Machine learning-based classification of Post-Traumatic Stress Disorder (PTSD) states in military personnel for personalized treatment approach.

Pablo Tejero Sanz

I6329819

INTRODUCTION

Post-traumatic stress disorder (PTSD) is a psychological condition that is usually observed in soldiers after extreme experiences (such as combat)[1]. These people may experience pessimistic symptoms after such experiences, such as nightmares, migranes, loss of concentration and memory, among others[2]. These conditions significantly affect the quality of life of soldiers, making it difficult for them to reintegrate into civilian life.

Knowing this, and in order to improve treatments, this project is carried out, which uses two databases to train a model to classify future cases in the different stages of the disease, in order to improve the treatment of patients.

The databases we have refer to the condition of Post-traumatic Stress Disorder (PTSD) in military personnel. In the file Phenotype_essentials.tsv there are columns with different classifications of this condition and different types of data collection. In this case, it has been decided to focus on the study of the long-term evaluation of PTSD symptoms, located in the traject column. These data are classified according to the time it took for patients to recover from their symptoms, being classified as 1 those who recovered after one year, 2 those who showed no symptoms after one year, or even more, and 3 those who maintain PTSD symptoms[3].

Regarding the data within the PGS_results.tsv file, they were generated using an algorithm called LDAK[4]. A model developed by the same, was also used to make the PGS score[5], this being what is meant for the name "BayesR". It should be noted that for the correct performance of the analysis, the data are normalized and standardized.

Once knowing what data there is to do the analysis, it was decided to opt for a classification approach thanks to Machine Learning. The main reason, is that, it would be very useful for the treatment of patients to find more efficient ways to be able to classify their condition, and thus be able to develop more personalized treatments, in addition, that it would be possible to start with such treatment in an earlier way, since the relevant tests, are possible to carry out in a short time, apart from not being too intrusive for the patient (BMI, cholesterol, or height among others).

In summary, this project aims to analyze a new approach that will allow us to more easily detect the type of state in which the patient is, and thus be able to provide a more personalized treatment, which will help him with his recovery as much as possible.

METHODS

Pre-processing.

In order to carry out an appropriate classification, it is necessary to pre-process the available data in order to obtain more accurate results.

As a complement to the project to be carried out, a small description of the different features was made, where you can see their distribution, among other characteristics. This was done using the ProfileReport function[6] form pandas. As it is very time consuming, the result has been added to the delivery of this test.

First I loaded both datasets in two variables, comparing their samples by the SampleID column, and keeping the samples that are in both of them. After eliminating those samples that are not represented in Phenotype_essential, I searched for repeated values, finding two samples with the same SampleID, so I decided to eliminate one of them and keep the other one. Also the next step is to eliminate the columns of both SampleID and UniqueID, since, once used to compare the data, they do not provide extra information.

Also, for the target variable (y), I obtained only the data of the Traject column, but it can be observed that there are data that appear as NaN. To avoid deleting these samples and losing that information, a commonly used technique is followed, which is to replace the values by the median of the other values in that column. This is the end for the pre-processing of the target variable.

With respect to the feature matrix (X), the data pre-processing is not yet finished, since there may still be outliers that affect subsequent studies. To eliminate the outliers, a function is used to determine the outliers for each column, replacing this value by the median value of that specific column[7]. This way of proceeding was chosen instead of eliminating the samples with outliers, in order to avoid substantial loss of data.

Finally, with the data already pre-processed, the data is divided into X_train, X_test, y_train and y_test, in order to carry out the subsequent classification.

Classification.

The objective is to be able to classify between the different types of trajects the future data that we obtain, for this, different methods are carried out, in order to evaluate their predictive power.

One factor to take into account is that the number of trajects equal 3 is much higher than the other two types, so it can affect the correct classification, even falsify the accuracy, so it has been decided to use two main measures, the normal **accuracy**, and the **balanced accuracy**, which allows to solve this problem of devaluation in terms of the different classes.

Also, for every model used, was checked the confussion matrix, allowing to plot how the test samples that are executed in the model are classified, observing their accuracy, and also in a certain way their distribution with respect to the real preconditions and those that are not[8]. Another measurements were calculated using classification_report function of sklearn.metrics[9] obtaining results of Precision score for each class, which indicates the proportion of correct positive predictions[9,10], the Recall score, which is the proportion of correctly predicted positive cases[9], and the F1-score, which is a combination of both[9]. Finally, the support column indicates the number of samples tested for each class.

First I started implementing a decision tree, checking both, accuracy and balanced accuracy, as well as other measurements and the confusion matrix.

Trying to improve the data, and also understanding, the hyperparameters were found, doing a decision tree using them, and calculating the accuracy and balanced accuracy, as well as the confusion matrix and the measurements explained before.

Also a Random forest was implemented, assuming that the combination of multiple decision trees would improve the results. As it was done previously, the same parameters were calculated.

According to other researchers, the support vector machine (SVM) is also a good model for data classification, which is based on finding an optimal hyperplane, within a multidimensional space, that allows to divide the available data into different groups[11].

In order to better understand the features we are analyzing, a study of their importance was carried out, classifying them from most to least important, and then plotting them to understand in a simpler way that importance.

Once the study of the traject variable was completed, it was decided to implement the same technique for the height_A column, in order to analyze the existing differences, and to see which characteristics are better to classify.

For this, as discrete variables are needed, the lower threshold was set at the 15 percentile, with samples below this threshold being stored as 1, the upper threshold at the 85 percentile, with samples above being classified as 3, and those between the two thresholds as 2. After that, the same methods were followed, eliminating the NaN values by the median, eliminating the outliers, and in this case, only the decision tree models with hyperparameters and a random forest were used, calculating for both also the accuracy, balanced accuracy, confusion matrix and the other measures mentioned above.

RESULTS

As mentioned above, a multitude of models were implemented, analyzing a great variety of results. As so many data were available, it was decided to add a table with both, accuracy and balanced accuracy obtained from all the models, but just the confusion matrix and the other measurements of the model with the best results are explained, in order to better explain it, and in this way, in the script you can see the rest of the most suboptimal models.

Accuarcy and Balanced accuracy

	Decision tree	Hyperparameters	Random forest	SVM
Accuracy	0.7619	0.7738	0.8809	0.6548
Balanced accuracy	0.3117	0.2928	0.3333	0.3165

Table 1. table with the accuracy and balanced accuracy obtained for every model for traject.

Taking into account both the accuracy and the balanced accuracy observed in table 1, random forest model turns out to be the one that obtains the best results for the classification of the data with respect to the types of trajects. The obtained accuracy is 0.8809, being a good accuracy, however, as previously mentioned, the presence of excessive data classified in group 3 in the database could have affected its performance, that is why it was used to analyze also the balanced accuracy, in this case the result is 0.33333, which means that it does not classify correctly.

Applying the study to the other measures, Table 1 shows the Precision score for each class, which indicates the proportion of correct positive predictions, the Recall score, which is the proportion of correctly predicted positive cases, and the F1-score, which is a combination of both. Finally, the support column indicates the number of samples tested for each class.

Precision, Recall, F1-score and Support

	Precision	Recall	F1-score	Support
1.0	0.09	0.08	0.08	13
2.0	0.00	0.00	0.00	7
3.0	0.88	0.87	0.87	148

Table 2. Scores obtained for Precision, Recall and F1-score for every class for traject.

As shown in Table 2, class 3 has the best classification ratio, while the other two have very poor ratios. It is worth noting that if we look at the support column, we notice that most of the data belong to class 3.

Finally, a confusion matrix was plotted to allow a more graphic understanding of these results (Figure 1).

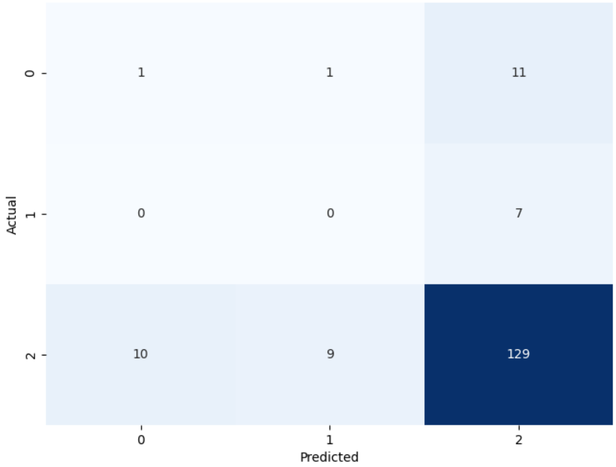

Figure 1. Confusion matrix obtained in Random Forest

In this case, we see that the model classifies practically all the data within class 3 (in this plot expressed as 2), confirming previous suspicions that the large number of samples classified in traject as 3 affects the prediction of the model, since it classifies very well for those that are class 3, but poorly for those that are of another type.

Finally, for a better understanding, we analyzed the features that are most important for classification, plotting the results in Figure 2.

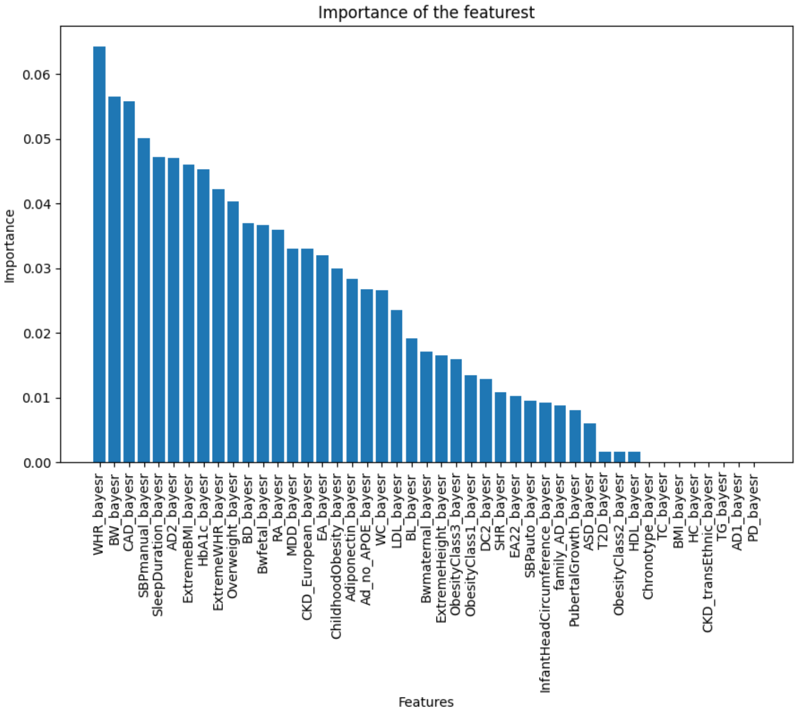

Figure 2. Importance features

The last ones show that aren't important for classification at all, meanwhile, sleep duration, extreme BMI among others show that are important.

Height results.

As explained above, a similar study was carried out, this time with the data from the height column, in order to check whether or not they classify correctly in this way. For this purpose, it was decided to use only the two models that were most optimal in the previous study (decision tree using hyperparameters and random forest).

In this case, as before, the random forest model is the one that classifies best, and, as before, the results are also suboptimal. The accuracy obtained is 0.6904, which is lower than that obtained when calculating for traject. For the balanced accuracy, 0.3333 is obtained, as with traject.

As before, we have also studied the different measures, obtaining this table (Table 3).

Precision, Recall, F1-score and Support

	Precision	Recall	F1-score	Support
1.0	0.00	0.00	0.00	20
2.0	0.69	0.97	0.81	116
3.0	0.00	0.00	0.00	32

Table 3. Scores obtained for Precision, Recall and F1-score for every class for height.

As shown in Table 3, class 2 has the best classification ratio, while the other two have very poor ratios. It is worth noting that if we look at the support column, we notice that most of the data belong to class 2.

Finally, a confusion matrix was plotted to allow a more graphic understanding of these results (Figure 1).

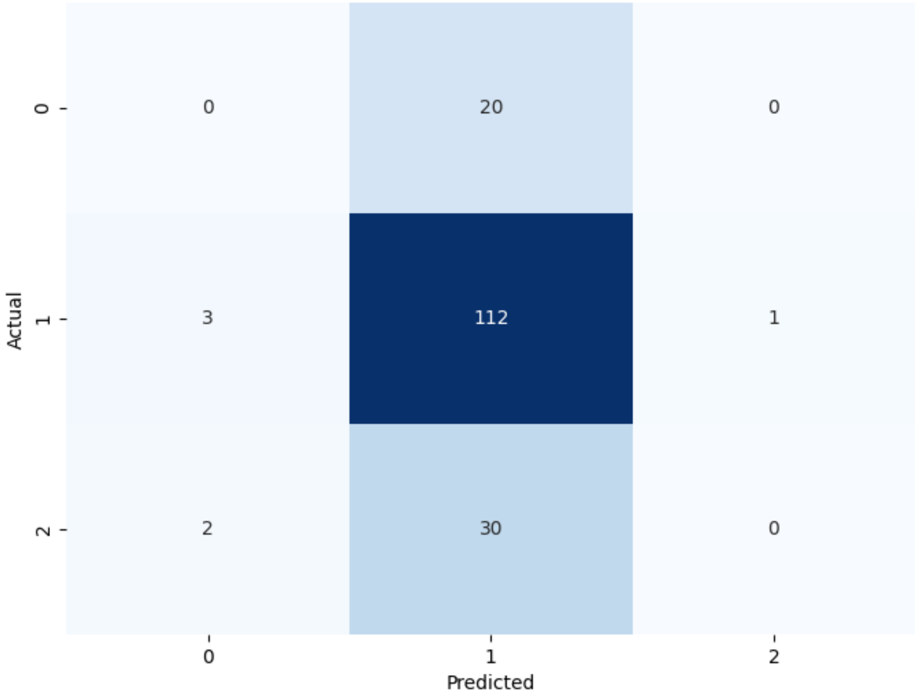

Figure 3. Confusion matrix obtained in Random Forest for height

In this case, we see that the model classifies practically all the data within class 2 (in this plot expressed as 1), confirming previous suspicions that the large number of samples classified in height as 2 affects the prediction of the model, since it classifies very well for those that are class 3, but poorly for those that are of another type.

DISCUSSION

The results obtained in the study have not been entirely positive, since as we have seen, it classifies well for those belonging to traject 3, however, it does not do the same for the other two groups. This allows us to question not only what may have affected the classification power, but also future ways to improve it, and future studies.

Starting with the explanations of why this difference, first, it should be noted that there is a large number of values equal 3 in the traject column, while there is a smaller number of the rest, also, when filling in the NaN values, it has taken the median, which despite making sense, could have also affected a correct classification, further increasing the number of these.

As it has been observed, for height, the accuracy is even worse than with respect to traject, this may be due to the way in which the data were previously classified, or due to the lack of quality of these, also for including features that do not present great importance, or because it is not an optimal target variable to consider in the classification.

Thus, this project could be improved by having a more updated database, with a greater number of classes 1 and 2, which are in the minority in this project.

During this study, an analysis of the most important features was carried out, but due to lack of time it was impossible to continue with the classification, only making use of these, so it would be a possible point to follow, in a future research, as the results obtained are expected to improve as a result of this.

Continuing with future research ideas, an interesting idea, based on the results obtained, which have shown a great accuracy when classifying correctly traject 3, would be to consider this a binary classifier, which means that in the future, this project could be used to classify the data, not between traject 1, 2 and 3, but between traject 3 and the other two together.

Another possible study, to see if it improves, although more data would be needed, would be to omit those samples that classify as traject 3, and observe the classification power available to classify traject 1 and traject 2.

The target variable height was also analyzed to see how it classified, but the results were not optimal. In this case, it is proposed for future research to look for upper and lower thresholds with greater biological significance, and also to expand the database, since the individuals presented very similar heights (due to army requirements, a minimum height is established).

BIBLIOGRAPHY

1. Steenkamp, M. M., Nash, W. P., & Litz, B. T. (2013). Post-traumatic stress disorder: Review of the Comprehensive Soldier Fitness program. *American Journal of Preventive Medicine*, *44*(5), 507-512.
2. Rosenthal, J. F., & Erickson, J. C. (2013). Post-traumatic stress disorder in US soldiers with post-traumatic headache. *Headache: The Journal of Head and Face Pain*, *53*(10), 1564-1572..
3. Eekhout, I., Reijnen, A., Vermetten, E., & Geuze, E. (2016). Post-traumatic stress symptoms 5 years after military deployment to Afghanistan: an observational cohort study. *The Lancet Psychiatry*, *3*(1), 58-64.
4. https://dougspeed.com/ldak/
5. https://dougspeed.com/bayesr-predict/
6. https://towardsdatascience.com/exploratory-data-analysis-with-pandas-profiling-de3aae2ddff3
7. https://www.kaggle.com/code/jonaspalucibarbosa/removing-outliers-within-a-pipeline
8. Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics*, *710*(1), 120-127.
9. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
10. Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).
11. Noble, W. What is a support vector machine?. *Nat Biotechnol* **24**, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565