Predicting empathy score for employee recruitment using eye-tracker technology

April 23, 2023

Registration number:	2201578
Link to GitHub:	`https://github.com/mv22003/PredictingEmpathy`

Executive summary (max. 200 words)	189
Main findings and Discussion (max. 600 words)	546
Conclusions (max. 300 words)	333
Total word count	1068

Contents

Abstract

Eye-tracking technology offers information such as pupil diameter, gaze point, and gaze vectors; this data can contribute to emotional markers like empathy scores. In this paper, we developed a classifier that uses eye-tracking data to predict an empathy score. This estimator intends to provide an empathy assessment for other companies, which can be use during employee recruitment.

We created several classifiers based on different methodologies to recognize the best for predicting empathy. Our target/label was discretize in four classes to upgrade our accuracy and to provide a more understandable predictor for our clients. A classification model based on Stochastic Gradient Descent reached an accuracy of 0.36 +/- 0.22 for the test group, and a model based on Logistic Regression threw an accuracy of 0.39 +/- 0.10 for the control group. Subsequently, by using Recursive Feature Elimination (RFE), we created two optimized classifiers that threw accuracies of 0.39 +/- 0.11 and 0.53 +/- 0.03, respectively.

Finally, our results show a promising start to developing empathy score classifiers. Furthermore, we analyzed the performance of our classifiers and provided recommendations to enhance the data acquisition process, the pre-processing of the features and the methodologies developed.

1 Main Findings

1.1 Feature extraction

For the feature extraction, all the recordings from each participant are considered. Figure 1 contains the dimensions of the final feature arrays for both groups. The values of the rows are represented by the **recordings per group** and the columns are the **features created per recording**.

It is important to notice that the number of recordings per participant is different, resulting in an **unbalanced feature matrix**. These dimensions were created by the extraction of the features explained in table 1. In this figure, the features were divided depending on their origin and describe each feature briefly.

For the prediction of the empathy score, the **label was categorized into four ranges** to create a more understandable classifier. In figure 2, the ranges used to transform our target from a continuous to a discrete variable are presented.

It is important to notice that, with the proposed ranges, **the control group only has scores from the first three classes (0,1,2) in comparison to the test group (0,1,2,3)**, and this creates differences in the dynamic of both estimators. For a better understanding, see figure 2.

1.2 Comparison between predictions from test and control groups with several classifiers (intermediate results)

Several classification models were selected to train and predict our target. The results obtained are displayed in table 2. We can find the best performer from each group highlighted. In this comparison, we can appreciate that the test group is harder to classify based on an estimate average from all the accuracies shown in table 2. For the control group, the classifiers perform better but their is no drastic difference.

Figure 1: Dimensions of the features extracted per group

Groups	
Test	Control
(361,19)	(147,19)

Table 1: Features extracted per recording explained

Identifier features	
1. Participant name	Identifier for groups per participant, not used in training
Pupil diameter features	
2. Mean Pupil diameter left	Average of the left pupil diameter in the time-series
3. Std Pupil diameter left	Standard deviation of the left pupil diameter in the time-series
4. Mean Pupil diameter right	Average of the right pupil diameter in the time-series
5. Std Pupil diameter right	Standard deviation of the right pupil diameter in the time-series
Eye movement features	
6. Num. of Fixations	Number of Fixations per recording
7. Num. of Saccades	Number of Saccades per recording
8. Num. of Unclassified	Number of Unclassified eye movements per recording
Duration features	
9. Recording duration	Duration of the recording in seconds
10. Mean Gaze event duration	Average gaze event duration in seconds
Gaze point features	
11. Mean Gaze point X	Average of the gaze point X in the time-series
12. Std Gaze point X	Standard deviation of the gaze point X in the time-series
13. Mean Gaze point Y	Average of the gaze point Y in the time-series
14. Std Gaze point Y	Standard deviation of the gaze point Y in the time-series
Fixation point features	
15. Mean Fixation point X	Average of the fixation point X in the time-series
16. Std Fixation point X	Standard deviation of the fixation point X in the time-series
17. Mean Fixation point Y	Average of the fixation point Y in the time-series
18. Std Fixation point Y	Standard deviation of the fixation point Y in the time-series
Prediction Label	
19. Empathy Score	Label/Target to predict

1.3 Comparison of the best classifier for the test and control group (final results)

Following section 1.2, we selected the best classifier from each approach and performed **Recursive Feature Elimination (RFE)** to increase our performance. The results are in figure 4. The accuracies thrown by our classifiers are better than our intermediate results, with the control group being the best classified.

1.4 Features selected from best classifiers for each group using RFE

As seen in figure 4, we obtained better results in the RFE estimators compared to the intermediate results. With this in mind, we extract the features selected in both classifiers, as shown in figure 5. Here we mentioned the features that estimate our label and gave a score to each participant.

2 Discussion

2.1 Discovery from intermediate results

The results for the intermediate predictions in the **testing group** (table 2) show the following:

- The Stochastic Gradient Descent Classifier has the best performance.
- Every classifier has a high standard deviation.
- Every classifier performs better than the dummy classifier.

The intermediate estimations from the **control group** (table 2) demonstrate:

Figure 2: Discretization of empathy label

Empathy Score Range	**Name of the Class**	**Numerical Value**
<100	Bad	0
100-110	Average	1
110-120	Good	2
>120	Outstanding	3

Table 2: Accuracy of classification estimators

Model	**Accuracy**	
	Test	Control
Dummy Classifier	0.12 +/- 0.15	0.40 +/- 0.08
Stochastic Gradient Descent	0.36 +/- 0.22	0.37 +/- 0.05
Nearest Centroid	0.34 +/- 0.12	0.28 +/- 0.07
Logistic Regression	0.29 +/- 0.18	0.39 +/- 0.10
Decision Tree	0.24 +/- 0.15	0.30 +/- 0.09
Random Forrest	0.28 +/- 0.14	0.33 +/- 0.09
KNearest-Neighbors	0.30 +/- 0.21	0.38 +/- 0.11
Support Vector Machine	0.23 +/- 0.15	0.32 +/- 0.06
AdaBoost Classifier	0.28 +/- 0.10	0.29 +/- 0.09
Neural Network	0.34 +/- 0.21	0.35 +/- 0.09

- The Dummy Classifier has the best performance, but this can be affected by the categorization of the label; in the test split, the label repeats itself 40% of the time.
- The Logistic Regression Classifier is the second best classifier and used for the final predictions.
- Lower values of standard deviation compared to test predictions.

Overall, there could be a better classifier for both groups. The nature of the data created from the recordings and the ambiguity of our label makes it harder for the classifiers to perform well.

2.2 Correlation between labels and features from best classifiers in intermediate results

After gathering the intermediate results, we continue observing the behavior of the classifiers, specifically the best performers from both groups. Figure 3, demonstrates that the *pupil diameter, the gaze point, and the fixation point* are the best features to estimate the empathy score in the **test prediction**. While for the **control group**, the best features to estimate the score are the *gaze point, the fixation point and the number of saccades per recording.*

3 Conclusions and Recommendations

For the conclusions and recommendations of the methodology, we present the crucial discoveries and their meaning to the models created.

Intermediate results of test group predictions (figure 3):

1. The mean value of the **pupil diameter** is relevant in predicting the worse empathy scores (<100).
2. The standard deviation of the **gaze point** and the **fixation point** are important differentiators for the middle empathy classes (100-120).
3. The standard deviation of the **pupil diameter** has a high impact in predicting the outstanding empathy scores (>120) .

Figure 3: Weights and coefficients of features in the predictors

Group	**Model**	**Features**	**Weights/Coefficients**	**Class**
Test	Stochastic Gradient Descent	Mean Pupil diameter left	6.40	0
		Std Gaze point X	17.41	1
		Std Fixation point Y	5.18	2
		Std Pupil diameter left	21.66	3
Control	Logistic Regression	Std Gaze point X	0.81	0
		Mean Fixation point X	0.71	1
		Num. of Saccades	0.94	2

Figure 4: Accuracy of estimators with Recurrent Feature Elimination (RFE)

Group	**Model**	Without RFE	With RFE
Test	Stochastic Gradient Descent	0.36 +/- 0.22	0.39 +/- 0.11
Control	Logistic Regression	0.39 +/- 0.10	0.53 +/- 0.03

Intermediate results of control group predictions (figure 3):

1. The standard deviation of the **gaze point** gives us relevant information about bad empathy scores (<100).
2. The mean **fixation point** has a high coefficient value and helps estimate average empathy scores (100-110).
3. The **number of saccades** has a massive coefficient and appears to be an important feature to distinguish good empathy scores (110-120).

Finals results of **test group** predictions:

1. Features selected by methodology backup the literature, **pupil diameter is a good predictor for empathy**.
2. Difference between intermediate and final results (figure 4) is not that impressive.
3. The **Gaze point** features are a common in all predictors.
4. All score labels are present in test data (figure 2).

Finals results of **control group** predictions:

1. Best classifier for the control group only uses the **average duration of the gaze event** per recording as a feature.
2. Relevant difference[1] between intermediate and final results (figure 4).
3. The nature of the experiment gives a bigger important to **gaze events** in comparison to other features.
4. Only first **three classes of score** (figure 2) present in control data.

Finally, as a data scientist, I recommend the following:

- Reduce the number of trails by increasing the duration of the recording.
- Propose one more methodology to acquire an empathy score, some sort of ground truth performed by the HR department so that the scores suits the profiles the company is looking for.

Figure 5: Features selected in RFE estimators

Group	**Model**	Features Selected
Test	Stochastic Gradient Descent	Mean Pupil diameter left Mean Pupil diameter right Std Pupil diameter left Std Pupil diameter right Std Gaze point X
Control	Logistic Regression	Mean Gaze event duration (s)

- Increase the number of participant on both the test and control experiments, specifically in the control group because it did not have all the score labels proposed.

References

[1] N. A. Harrison, C. E. Wilson, and H. D. Critchley. Processing of observed pupil size modulates perception of sadness and predicts empathy. *Emotion*, 7(4):724–729, 2007.

[2] P. Lencastre. Code to read data. 2022.

[3] P. Lencastre. Eye tracker data. 2022.

[4] P. Lencastre. Questionnaires. 2022.

[5] P. Lencastre. Raw_Data. 2022.

[6] P. Lencastre, S. Bhurtel, A. Yazidi, G. B. e Mello, S. Denysov, and P. G. Lind. Eyet4empathy: Dataset of foraging for visual information, gaze typing and empathy assessment. *Scientific Data*, 9(1), 2022.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[1] The results were done by using *StratifiedKFold*, instead of *GroupKFold* like the intermediate approach, because of a bug in the *GridSearchCV library from sklearn.* For more explanation, see `https://github.com/mv22003/PredictingEmpathy/blob/main/rfe_models_control.ipynb`