Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 177 (2020) 556–561

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
November 2-5, 2020, Madeira, Portugal

Prediction of Mental Illness Associated with Unemployment Using Data Mining

Carina Gonçalves[a], Diana Ferreira[b], Cristiana Neto[b], António Abelha[b], José Machado[b,*]

[a]*University of Minho, Campus Gualtar, Braga 4710, Portugal; a81247@alunos.uminho.pt*
[b]*Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal; diana.ferreira@algoritmi.uminho.pt (D.F.); cristiana.neto@algoritmi.uminho.pt (C.N.); abelha@di.uminho.pt (A.A.)*

Abstract

Mental illness is a concern these days, affecting people worldwide and across all kinds of ages. This article aims to predict mental illness and discover its association with unemployment as well as other possible causes behind the illness. In order to accomplish this goal, a Data Mining (DM) process was performed using the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology and the RapidMiner Studio software. In the end, the results obtained were considered promising since all the evaluation metrics, namely accuracy, sensitivity, and specificity, obtained values above 90%. The study also allowed, in the end, to identify the factors associated with the prediction of mental illness.

Keywords: Mental Illness; Unemployment; Data Mining; CRISP-DM; Classification.

1. Introduction

Nowadays, there are large amounts of data being collected every day, which makes it impossible for humans to analyze it. DM appears as a process that allows anomalies, patterns, and correlations to be found in large datasets and has a wide variety of techniques aimed at different applications. At the health level, it can improve healthcare and reduce costs by predicting variables of interest [1].

For all people, mental, physical, and social health are strands of life that are closely intertwined and deeply interdependent. As the understanding of this relationship grows, it becomes increasingly evident that mental health is indispensable for the general well-being of individuals, societies, and countries [2]. From a cross-cultural perspective, it is almost impossible to define mental health in a complete way. However, it is agreed that mental health is more

* Corresponding author. Tel.: +351253604430; fax: +351253604471.
E-mail address: jmac@di.uminho.pt

1877-0509
10.1016/j.procs.2020.10.078

than the absence of mental disorders. Hence, it is important to understand mental health and, more generally, mental functioning, because therein lies the foundation on which to form a more complete understanding of the development of mental and behavioral disorders [2].

Mental disorders are associated with decreased activity in the labor market as people diagnosed with mental illness have significantly high rates of unemployment compared to the general population [3]. The truth is that when a person loses his/her job, he/she not only loses his/her source of regular income but also his/her personal relationships with co-workers, daily routines, and the desire to strive to self-overcome. Unemployment can be, and often is, a shock to the entire personal system.

Hence, this article intends to predict the existence of mental illness associated with unemployment. This research intends not only to assess whether mental illnesses may be behind unemployment but also to assess other causes that may influence mental illness. The remainder of the paper is organized as follow: the next section describes the DM process carried out, being organized according to each step of the CRISP-DM methodology, then the discussion of the results is presented and, finally, the main conclusions outlined with this work are disclosed.

2. Methodology

This study followed the CRISP-DM methodology, which aims to increase the success of DM projects, making them less expensive, more reliable, repeatable, manageable, and faster [4,5]. It is divided into six phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [4,6], which will be described throughout this section. The RapidMiner tool was also used to conduct this study, which contains a multitude of operators to import and prepare datasets as well as to apply machine learning algorithms and evaluate them.

2.1. Business Understanding

This initial phase focuses on realizing the objectives and requirements of the project from a business perspective and then on transforming that knowledge into a definition of the DM problem by also making a preliminary plan for the project to achieve the objectives outlined [5,6]. Thus, the objective of this study is to predict whether individuals have mental illness, where the target attribute "I identify as having a mental illness" can be "0" or "1", where “0” means absence of mental illness and “1” means presence. This prediction fits in the scope of classification problems and will help to understand whether having a mental illness may or may not be related to unemployment. If this connection is verified, at the end of this study, measures may be taken by the competent entities so that the people with mental illnesses do not get further harmed in their lives.

2.2. Data Understanding

At this stage, it is important to explore the data, thus allowing to identify potential problems in its quality [4,6]. The dataset used in this work is the result of a research survey carried out by Michael Corley in 2018 using the Survey Monkey tool [7]. The dataset contains 334 instances and 31 attributes, which are presented in Table 1. The attributes cover demographic characteristics such as *Gender*, clinical conditions such as *Anxiety*, mental illness history such as *I have been hospitalized before for my mental illness*, employment/income situation such as *I am unemployed* and also familiarity with technology such as *I have my regular access to the internet*. Analyzing the attributes in greater detail, it is noticed that the *Age* attribute is divided into age ranges instead of being in the exact values, i.e, this attribute is already discretized. It is also important to note that the *Gender* attribute is balanced with 176 female and 158 male individuals. Furthermore, a total of 45 missing values were detected in the following attributes: *Days of hospitalization, Region, Lack of concentration, Obsessive thinking, Mood swings, Panic attacks, Compulsive behavior*, and *Tiredness*.

Table 1. Attributes.

Attribute	Type
I am currently employed at least part-time	Binominal
I identify as having a mental illness	Binominal
Education	Polynomial
I have my own computer separate from a smart phone	Binominal
I have been hospitalized before for my mental illness	Binominal
How many days were you hospitalized for your mental illness	Integer
I am legally disabled	Binominal
I have my regular access to the internet	Binominal
I live with my parents	Binominal
I have a gap in my resume	Binominal
Total length of any gaps in my resume in months	Binominal
Annual income (including any social welfare programs)	Integer
I am unemployed	Binominal
I read outside of work and school	Binominal
Annual income from social welfare programs	Integer
I receive food stamps	Binominal
I am on section 8 housing	Binominal
How many times were you hospitalized for your mental illness	Integer
Lack of concentration	Binominal
Anxiety	Binominal
Depression	Binominal
Obsessive thinking	Binominal
Mood swings	Binominal
Panic attacks	Binominal
Compulsive behavior	Binominal
Tiredness	Binominal
Age	Polynomial
Gender	Polynomial
Household Income	Polynomial
Region	Polynomial
Device Type	Polynomial

2.3. Data Preparation

The data preparation phase covers all activities that are important for building the final dataset from the initial raw data [5,6]. These tasks include the selection of data for inclusion or exclusion, the possibility of creating new attributes or transforming the existing ones, as well as data cleaning [1].

As previously mentioned, the dataset had 45 missing values, 37 out of which were from the *Days of Hospitalization* attribute. Hence, the missing values of this attribute were replaced by the average value through the *Replace Missing Values* operator. Since the remaining missing values were contained in the same instances, the *Filter Examples* operator with the condition class equal to "*no_missing_attributes*" was used for those instances. Thus, after this step, the dataset had 331 instances instead of 334.

The label attribute was not balanced, as it can be seen in Figure 1. So, it was necessary to use an oversampling method, through the *Sample (Balance)* operator, to achieve an equal distribution by replicating the cases of the minority class. In this way, the final dataset had a total of 504 instances.

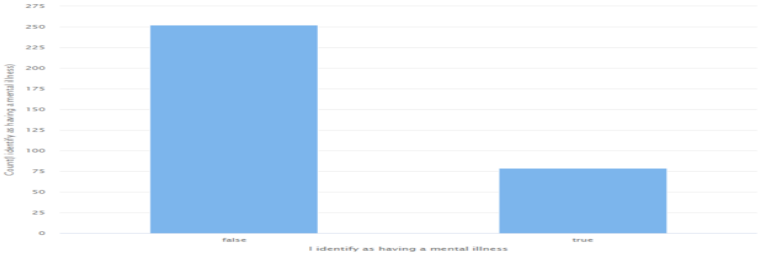

Fig.1. Distribution of the target attribute *I identify as having a mental illness*.

2.4. Modeling

In this step, the machine learning algorithms to be used must be chosen to construct different DM models [4,8]. In order to find the best classifiers for the problem at hand, the *Compare ROCs* operator was used, which allows an initial filtering of the available algorithms. A ROC (Receiver Operating Characteristic) curve plots two parameters, True Positives Rate and False Positives Rate, at different classification thresholds. The *Compare ROCs* obtained the following best algorithms: Random Forest (RF), Gradient Boosted Trees (GBT), K-Nearest Neighbor (kNN) and Decision Tree (DT). For each of these DM techniques, two sampling methods were tested: Split Validation with 70% of the data for training and 30% for testing and Cross Validation with 10 folds.

The weights of the attributes were evaluated with the operator *Weight by Correlation*, which indicates the weight of the correlation of each attribute with the label attribute. According to these results, 3 scenarios were defined in order to assess which attributes were more related to the prediction of the label attribute. The first scenario (S1) contains all the attributes. In the second scenario (S2), all attributes weighing less than 0.09 were removed, leaving behind S2 = {*I have my own computer separate from a smart phone, I have been hospitalized before for my mental illness, How many days were you hospitalized for your mental illness, I am legally disabled, I live with my parents, I have a gap in my resume, Total length of any gaps in my resume in months, I am unemployed, How many times were you hospitalized for your mental illness, Lack of concentration, Anxiety, Depression, Obsessive thinking, Mood swings, Panic attacks, Compulsive behavior, Tiredness, Age*}. On the other hand, in the third scenario (S3) only the *Depression* and *Anxiety* attributes were removed, as they had a very high correlation weight with the label attribute so they will take a higher influence in the label's prediction, which can be misleading or disguise the other features.

Two data approaches were also tested, one with oversampling and other without applying sampling techniques. In total, 48 models were tested.

2.5. Evaluation

In the final phase, it is necessary to evaluate the results obtained and review the steps performed in detail [1,8].

The performance of all the models tested was evaluated using the confusion matrix, which contains the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). With these values, it is possible to obtain the metrics of accuracy, sensitivity, and specificity. Accuracy measures the model's ability to capture true positives as being positive and true negatives as being negatives [7]. The sensitivity calculates how many of the true positives the model captures as being positive [7]. Specificity calculates how many of the true negatives the model captures as being negatives [7]. The best results of these three metrics for each algorithm are shown in Table 2, 3, and 4.

Table 2. DM models with the best accuracy results for each DM technique.

DM Technique	Scenario	Sampling Method	Data Approach	Accuracy
RF	S1	Cross Validation	With Oversampling	94.24%
GBT	S1	Cross Validation	With Oversampling	91.67%
K-NN	S2	Cross Validation	With Oversampling	84.10%
DT	S1	Split Data	With Oversampling	92.76%

Table 3. DM models with the best sensitivity results for each DM technique.

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
RF	S1	Cross Validation	With Oversampling	98.78%
GBT	S2	Split Data	With Oversampling	94.74%
K-NN	S2	Cross Validation	With Oversampling	83.78%
DT	S2	Split Data	With Oversampling	97.37%

Table 4. DM models with the best specificity results for each DM technique.

DM Technique	Scenario	Sampling Method	Data Approach	Specificity
RF	S3	Cross Validation	Without Oversampling	95.22%
GBT	S1	Split Data	Without Oversampling	93.42%
K-NN	S1, S2, S3	Split Data	Without Oversampling	96.05%
DT	S3	Cross Validation	Without Oversampling	94.80%

Looking at the tables presented, it is noticed that the highest values of accuracy are between 84% and 95%, those of sensitivity range from 83% to close to 99% and those of specificity are between 93% and 96%. Table 5 shows the 3 best models, which were chosen based on the best values of accuracy.

Table 5. DM models with the best accuracy results.

DM Technique	Scenario	Sampling Method	Data Approach	Accuracy	Sensitivity	Specificity
RF	S1	Cross Validation	With Oversampling	94.24%	98.78%	89.66%
DT	S1	Split Data	With Oversampling	92.76%	97.37%	88.16%
RF	S2	Cross Validation	With Oversampling	92.26%	96.40%	88.12%

3. Discussion

Analyzing the results shown in Tables 2, 3, 4, and 5, it is easily noticed that the best values of accuracy and sensitivity were obtained using Oversampling. This is an expected behavior since the dataset was not balanced. With the application of oversampling techniques, a balanced dataset is obtained but with little data variability because the instances are replicated. In Table 2, it is also observed that most of the best accuracy results used Cross Validation. This is due to the fact that in the Cross Validation technique all data are used for training, whereas with the Split Validation technique only a percentage of the data is used.

On the other hand, looking at Table 4, where the specificity values are presented, it is noted that the best results were all obtained without oversampling, which means that, in this case, the unbalanced dataset obtained better values. This is expected as specificity summarizes how well the negative class was predicted, which in this case corresponds to the majority class that is the one without mental illness, leading to a higher rate of true negatives, which in turn leads to higher specificity values.

The algorithm that obtained the best results was the RF followed by the DT. Of all the algorithms, the worst was kNN, but its results were not bad, just slightly worse than the others. Table 5 presents the best models according to their accuracy. It was thought that this metric would better to evaluate the models, since here, in addition to wanting to predict mental illness, we also want to identify possible causes that are behind the disease, that is, it is not only essential to evaluate the calculation of true positives or true negatives, but both, so that in this way it is possible to understand what is associated with the disease and what is not. These results, besides giving the best accuracy results, also contain the best sensitivity results. Thus, it is possible to realize that S1 was the best scenario, which in turn was the scenario that contained all the attributes, this leads to the desire that all these attributes are more or less related to the label attribute and therefore all of them in a way are important for predicting the disease. S2, the scenario where attributes with lower correlation weights had been eliminated, also obtains good results, this is not surprising since it is normal that the more related attributes help to better predict the disease and moreover this is important so that if eventually other studies are done it is already known that attributes such as Gender, Region and Device type for example do not need to be used. The S3 scenario had the worst results, since in this scenario two attributes had been removed that were strongly correlated with the label attribute, this means that there is really a great possibility that people with the Depression and Anxiety attributes, develop mental illness.

Finally, considering one of the initial objectives, the unemployment attribute's results were analyzed, with which was possible to infer that no linear relationship between the disease and unemployment was identified, meaning that most predictions referring mental illness did not correspond to unemployed people. However, it was found that most predictions with mental illness were associated with people suffering from depression and anxiety.

4. Conclusion

Nowadays, in the society we live in, it is increasingly important to worry about mental illness as it is a disease often underestimated. Studies of this nature are very important for the forecast of mental illness to be made in time in order to avoid greater evils and for finding the factors that are more easily associated with mental illness. The results obtained with this study were satisfactory, achieving accuracy, sensitivity, and specificity values above 90%. The best model achieved a value of 94.24% of accuracy, 98.78% of sensitivity and 89.66% of specificity. This model used the Random Forest, the first scenario, 10 folds Cross Validation, and oversampling. In future studies, the number of instances of the dataset should be increased, including more cases of the minority class, so that there is no need to apply oversampling techniques and therefore more realistic results can be obtained. Finally, although the model has obtained positive results, if it is used as clinical decision support, different DM techniques and new models should be considered for further testing.

Acknowledgements

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/00319/2020.

References

[1] Ferreira, Diana, Sofia Silva, António Abelha, and José Machado. (2020) “Recommendation System Using Autoencoders.” *Applied Sciences* **10** (**16**): 5510.

[2] World Health Organization. (2002). “World Health Report. Mental Health - new conception, new hope.”. World Health Organization.

[3] Luciano, Alison, and Ellen Meara. (2014) "Employment status of people with mental illness: national survey data from 2009 and 2010." *Psychiatric Services* **65** (**10**): 1201-1209.

[4] Neto, Cristiana, Maria Brito, Vítor Lopes, Hugo Peixoto, António Abelha, and José Machado. (2019). “Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients.” *Entropy* **21** (**12**): 1163.

[5] Wirth, Rüdiger, and Jochen Hipp. (2000) "CRISP-DM: Towards a standard process model for data mining." Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, UK: Springer-Verlag.

[6] Pereira, Sónia, Filipe Portela, Manuel F. Santos, José Machado, and António Abelha. (2015) “Predicting type of delivery by identification of obstetric risk factors through data mining.” *Procedia Computer Science* **64**: 601-609.

[7] Corley, Michael. (2019). Unemployment and mental illness survey, Exploring the causation of high unemployment among the mentally ill. Version 2. Retrieved on June 10, 2020 from *https://www.kaggle.com/michaelacorley/unemployment-and-mental-illness-survey*

[8] Silva, Cristiana, Daniela Oliveira, Hugo Peixoto, José Machado, and António Abelha. (2018) “Data Mining for Prediction of Length of Stay of Cardiovascular Accident Inpatients”. *International Conference on Digital Transformation and Global Society*. Springer, Cham.