MSE-performance. One could think of the MLP and RF classifiers as more robust than their counterparts, as they appear to generally give more similar and consistent results accross multiple different testing rounds.

When observing the boxplot of the SVC WF1-score, it seems to generally obtain a WF1-score in a narrower range than the counterparts - even though the analysis implies that it may generally perform worse than some of the other classifiers, it is easier to predict, in what range an observation will be found, which may be preferred in some cases. The RF and MLP model do appear to give a wider range of results, making them less consistent, but in return sometimes scoring relatively well on the WF1-metric.

Taking all data into account, it would indicate that a RF, SVC, or MLP-Classifier would be preferred models for this specific issue. This is due to general consistency in results, and overall relatively good performance on both metrics. This is the case both when observing confidence intervals for mean metric performance, median metric performance and when observing relatively low standard deviations of the results. There is evidence to suggest that these models outperform the BASE and NB-classifiers on both metrics.

References

SciKit-Learn. Ensembles: Gradient boosting, random forests, bagging, voting, stacking. `https://scikit-learn.org/stable/modules/ensemble.html#random-forests-and-other-randomized-tree-ensembles`, n.da.

SciKit-Learn. Naive Bayes. `https://scikit-learn.org/stable/modules/naive_bayes.html`, n.db.

SciKit-Learn. Support Vector Machines. `https://scikit-learn.org/stable/modules/svm.html#svm`, n.dc.

SciKit-Learn. Neural network models (supervised). `https://scikit-learn.org/stable/modules/neural_networks_supervised.html`, n.dd.

Sneha Das, Nicklas Leander Lund, Carlos Ramos González, and Line H Clemmens. Emopair-compete - physiological signals dataset for emotion and frustration assessment under team and competitive behaviors. ICLR, 2024.

Kenneth Leung. Micro, Macro Weighted Averages of F1 Score, Clearly Explained. Towards Data Science, 2022.

2 Appendix

2.1 Code

The code utilized in this project can be found at this GitHub repository:
`https://github.com/PeterVL02/02445-Individual`

2.2 Log-transformed Metrics

Model	Log MSE Shapiro-Wilk Test p-value	Log WF1 Shapiro-Wilk Test p-value
Baseline	6.459e-36	3.092e-45
MLP	3.755e-16	1.174e-16
Naive Bayes	8.042e-35	7.617e-33
Random Forest	1.036e-16	1.753e-20
SVC	8.649e-27	4.112e-31

Table 6: Shapiro-Wilk Test Results on log-transformed MSEs