Binary Classification: Depressed vs. Non-Depressed

Using Actigraph data from Smart Wearable devices

Chukwudi Onyema Ajoku

A Report Submitted in Partial Fulfilment of the Requirement for the Degree of
Master of Science

Module Leader: Dr. Alessandro Di Stefano

Applied Artificial Intelligence
Teesside University
Middlesbrough, England, United Kingdom
May 14, 2021

Classifying Depressed vs. Non-Depressed, Using Actigraph data from Smart Wearable devices

Chukwudi Onyema Ajoku
A0151658@tees.ac.uk

Department of Applied Artificial Intelligence
School of Computing, Engineering and Digital Technologies(SCEDT)
Teesside University, England, United Kingdom.

May 14, 2021

Abstract

The ability to bring a Machine Learning academic paper to life by understanding and implementing it in code seems to be the desire of every Machine Learning beginner. It is a skill highly sort after. This paper and its accompanying code(in Python) is with the intention to improve upon a previously published paper (Garcia-Ceja et al., 2018) which tried to understand the relationship between motor activity and the diagnoses of depression. Efforts have been made to improve upon the authors' accuracy, precision, recall and F1 scores. They have been compared side by side while justifying methodologies, algorithms and techniques chosen to do so.

1 Introduction

> "...the world is facing yet another pandemic that spares no age group, no race and no climate. It is the depression Pandemic".
>
> *(Owobu, 2021)*

Recently, there has been a massive concern over mental health-related issues. According to (WHO, 2017), depression is a leading cause of disability worldwide and contributes significantly to the global burden of disease. In most developing countries, about half of people suffering from depression are neither diagnosed nor treated, while in less developed countries, the figure ranges between 80 and 90

Using data from wearable devices to bring mental health diagnoses to everyone will solve this problem as early detection and management of depression can be a highly successful method of reducing suicide deaths.

1.1 About the data

Data used was downloaded from this website and was sourced from smart actigraph wrist watches belonging to 55 different individuals out of which 23 were unipolar bipolar and depressed patients, and 32 were healthy people. This actigraph wristwatch takes patients health details such as patients motor activity, sleep/inactivity and heart rate. In this case, each wearable device/observation is a CSV file that contains the following columns:

1. timestamp (one minute intervals)
2. date (date of measurement)
3. activity (activity measurement from the actigraph watch).

The data set is already separated into two folders, containing actigraph data (CSV files) of patients collected over time. The first group is stored in a folder called "condition", which contains actigraph data of 23 depressed patients who suffer from bipolar/unipolar. In contrast, the second folder, "control", contains actigraph data of 32 regular people with no signs of depression.

For every sample, actigraph data is taken every minute, making it 1440 observations(minutes) in one day.

2 Data exploration and preparation

Here is an example of what an observation looks like from the **control group**:

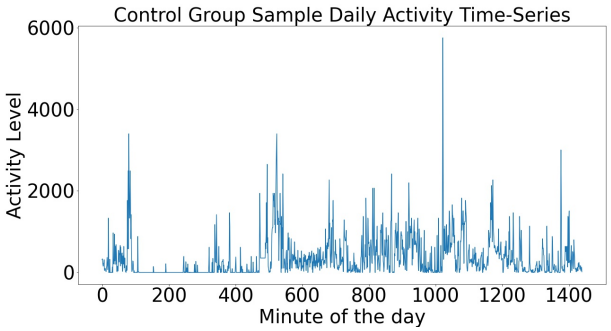

Figure 1: *An observation belonging to the* ***control group****.*

Similarly, that of the **condition group** looks the same but, the y axis is not scaled equally. This shows that the control group has higher activity than the condition group. With these differences, there is hope that separating them will be possible using a trained model when these two groups are mixed up. The trained model will likely leverage this discrepancy.

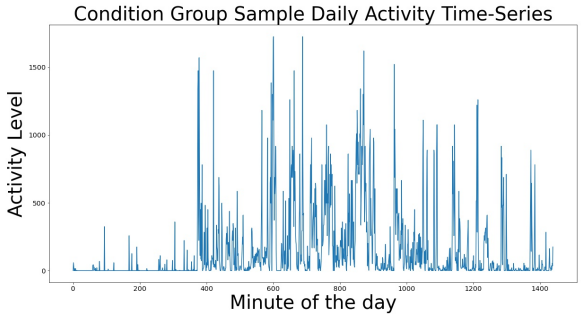

Figure 2: *An observation belonging to the* ***condition group****.*

However, looking at one random sample from each group is insufficient to make the conclusion stated above. Below is an averaged activity visual of both groups plotted in one graph for comparison:

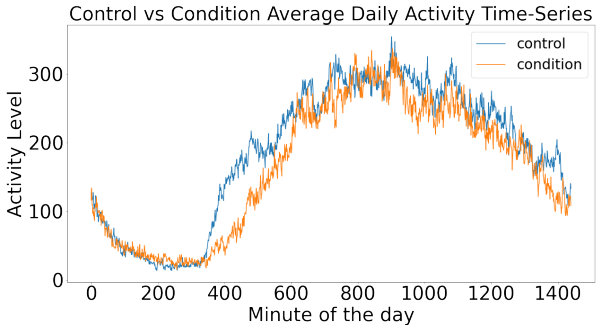

Figure 3: *An* ***averaged*** *observation of* ***both groups*** *in totality.*

Here, it can be seen that the problem gets a little bit more complex. Both groups are similar except at night-time, between the 400th and 700th minutes of the day, the control group show signs of more activity. On the y-axis, it can be seen that the maximum scale reduces to 300. This is very understandable; actigraph data are usually spiky. While one observation might be spiking in one minute of instance, other observations could be low at that exact instance. Hence, averaging eliminates the sudden spikes. Zooming in further into the area where this discrepancy occurs:

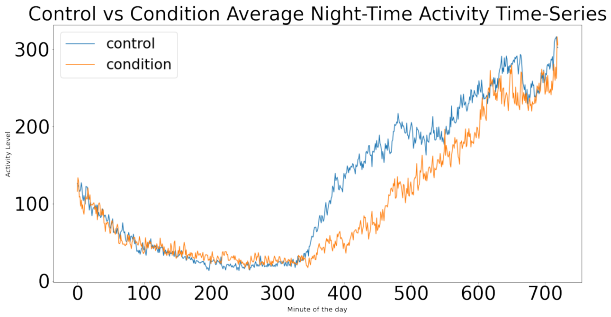

Figure 4: *An* ***averaged*** *observation of* ***both groups*** *at night-time.*

3 Experiments

Data from each group were read. The condition group were labelled "1" while the control group were labelled "0". Each observation contained 1440 time-series variables that showed the activity level per minute per day(there are 1440 minutes in a day) with its corresponding label(1 or 0).

Data from both groups were merged and shuffled. In the end, each observation in the data looked like a NumPy array with 1440 variables(showing activity levels per minute) mapped to a label(1 or 0). At this stage, completes the data cleaning and manipulation process. A few observation with less than 1440 minutes of monitoring were eliminated. Finally, after data preparation, merging and cleaning, the whole dataset translated into 1029 days of observation, with 359 (35%) belonging to the condition group and 670 (65%) belonging to the control group. This is an imbalanced class problem.

3.1 Evaluation Metrics

Accuracy score is not usually the best form of evaluation in Machine Learning (Mishra, 2018; Koehrsen, 2018). More importantly, in cases where there are oversampling or imbalanced classes as seen in the dataset. 65% of the observations are from the control group. If a model classifies **all** the observations as belonging to the control group, such a model will already be achieving 65% accuracy while being completely useless. To avoid this and to properly evaluate the performances of the models, the following evaluation metrics have been chosen:

3.1.1 Precision

Precision shows the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{True\_Positives}{True\_Positives \quad + \quad False\_Positives}$$

From the equation above, It is clear that for a good model, $False\_Positives$ should be as small as possible. Precision lies between 1(good) and 0(bad).

3.1.2 Recall

This is the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{True\_Positives}{True\_Positives \quad + \quad False\_Negatives}$$

Here, for a good model, $False\_Negatives$ should be as small as possible. Recall also lies between 1(good) and 0(bad).

3.1.3 Accuracy

Accuracy summarises the whole model. It is the ratio of the correctly classified prediction to the entire prediction. Mathematically:

$$Accuracy = \frac{Correct\_Predictions}{All\_Predictions}$$

3.1.4 F1 score

The F1 score seems to be the most suitable metrics for evaluating imbalanced data problems (Seo, 2013). This is because it incorporates both the Precision and the recall scores.

Mathematically:

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Just like Precision and Recall, the best score is 1 and the worst is 0.

3.1.5 Matthew's Correlation Coefficient(MCC)

Even more reliable is the Mathews Correlation Coefficient (MCC). This is because it takes into consideration all 4 different Confusion Matrix classes($True\_Positives(TP)$, $True\_Negatives(TN)$, $False\_Positives(FP)$ and $False\_Negatives(FN)$) (Chicco and Jurman, 2020; Shmueli, 2019). Mathematically, it is given as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC values ranges between -1 and 1. A value of 0 means that there is no correlation and the relationship was random. The further aware from zero MCC is the the better.

3.2 Procedure

3.2.1 Any further processing?

Further data preprocessing steps such as normalisation were not necessary. Even when it was implemented, the model's performance became worse because it brought many of the actigraph values to zero or too close to zero. There was no need to normalise the data as it was already scaled. The idea of ratio made sense since it can be said that someone with 300 units of activity level was twice as active as someone with 150. Normalisation was not necessary.

3.2.2 Hyper-parameter tuning

For some models, choosing the best hyper-parameter can be very daunting, but there exist some heuristic rules and guidelines one could follow. Understanding how the algorithm works and what the hyper-parameters mean is very important. Although Grid Search cross-validation was used, it was not a blind guess. For instance, In the Support Vector Machines model, choosing a range of hyperparameters in exponentially growing sequences (for example, C = $2^{-4}, 2^{-3}, 2^{-2}, ...2^{10}$ and $\gamma = 2^{-10}, 2^{-9}, 2^{-8}...2^{4}$) is a practical method to identify good parameters(Chih-Wei Hsu and Lin, 2016). Also, when one finds a good parameter, one can narrow down further and still search for better parameters within the range of the found good parameter (for instance. if C = 2^{-3} is the best parameter, a better parameter can yet exist between 2^{-2} and 2^{-4} when searched through a range of smaller intervals)

3.2.3 Crossvalidation

K-fold cross-validation of 10 n_splits was also carried out in all the training with the weighted average of the evaluation metrics collected.

3.2.4 Models Used

The main reason behind choosing this experiment and dataset is to serve as a personal evaluation, to compare personal ability with existing published results. For this reason, algorithms used in this work are the same as those used in the paper(Garcia-Ceja et al., 2018). However, it is worthy to note that the authors' exact methodologies, software, programming languages, libraries, and tools were not stated in their paper or anywhere else. For instance, while I chose LSTM for the deep learning model, what the authors used for theirs is unknown. The paper showed no technical details but only results, which I desire to improve upon.

The models used were, **K Nearest Neighbors**, **Linear Support Vector Machine(SVM)**, **RBF SVM**, **Gaussian Process**, **Decision Tree**, **Random Forest**, **AdaBoost**, **Naive Bayes**, **Quadrant Discriminant Analysis** and **Deep Neural Networks**.

4 Results

Here is a comparison between what Garcia-Ceja et al., 2018 got versus what I got.

Table 1: ***Results***

Classifier	Compare	Class	PREC	REC	ACC	MCC	F1
Nearest Neighbors	Garcia-Ceja, et al	depressed	0.395	0.705	0.675	0.318	0.5
		nondepressed	0.878	0.669	0.675	0.318	0.758
		weighted average	0.752	0.678	0.675	0.318	0.691
	Ajoku, Chukwudi	depressed	0.577	0.562	0.701	0.342	0.568
		nondepressed	0.768	0.776	0.701	0.342	0.772
		weighted average	0.673	0.669	0.701	0.342	0.670
Linear SVM	Garcia-Ceja, et al	depressed	0.577	0.721	0.727	0.433	0.638
		nondepressed	0.836	0.734	0.727	0.433	0.78
		weighted average	0.735	0.729	0.727	0.433	0.724
	Ajoku, Chukwudi	depressed	0.440	0.401	0.614	0.132	0.418
		nondepressed	0.695	0.728	0.614	0.132	0.711
		weighted average	0.568	0.565	0.614	0.132	0.564
RBF SVM	Garcia-Ceja, et al	depressed	0.546	0.732	0.724	0.426	0.622
		nondepressed	0.853	0.724	0.724	0.426	0.782
		weighted average	0.733	0.727	0.724	0.426	0.719
	Ajoku, Chukwudi	depressed	0.675	0.502	0.736	0.4	0.567
		nondepressed	0.765	0.863	0.736	0.4	0.809
		weighted average	0.720	0.682	0.736	0.4	0.688
Gaussian Process	Garcia-Ceja, et al	depressed	0.543	0.733	0.723	0.424	0.619
		nondepressed	0.853	0.723	0.723	0.424	0.781
		weighted average	0.732	0.727	0.723	0.424	0.718
	Ajoku, Chukwudi	depressed	0.713	0.524	0.758	0.446	0.598
		nondepressed	0.778	0.884	0.758	0.446	0.826
		weighted average	0.745	0.704	0.758	0.446	0.712
Decision Tree	Garcia-Ceja, et al	depressed	0.561	0.689	0.707	0.391	0.615
		nondepressed	0.813	0.72	0.707	0.391	0.763
		weighted average	0.711	0.707	0.707	0.391	0.702
	Ajoku, Chukwudi	depressed	0.584	0.477	0.697	0.309	0.518
		nondepressed	0.746	0.815	0.697	0.309	0.777
		weighted average	0.665	0.646	0.697	0.309	0.647
Random Forest	Garcia-Ceja, et al	depressed	0.509	0.702	0.7	0.375	0.585
		nondepressed	0.838	0.704	0.7	0.375	0.764
		weighted average	0.738	0.703	0.7	0.375	0.709
	Ajoku, Chukwudi	depressed	0.768	0.535	0.778	0.494	0.625
		nondepressed	0.785	0.908	0.778	0.494	0.841
		weighted average	0.777	0.721	0.778	0.494	0.733
AdaBoost	Garcia-Ceja, et al	depressed	0.523	0.706	0.706	0.387	0.595
		nondepressed	0.838	0.71	0.706	0.387	0.767
		weighted average	0.733	0.709	0.706	0.387	0.71
	Ajoku, Chukwudi	depressed	0.608	0.510	0.713	0.349	0.549
		nondepressed	0.760	0.822	0.713	0.349	0.789
		weighted average	0.684	0.666	0.713	0.349	0.669
Naive Bayes	Garcia-Ceja, et al	depressed	0.663	0.63	0.694	0.379	0.645
		nondepressed	0.716	0.747	0.694	0.379	0.731
		weighted average	0.69	0.69	0.694	0.379	0.688
	Ajoku, Chukwudi	depressed	0.437	0.713	0.576	0.21	0.540
		nondepressed	0.767	0.503	0.576	0.21	0.603
		weighted average	0.602	0.608	0.576	0.21	0.571
QDA	Garcia-Ceja, et al	depressed	0.694	0.634	0.7	0.397	0.66
		nondepressed	0.704	0.761	0.7	0.397	0.73
		weighted average	0.699	0.697	0.7	0.397	0.695
	Ajoku, Chukwudi	depressed	0.355	0.992	0.368	0.07	0.523
		nondepressed	0.847	0.034	0.368	0.07	0.065
		weighted average	0.601	0.513	0.368	0.07	0.294
Neural Nets	Garcia-Ceja, et al	depressed	0.557	0.711	0.719	0.413	0.621
		nondepressed	0.836	0.724	0.719	0.413	0.775
		weighted average	0.727	0.719	0.719	0.413	0.715
	Ajoku, Chukwudi	weighted average	0.620	0.610	0.760	0.430	0.600

It can be seen that in most cases, the discrepancies are negligible and that the results are close but not the same. This is expected because the methodologies used here may not have been the same used by the authors of ibid. Moreover, it is largely a random space, right from the n_split cross-validation, some tree-based algorithms can start off at different nodes starting at any feature. KNN, SVM and other algorithms can start off at any variable. Without setting a random seed, different results will be produced between multiple function calls.

5 Discussion

5.1 Justify 3 algorithms:

As a requirement for this ICA, I will briefly justify 3 of the best performing algorithms from the results obtained above.

5.1.1 Random Forest:

Random forest is a very robust classification algorithm because it is not prone to over-fitting. Further more, because it is a tree based algorithm, it is can automatically handle datasets with imbalanced classes as evidenced above.

5.1.2 Gaussian Process:

Gaussian Processes is known to perform wonderfully for time series related data(Roberts et al., 2012; Wikipedia, 2021). It is based on Gaussian (Normal) distribution and as such, inherits its properties as well. This was a big advantage.

5.1.3 Neural Networks:

The main advantage Neural Networks had was its ability to model non-linear and complex relationships. However, it did not seem as though there was a strong relationship.

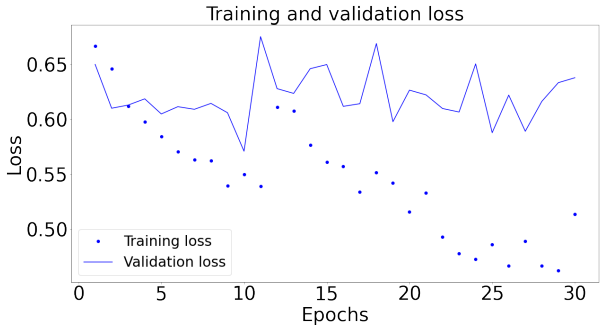

Figure 5: *Graph of validation and training loss showing randomness*

6 Conclusion

Many factors could have affected the results. Depression can be seen as a spectrum (Angst and Merikangas, 1997), with different levels of intensities ranging from mild to severe. Some can be seasonal. Those who are at the threshold or about getting depressed can be easily misclassified.

There is a need for more data for the condition group as they were in the minority. It was not said if both groups were subjected to the same daily activity because a unipolar/bipolar/depressed person might indulge in vigorous activity while the control group rests.

Another type of data that may have helped was the one containing the MADRS scores however; the scores were only available for the condition group and not in the control group. Using this dataset would have given the model 100% accuracy as the model will automatically learn to classify missing values as belonging to the control group.

Although Random forest performed the best, other algorithms can also stand a chance if the classes were balanced and bigger.

References

Angst, J and K Merikangas (1997). "The depressive spectrum: diagnostic classification and course". In: *Journal of Affective Disorders* 45.1, pp. 31–40. ISSN: 0165-0327. DOI: https://doi.org/10.1016/S0165-0327(97)00057-8. URL: https://www.sciencedirect.com/science/article/pii/S0165032797000578.

Chicco, Davide and Giuseppe Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC Genomics* 21.1, p. 6. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. URL: https://doi.org/10.1186/s12864-019-6413-7 (visited on 2021).

Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin (2016). *A Practical Guide to Support Vector Classification*. Department of Computer Science, National Taiwan University, Taipei 106, Taiwan.

Garcia-Ceja, Enrique et al. (2018). "Depresjon: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients". In: *Proceedings of the 9th ACM on Multimedia Systems Conference*. MMSys'18. Amsterdam, The Netherlands: ACM. DOI: 10.1145/3204949.3208125. URL: http://doi.acm.org/10.1145/3204949.3208125.

Koehrsen, Will (2018). *Beyond Accuracy: Precision and Recall*. URL: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c (visited on 2021).

Mishra, Aditya (2018). *Metrics to Evaluate your Machine Learning Algorithm*. URL: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (visited on 2021).

Owobu, Adaugo (2021). *Another Pandemic! Smile with Medicine*. YouTube. URL: https://www.youtube.com/watch?v=537t2qADOSU (visited on 2021).

Roberts, S. et al. (2012). "Gaussian Processes for Time-series Modelling". In: *Philosophical Transactions of the Royal Society (Part A.*

Seo, Minkoo (2013). *How to interpret F-measure values?* Cross Validated. URL:https://stats.stackexchange.com/q/49244 (version: 2015-10-05). eprint: https://stats.stackexchange.com/q/49244. URL: https:

//stats.stackexchange.com/q/49244 (visited on 2021).

Shmueli, Boaz (2019). *Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of*. URL: https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a (visited on 2021).

Siddique, Haroon (2016). *Tiny minority of people with depression get treatment, study finds*. URL: https://www.theguardian.com/society/2016/dec/01/minority-depression-treatment-study-finds (visited on 2021).

WHO (2017). *Depression*. URL: https://www.who.int/health-topics/depression#tab=tab_1 (visited on 2021).

Wikipedia (2021). URL: https://en.wikipedia.org/wiki/Gaussian_process.