MINI PROJECT – I

(2019-20)

Sentimental Analysis on Depression Dataset and Prediction on it

SYNOPSIS

Team Members

Aditi Bhatia (171500016)

Dolly (171500105)

Supervised by

Piyush Vashishth Sir

Depart of Computer Engineering & Applications

Contents

Introduction

According to Our World in Data Website, *Depressive disorders occur with varying severity. The WHO's International Classification of Diseases (ICD-10) defines this set of disorders ranging from mild to moderate to severe. The Institute for Health Metrics and Evaluation(IHME) adopt such definitions by disaggregating to mild, persistent depression (dysthymia) and major depressive disorder (severe).*

All forms of depressive disorder experience some of the following symptoms:

- *reduced concentration and attention*
- *reduced self-esteem and self-confidence*
- *ideas of guilt and unworthiness (even in a mild type of episode)*
- *bleak and pessimistic views of the future*
- *ideas or acts of self-harm or suicide*
- *disturbed sleep*
- *diminished appetite*

Materials and Methods

The methodology followed in this work is presented in Figure 1. Figure 1A shows the data that were obtained from the Depresjon database. Figure 1B presents the data preprocessing stage, consisting in the selection of samples and subjects from the original dataset, the normalization of data and the elimination of incomplete cases. The feature extraction, as shown in Figure 1C, was performed in order to extract a series of statistical features, which were subsequently subjected to a classification analysis (Figure 1D) through a random forest (RF) technique. Finally, a validation step as performed to evaluate the results obtained by measuring the receiver operating characteristic (ROC) curve and its area under the curve (AUC) correspondent value (Figure 1E).

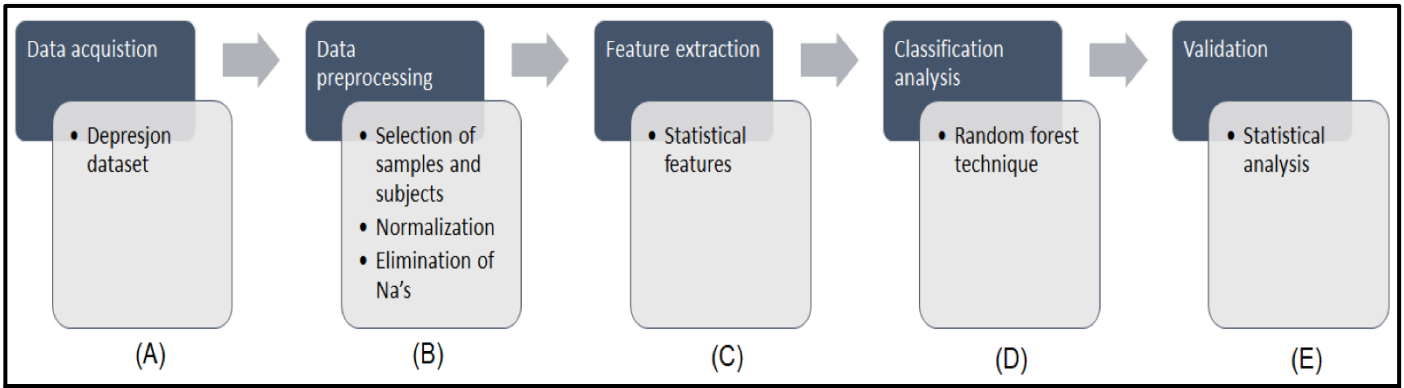

***Figure 1**. Flowchart of the methodology followed. Blue squares refers to the data processing step and gray square details task dones in each step (**A-E**).*

Data Description

The Depresjon dataset contains information of patients with absence of depression (controls) vs. patients with presence of depression (cases). In this dataset, the activity levels were monitored through an actigraph watch worn on the right wrist. The actigraph watch used was the "Actiwatch" (Cambridge Neurotechnology Ltd., Cambridge, UK, model AW4), which has a sampling frequency of 32 Hz and records movements over 0.05 g. Movements are stored in the memory unit of the watch based on the corresponding voltage that is produced, thus the number of counts is proportional to the intensity of the movements. The total activity counts were recorded in intervals of one minute.

The features collected for each subject were divided in two categories: actigraph data recorded over time and Montgomery–Åsberg Depression Rating Scale (MADRS) scores. For this work, only the features over time were used, which include: timestamp (one minute intervals), date (date of measurement) and activity (activity measurement from the actigraph watch).

***Figure 2** presents a graph with the samples collected using the Actiwatch from a control subject and from a case subject to identify the energy level over time, where the difference in energy presented by each subject is evident, with the control subject presenting higher levels.*

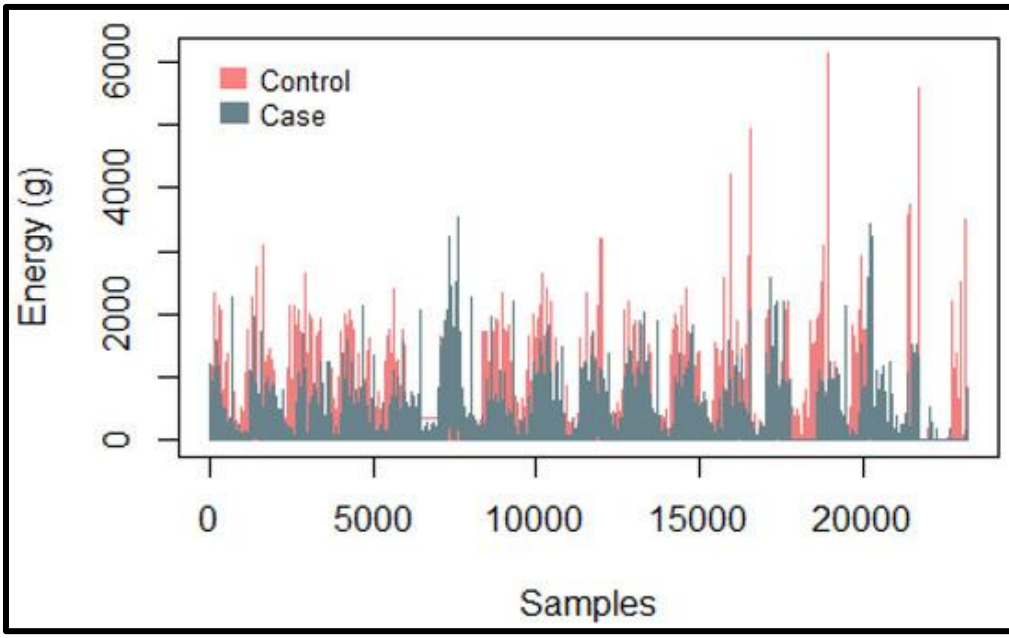

***Figure 2** Example of samples collected with the Actiwatch of a control and a case of the Depresjon database.*

The total subjects contained in the Depresjon data set were 5895 (2112 cases/3783 controls).

Data Pre-processing

The data preprocessing consisted of three main steps: the selection of samples and subjects, the elimination of incomplete cases and the normalization of data.

Is is important to mention that these preprocessing steps were used to collect a balanced volume of data to carry out the proposed methodology, as well as to make the data present a standard distribution, that is, with an average of zero and a standard deviation of one. Besides, the elimination of incomplete cases allowed avoiding bias problems and reducing the value of uncertainty due to missing values.

The number of samples collected in the original datasets is not consistent, differing in the number of minutes recorded for each subject, thus a selection of subjects and samples was made to present a balanced volume of data referring to controls and cases. In the selection of the samples, only the first value of the 60 values acquired during 1 h was kept, equivalent to the minutes correspondent to that time lapse, counting now the activity in intervals of 1 h. This procedure was performed for each hour of the total data. On the other hand, the selection of subjects depended on the volume of data resulting from the selection of samples, looking for the most balanced volume of data possible,. The first four controls present in the dataset and the first five cases were selected, thus balancing the number of samples.

The elimination of incomplete cases consisted in removing all rows where any missing value was found, represented as NA (not available).

Then, the normalization was calculated with Equation (1), where zi represents the current value normalized, xi represents the original value, μ represents the mean of the column where the value is located and σ represents the standard deviation. This step was performed to avoid overfitting problems.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Feature Extraction

The feature extraction was performed to obtain a series of 14 statistical parameters, as presented in Table 1, which were subsequently analyzed. These features were extracted from the time dependent features of the database, which were collected from the activity of the subjects through the actigraph watch.

Table 1. Statistical features collected.

Feature	Description
Mean	$\mu = 1/n \sum_{i=1}^{n} x_i$
Standard deviation	$sd = \sqrt{\sum_{i=1}^{N} (x_i - \mu)^2 / (N-1)}$
Variance	$sd^2 = 1/n \sum_{i=1}^{n} (x_i - \mu)^2$
Trimmed mean	Mean with outliers trimmed.
Coefficient of variation	$CV = sd/\mu$
Inverse coefficient of variation	$ICV = \mu/sd$
Kurtosis	$K = \mu/\sigma$
Skewness *	$S = (\mu - \upsilon)/\sigma$
Quantile * 1, 5, 25, 75, 95, 99%	$Q[i](p) = (1 - \gamma x[j] + \gamma x[j+1]$

* υ represents the median value; $1 \leq i \leq 9$, $(j-m)/n \leq p < (j-m+1)/n$; $x[j]$ represents the jth order statistic; n represents the sample size; γ is in function of j and g, where $j = floor(np+m)$ and $g = np + m - j$; and m represents a constant determined by the sample quantile type.

These statistical features were chosen because they are the first, second, third and fourth moments of an aleatory variable, which represent the descriptive measures that may be used for the characterization of the probability distribution of that variable. In other words, they describe the characteristics of the time courses of the activity measured.

Classification Analysis

In the classification analysis, the machine learning technique Random Forest (RF) was used for the classification of subjects in two different states: depressed (labeled as "1") or not depressed (labeled as "0").

RF is a non-parametric statistical method introduced by Breiman et al. , which has been widely used in different health approaches, such as the in development of models to identify high-risk surgical patients through the electronic health record data, the definition of the individual double minimum-distance of protein–RNA for the structure-based prediction , the prediction of plant-derived xenomiRs from plant miRNA sequences, the modeling of the groundwater nitrate exposure in private wells for the Agricultural Health Study, the classification of neuroimaging data in Alzheimer's disease, and the development of a three-level hepatotoxicity prediction system based on adverse hepatic effects, among others.

Validation

The evaluation of the results obtained was carried out in the validation stage through a ROC curve-based approach. The ROC curve has been widely used to measure or visualize a classifier's performance in conjunction with the AUC value to select a suitable operating point, called as decision threshold .

The two possible outputs in a classification problem are "correct" and "incorrect" for each class. This information can be represented in a confusion matrix, which is a table that shows the differences between the true and predicted classes for a set of labeled examples. This table mainly contains the true positives (Tp), true negatives (Tn), false positives (Fp) and false negatives (Fn) values, as well as the row totals with the truly negatives (Cn) and truly positives (Cp) examples, and the column totals with the predicted negative (Rn) and the predictive positive (Rp) examples. From these parameters, more meaningful measures can be extracted to have certain performance criteria, such as the accuracy, which refers to the degree to which the result of a calculation conforms to the correct value, shown in Equation (4),

$$accuracy\,(1-error) = \frac{Tp+Tn}{Cp+Cn};$$

the sensitivity, which is referred to the ability to correctly identify those with a condition, shown in Equation (5),

$$sensitivity\,(1-\beta) = \frac{Tp}{Cp}; \qquad (5)$$

the specificity, which is referred to the ability to correctly identify those without a condition, shown in Equation (6),

$$specificity\,(1-\alpha) = \frac{Tn}{Cn}; \tag{6}$$

the positive predicted value (PPV), which is the proportion of true positives results, shown in Equation (7),

$$PPV = \frac{Tp}{Rp}; \tag{7}$$

and the negative predicted value (NPV), which is the proportion of true negatives results, shown in Equation (8).

$$NPV = \frac{Tn}{Rn}. \tag{8}$$

The plotted values of the sensitivity and the specificity as the decision threshold is called ROC curve, and the simplest way to calculate the AUC is through trapezoidal integration, shown in Equation (9),

$$AUC = \sum_i (1-\beta_i \cdot \Delta\alpha) + \frac{1}{2}[\Delta\,(1-\beta) \cdot \Delta\alpha] \tag{9}$$

where $\Delta\,(1-\beta) = (1-\beta_i) - (1+\beta_{i-1})$ and $\Delta\alpha = \alpha_i + \alpha_{i-1}$.

All analyses performed in this work were carried on in "R" (version 3.4.4), which is a "free software environment for statistical computing and graphics". The libraries used were "randomForest" (version 4.6-14), "e1071" (version 1.7-0) [37], "pROC" (version 1.11.0), "caret" (version 6.0-79) and "rminer" (version 1.4.2).

Results and Discussion

In this section, the results obtained for each stage of the methodology are exposed.

Initially, from the data preprocessing step, the total subjects were reduced to 4125 (controls = 2176/cases = 1949) and the total data were standardized, causing them to have a mean = 0 and a standard deviation = 1.

According to these results, the number of data selected from the total dataset was adequate to be able to carry out the proposed methodology obtaining significant results, even when the data were slightly unbalanced, because the volume of data originally acquired was presented in greater quantity for controls than for cases.

Then, a set of 14 statistical features were extracted based in the statistical moments to know the performance of the data through the time, looking for alterations at specific times that could provide meaningful information.

Then, from the classification analysis through the RF technique, an OOB estimate of error rate of 8.95% was obtained, which implies that this percentage of the OOB samples were incorrectly classified through the internal cross-validation of RF.

It is worth mentioning that some of the main reasons for choosing this technique for the classification analysis were that it can be used for handling high-dimensional data, it performs an internal cross-validation, it only has a few tuning parameters,

it is easy to interpret even when the relationships between predictors are complex, it uses all available input variables simultaneously, and it has an intuitive structure. Besides, as it is a non-parametric method, it is not necessary to comply with any specific distribution, thus it requires less preprocessing of data compared to other statistical learning methods and it is not greatly influenced by outliers .

Then, the sensitivity, specificity and error rates were also calculated, as shown in the confusion matrix of Table 2 (where the top indices represent the predicted values, the lateral indices represent the reference values and the last column presents the error rate for the specificity and sensitivity, respectively), to measure the performance of the learning stage of the algorithm, where it is possible to observe that, from the in-bag samples, which correspond to controls = 1483/cases = 1267, the error rate of the specificity is 0.077, meaning that at least 92.3% of the controls were correctly classified, and the error rate of the sensitivity is 0.104, meaning that at least 89.6% of the cases were correctly classified.

Table 2. Confusion matrix of the subjects classification based in the RF approach.

Table 2. Confusion matrix of the subjects classification based in the RF approach.

	Control	Case	Error
Control	1369	114	0.077
Case	132	1135	0.104

These values represent statistically significant results, since a low percentage of subjects was misclassified in the learning process of RF, which implies that the information contained in

the extracted features is presenting values that allow distinguishing between the two possible classes of the subjects. This discussion is supported with the OOB error, validating through a test the performance of the learned model.

On the other hand, from the validation stage, Figure 3 is obtained, which shows the ROC curve of the performance of modeling the data through RF, obtaining an AUC of 0.893. As can be seen, the AUC value matches with the results obtained from the internal validation of RF; therefore, it also presents a significant specificity/sensitivity rate.

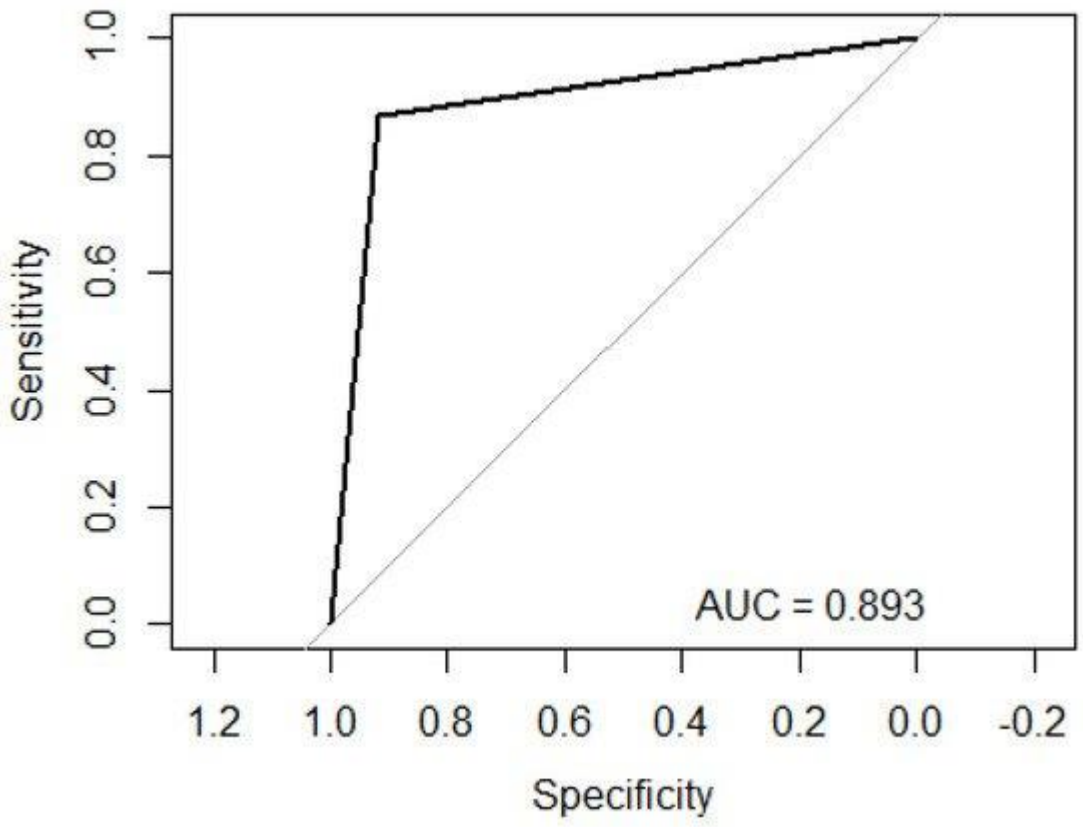

Figure 3. ROC curve obtained from the classification analysis based in RF.

Table 3 presents a confusion matrix based on a blind test, using 30% of the data (unlabeled) to evaluate the performance of the classification, where the top indices represent the reference values and the lateral indices represent the predicted values. For this test, the

subjects were randomly selected and balanced, using a total of 1375 (controls = 693/cases = 682).

Table 3. Confusion matrix of the validation through a blind test.

Table 3. Confusion matrix of the validation through a blind test.

	Control	Case
Control	637	91
Case	56	591

These results show true positives of 591 and true negatives of 637 (false positives of 56/false negatives of 91), also being statistically significant and supporting the results obtained previously.

Finally, Table 4 presents a set of measures calculated to validate the ability to classify subjects based in RF, allowing to complement and understand the significance of the performance criteria. The accuracy shows a value of 0.893, which means that, for any subject, there is a degree of 89.3% of being classified with a correct value. Then, the sensitivity shows a value of 0.867, referring that those subjects with presence of depression have a degree of 86.7% of being correctly classified. The specificity shows a value of 0.919, referring that those subjects with absence of depression have a degree of 91.9% of being classified with a correct response.

Table 4. Parameters obtained through validation.

Table 4. Parameters obtained through validation.

Parameter	Value
Accuracy	0.893
Sensitivity	0.867
Specificity	0.919
Balanced accuracy	0.892
PPV	0.875
NPV	0.931

It is important to remark that, in all results, the classification rate of controls presents better values than the classification rate of cases, which may be because in some time ranges the subjects with presence of depression can be presenting similar information to subjects with absence. Frequently, depressive patients have less activity in their daily life. This could cause confusion when a non-depressed patient shows

low activity in a specific time range, for example at night while sleeping, inflicting on the depressive patient being classified as non-depressive.

The balanced accuracy was calculated to support the value of the regular accuracy, because, even when the data were not significantly unbalanced, the number of controls exceeded the number of cases. This parameter obtained a value of 0.892, which is only 0.1% lower than the regular accuracy, validating the statistically significant result calculated previously.

The PPV and NPV, which obtained values of 0.875 and 0.931, respectively, were calculated to know the proportion of true positives and the proportion of true negatives, where the results agree with the other validation parameters and it is corroborated that the subjects with absence of depression were slightly better classified than the subjects with presence.

Finally, the results outperform the baseline performance proposed by García-Ceja et al. [41], which includes Nearest Neighbors, Linear kernel Support Vector Machine (SVM), Radial Basis Function kernel (RBF) SVM, Gaussian Process, Decision Tree, Random Forest, Neural Network, AdaBoost, Naive Bayes, and Quadratic Discriminant Analysis (QDA), as presented in Table 5.

Table 5. Machine learning techniques comparison.

Table 5. Machine learning techniques comparison.

Technique	Specificity	Accuracy
Nearest Neighbors	0.696	0.675
Linear SVM	0.726	0.727
Random Forest	0.703	0.700
Neural Net	0.716	0.719
AdaBoost	0.707	0.706
Naive Bayes	0.688	0.694
Our proposal (Feature extraction & RF)	0.919	0.893

Conclusion

This study proposed an analysis to find the relationship between a series of statistical features, based on continuous values acquired in a specific time, and the possible condition of depression.

Since the number of patients was adequate to carry out this research and the extracted features allowed describing the main characteristics of a patient's full-day activity, according to the modeling developed between them and the condition of presence or absence of depression, evaluated through statistical validation, it is possible to conclude that the results obtained through this methodology shown statistically significant values indicating that there is an association between the recorded daily activity of a patient and the condition of his depressive state.

Among the symptoms presented by patients with depression are the slowness of movement, poor body gesticulation and the feeling of fatigue, thus they tend to show lower levels of activity than subjects who do not have this condition, giving meaning to the results obtained.

Therefore, the main benefit presented in this study is a preliminary tool (bearing in mind that it is necessary to study in greater depth this approach, taking into account the regulations of the health system and characterizing the results) that may support the diagnose of specialists to know if a patient presents depression based on the level of activity he has in a full day through the automatic diagnosis of subjects obtained by submitting this information to the model developed in this

work, relating the total motor activity with the presence or absence of depression, which is shown, according to the results presented, to have a significantly high accuracy, allowing to reduce false positives and false negatives in the detection of this condition, thus improving the diagnosis of this disease..

Through this research is obtained a second automatic opinion of low cost, since the implementation of the developed model does not need any software or specialized hardware, so it is viable to be used in regions with limited access to health services.

Finally, it is important to mention that the main limitation of this study is the large volume of data on the motor activity required by each patient, which makes the analysis of the data a bit complex and delayed, causing a small number of subjects to be used. On the other hand, a limiting factor that could broaden the focus of this work would be to know information about patients in relation to any psychiatric or other treatment (e.g., sleeping pills) that they are currently undergoing, allowing us to know if this influences the amount of activity performed by the subjects, which could affect the controls by reducing their physical activity, causing false positives.

Future Work

As future work, it is proposed to increase the number of subjects in the experimentation to present results based on a greater diversity of data and thus highlight the robustness of the results. In addition, the inclusion of a stage of feature selection is proposed, comparing different machine learning tools, to know which are the features that have the greatest contribution in the classification of depressive and non-depressed subjects.

Applications

Our vision is that the available data may eventually help people to develop systems capable of automatically detecting depression states based on sensor data. This project can be suitable (but not limited to) for the following applications:

(i) Use machine learning for depression states classification;
(ii) MADRS score prediction based on motor activity data and
(iii) Sleep pattern analysis of depressed v.s. nondepressed participants.

This project can be used as the basis for evaluating different machine learning methods and approaches such as: cost-sensitive classification and oversampling techniques for imbalanced class problems. This dataset is also suitable for comparing different machine learning classification approaches such as feature based and deep learning based methods like convolutional neural networks and recurrent neural networks for time series.