Simulation of Suicide Tendency by Using Machine Learning

Hugo D. Calderon-Vilca
Department of Computer Science, School of Engineering
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
pcsihcal@upc.edu.pe

William I. Wun-Rafael
Department of Computer Science, School of Engineering
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
pcsihcal@upc.edu.pe

Roberto Miranda-Loarte
Department of Computer Science, School of Engineering
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
pcsihcal@upc.edu.pe

Abstract— **Suicide is one of the most distinguished causes of death on the news worldwide. There are several factors and variables that can lead a person to commit this act, for example, stress, self-esteem, depression, among others. The causes and profiles of suicide cases are not revealed in detail by the competent institutions. We propose a simulation with a systematically generated dataset; such data reflect the adolescent population with suicidal tendency in Peru. We will evaluate three algorithms of supervised machine learning as a result of the algorithm C4.5 which is based on the trees to classify in a better way the suicidal tendency of adolescents. We finally propose a desktop tool that determines the suicidal tendency level of the adolescent.**

Keywords— Suicide tendency, suicide, suicide attempt, suicide risk, machine learning, prevention, classification.

I. Introduction

Suicide is one of the main causes of death worldwide. According to the health World Organization (OMS) [1], around the world, about 800,000 people commit suicide each year, and what is most worrying is that, for each suicide that takes place, many other suicide attempts occur. By 2015 a statistic shows that 78% of suicides occur in low-income countries. According to the Health World Organization (OMS), suicide represents nearly half of all violent deaths, which is almost 1 million victims a year; as well as billions of dollars in economic costs. Estimations indicate that, by 2020, the amount of victims might increase to 1.5 million.

Days before the International Day for Suicide Prevention, the representative of the Mental Health National Institute in Peru, stated [13] that in the last decade, the suicide rate has remained between 3 and 4 per 100,000 inhabitants; however suicide attempts and suicide thoughts have increased. During the year 2016, 295 people committed suicide; 41% of them occurred due to domestic violence, 30% due to emotional problems and 28% due to physical and psychological bullying (through social networks).

Suicide is a complex problem, with several causes which are related to each other. The most relevant cause is the mental illness that comes from desperation, which has caused suicide tendency to increase enormously [2]. In addition, according to the severity of the problem and the result of the psychological analysis, it might be necessary to keep the person under observation, medicate it or internalize it so that it doesn't attempt to attack itself. For all these reasons the different branches of science and health (emphasizing psychology as the main branch to address this issue) have been working on identifying the traits related to suicidal tendency in order to prevent suicide from occurring.

Sometimes it's too difficult to prevent the people from having feelings of guilt, shame, and feelings of being a burden to others; which eventually can trigger suicidal thoughts or suicide itself. The situations that may be the cause of this problem can vary, for example, the aging of the person, the death of a loved one, the consumption of drugs or alcohol, emotional trauma, a serious illness, unemployment, economic problems, etc. [8]. Risk factors may also vary depending on age and gender. As for adolescents, the main causes are abandonment, mistreatment, and deliberate self-aggression. Other causes may involve a strong sentimental rupture or a family member who previously committed suicide.

There are studies that help to identify and prevent depressive and suicidal behaviors [6]. These studies [7] also allow predicting the mental state of the people. There is also an investigation on how the mental disorder is strongly related to suicide. Several investigations use real data to construct the model, however, since the profiles of registered cases of suicide and their causes are not shown in detail by the institutions like the National Police of Peru; it's not possible to construct a model of classification. Therefore, the reason of this investigation is to construct a model of classification of the suicide tendency in adolescents by the simulation and based on data generated stochastically and adjusted to reflect the Peruvian reality. The obtained data are very close to the reports presented in [12] where 25, 9% of adolescents have had suicidal thought at some point in their lives. According to the last investigation conducted in Lima and Callao in 2012, 4.9 % of the interviewed adolescents had concrete plans to commit suicide, and 3.6% of them tried to do it. To balance the data it was also taken into account the fact that 700 thousand of Peruvian people suffer from depression (according to the national institute of mental health). This disease causes 80% of suicides in the Peruvian country. Out of every 20 people with

978-1-5386-3483-7/17/$31.00 ©2017 IEEE

depression, one tries to commit suicide; and every 20 suicide attempts, one is committed.

The aim of the present investigation is to determine if a teenager has a tendency towards suicide using learning classification algorithms. The answers that the user provides are based on the suicidal risk factors during the adolescence, studied in [14]. The practical guide has been used, this guide helps to evaluate and take the behavior according to the score obtained.

In this article the Section I is presented as an Introduction, in Section II the State of Art, in Section III the Simulation and Dataset, Section IV the Experiment and Section V Results and discussion.

II. ESTATE OF ART

In the studies carried out in [6] the authors identify areas that might help identify and prevent depressive and suicidal behaviors; these studies rely on the connectivity that people are currently in. Basing the study on Smartphone devices and social networks, it can be inferred that people with depressive symptoms tend to group together, although many women have less social grouping than normal and tend to show emerging episodes of depressive and suicidal behavior. In this study machine learning was used in order to determine suicidal behavior in social networks such as Twitter. The authors categorized the danger of suicide in 4 levels, where "strongly concerning" is the most dangerous one (in the end 14% of people ended up in this strongly concerned level and they had depressive and suicidal behavior). Also, a supervised machine for vector support learning algorithm was used in which 80% of diagnostic accuracy was obtained. The study concludes that suicide prevention is a relatively new issue for engineering and its community. The authors recommend developing applications to take advantage of the technology for the prevention of people with suicidal behaviors. In addition, after the study, it is recommended to be very careful when addressing the issue of suicide since it is an extremely sensitive issue.

Other researchers such as [7] conducted a study to predict the mental state of people; they stored their data in a cloud environment with Machine Learning techniques. In this system it is proposed to analyze the data and symptoms in real time. The data are collected and stored as historical information, using the Machine Learning Viterbi algorithm, this algorithm has been implemented to analyze the data and generate the results. For the data collection they used electroencephalography, which are sensors that stick to the body and perform the analysis by waves. All data is stored on an IAAS server in the cloud (when patients tell hallucination stories, their fears, their problems), where the hospitals can access and keep their patients' information updated.

The results obtained show an accuracy of 48.11% without sensor observations, or an accuracy of 74.02% with EDA sensor on patients, or an accuracy of 83.21% with EEG and EDA with patient history, and an accuracy of 86.37% with EEG, EDA and BVP with sensors and patient history.

Finally, the authors state that these are new techniques and it is expected that they can be implemented to reduce the risk of suicidal tendency.

In the study conducted in [3] they investigated how the mental disorder is strongly related to suicide. They mention that, due to ignorance and social shame, this disease is usually ignored and undiagnosed. Their study determined that 1 out of 4 people are affected by anxiety disorder (10% is depression). They used Data Mining techniques to predict a patient's stress level. They used the Bayesian Red model to discover quickly and efficiently the different factors that affect the mental health of a patient, taking into account 140 data of students and obtaining 22.7% of positive results.

The study concludes that stress prediction and generated rules will act as a supportive tool to assist expert doctors; this would reduce the cost of various medical tests and would make it easier for patients to take preventive measures in advance.

In another research they used algorithms in order to classify the text related to the suicide in Twitter [4], because it's a social network where information can be propagated to millions of people in a matter of minutes. They have analyzed the suicidal thought that users express though social networks. They have used the Rotation Forest algorithm, in which they obtained an F measure of 0.728 in general, and 0.69 for the suicidal thoughts group. From their analysis, they found that the word lists and regular expressions extracted from discussion forums related to online suicide and other micro blogging websites, allowed to capture relevant keywords related to suicide.

III. SIMULATION AND DATASET

Taking into account the suicidal risk factors in adolescence presented in [14], the practical guide that helps to evaluate and adopt a behavior according to the obtained score, and adding the field "Attribute" nickname of the Item to generate.

TABLE I. PRACTICAL GUIDE TO DETECT SUICIDAL RISK IN ADOLESCENCE

Nº	*Item*	*Score*	*Attribute*
1	Come from a broken home	1	Provenir
2	Progenitors with mental illness	2	Progenitor
3	Family history of suicidal behavior	3	Afamiliar
4	History of learning disorders, school escapes, bad adjustment to scholarships or military regime	2	Historia
5	Personal history of self-destructive behavior	4	Conducta
6	Obvious changes in habitual behavior	5	Cambios
7	Friends with suicidal behavior	2	Amigos
8	Presence of suicidal thoughts and its variants (gestures, threats, suicidal plan)	5	Ideacion
9	Personal history of mental illness	4	Enfermedad
10	Current conflict (family, partner, school, etc.)	2	Conflicto

R language is used to generate 10000 instances, considering the items with their respective weights as attributes that allow the balancing, by repeating values is possible to increase the possibility of choosing of specific elements. This is by applying the random sampling proposed by [15]. The methods allows to randomly obtain data from any of the groups to establish a more balanced proportion between groups, weather by removing elements or by adding others.

```
    N = 10000;
    Provenir =
       sample(c('n','n','n','s'),
       N, replace = TRUE)
    Progenitor =
       sample(c('n','n','n','s'),
       N, replace = TRUE)
    Afamiliar =
       sample(c(1,1,1,2,3),
       N, replace = TRUE)
    Historia =
       sample(c('n','n','n','s'),
       N, replace = TRUE)
    Conducta =
       sample(c(1,1,1,2,2,3,4),
       N, replace = TRUE)
    Cambios =
       sample(c(1,1,1,1,1,2,2,2,3,3,4,5),
       N, replace = TRUE)
    Amigos =
       sample(c('n','n','n','s'),
       N, replace = TRUE)
    Ideacion =
       sample(c(1,1,1,1,1,2,2,2,3,3,4,5),
       N, replace = TRUE)
    Enfermedad =
       sample(c(1,1,1,1,1,1,2,2,3,4,4),
       N, replace = TRUE)
    Conflicto = sample(c('n','n','n','s'),
  N, replace = TRUE)
```

Fig. 1. Generating instances for Attributes

Although suicide risk assessment is a complex task, in [8] the suicide risk classification is presented as levels based on interview:

Minor risk (leve): there are suicidal thoughts without concrete plans to hurt itself. There is no obvious intention but there are suicidal thoughts. The person is able to rectify its behavior and become self-critical.

Moderate risk (moderado): there are plans with suicidal thoughts, possible antecedents of suicidal attempts, additional risk factors. There may be more than one risk factor without a clear plan.

Serious Risk (grave): There is a concrete determination to hurt itself. The person may have had a previous self-elimination attempt, there are more than two risk factors, hopelessness, the person rejects social support and it does not rectify its ideas.

Extreme Risk (extremo): several self-elimination attempts with several risk factors, self-aggression may be present as an aggravation.

The scores in the practical guide for the detection of suicidal risk in adolescence have been used in the group labeling in this research. For each instance the accumulated score for each attribute is evaluated defines the corresponding label.

```
    for(i in 1:N)
    {
     Puntaje = 0;
     if (Tendencia$Provenir[i]=="no")
       Puntaje = Puntaje + 0;
     if (Tendencia$Provenir[i]=="si")
       Puntaje = Puntaje + 1;
     ......

     if (Tendencia$Ideacion[i]==1)
       Puntaje = Puntaje + 1;
     if (Tendencia$Ideacion[i]==2)
       Puntaje = Puntaje + 2;
     if (Tendencia$Ideacion[i]==3)
       Puntaje = Puntaje + 3;
     if (Tendencia$Ideacion[i]==4)
       Puntaje = Puntaje + 4;
     if (Tendencia$Ideacion[i]==5)
       Puntaje = Puntaje + 5;

     if(Puntaje >= 0 & Puntaje <= 11 )
       Tendencia$Riesgo[i] = "Leve"
     if(Puntaje >= 12 & Puntaje <= 15 )
       Tendencia$Riesgo[i] = "Moderado"
     if(Puntaje >= 16 & Puntaje <= 18 )
       Tendencia$Riesgo[i] = "Grave"
     if(Puntaje >= 19 & Puntaje <= 30 )
       Tendencia$Riesgo[i] = "Extremo"
 }
```

Fig. 2. Labels to the suicidal tendency group.

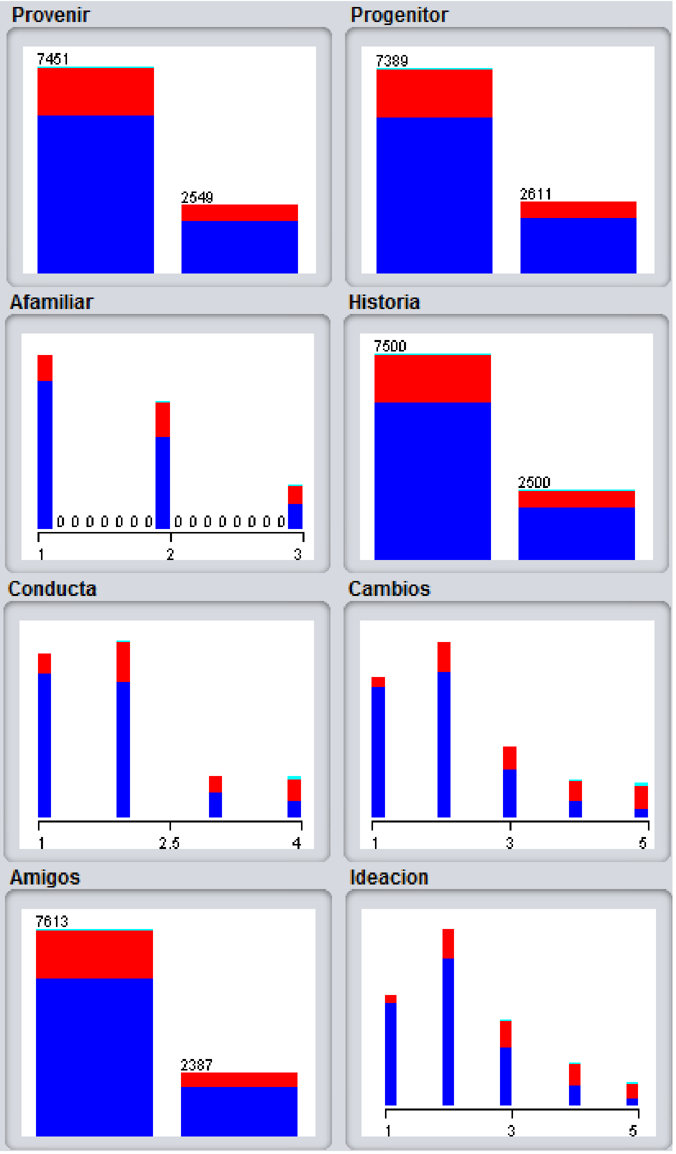
Writing the dataset compatible with Weka with the generated data.

```
    @relation Suicidio

    @attribute Provenir {n,s}
    @attribute Progenitor {n,s}
    @attribute Afamiliar numeric
    @attribute Historia {n,s}
    @attribute Conducta numeric
    @attribute Cambios numeric
    @attribute Amigos {n,s}
    @attribute Ideacion numeric
    @attribute Enfermedad numeric
    @attribute Conflicto {n,s}
    @attribute Riesgo {Leve,Moderado,Grave,Extremo}

    @data
```

```
n,n,1,n,4,1,n,2,3,n,Leve
s,n,1,n,4,3,n,4,1,n,Moderado
n,n,1,n,4,5,n,5,4,n,Grave
n,n,3,n,4,5,s,4,4,n,Extremo
……..
```

Fig. 3. Generated dataset compatible with Weka

In the pre-processing the data that reflect the behavior of the adolescence population are adjusted in a better way. In this simulation the following reports are shown:



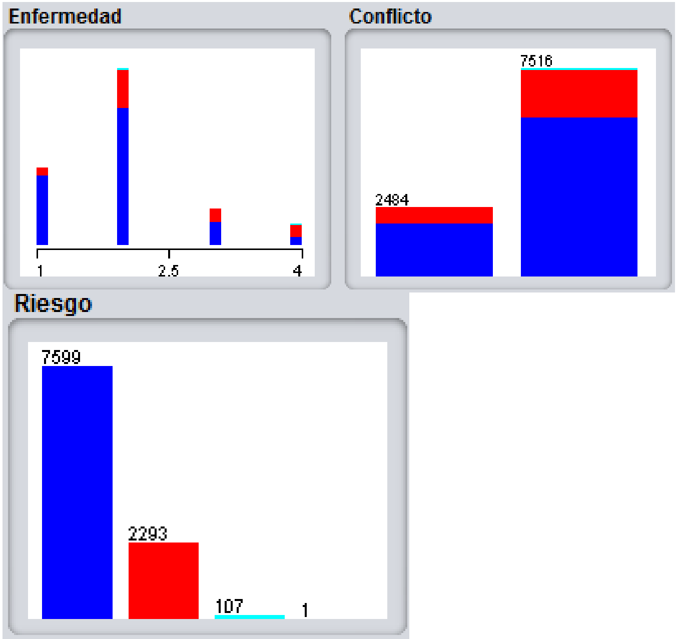


Fig. 4. Displaying data generated by attribute

The attribute Risk group that is labeled as minor or not-existent is observed 7599 people in this group. The rest has a tendency to suicide. This reflects that the population behavior is close to the shown report in [12] 25.9% of adolescents who have had desires of dying at some point of their lives, according to the study. The other attributes such as come from, Provenir, Progenitor, Historia, Amigos y Conflicto, also show a 75% each one. The ones who had concrete plans to commit suicide represent a 4.94%, this is counted by instances with the attribute Ideacion > = 5, when contrasting this with the study [13] of the interviewees, it can be seen that 4.9% had concrete plans for commit suicide. The one who suffer from depression are counted instances with the attribute Enfermedad >= 3 they ended up being 1878 people (this is 78.22% of 2401 that represents Moderado+Grave+Extremo), which approximates to 80% comparing with [13], and resulting in one million 700 thousand Peruvians who suffer from depression.

This disease causes 80% of suicides in Peru, out of every 20 people with depression, one tries to commit suicide; and every 20 suicide attempts, one is done.

From what is written above it can be confirmed that the generated data simulate the behavior of the adolescent population with a tendency towards suicide, these data are consistent for the experiment, therefore this dataset is taken for the training of the classification model of suicidal tendencies that is applied in the experiment.

IV. Experiment

For the experiment three algorithms have been evaluated, which indicate which of the algorithms constructs the best model of suicide tendency classification for adolescents, the first chosen algorithm was JRIP based on rules, the second was C4. 5 of the decision tree family and the last algorithm was Naive Bayes, based on probabilities.

For the three algorithms the Weka tool has been used for model training and test with the "Percentage Split" option. 80% of data that was used for training and 20% was used for testing.

By doing a test with the JRip algorithm with the parameters: pruning 3, minimum total weight per rule 2, amount of runs per optimization 2 and seed for random data 1, it was obtained 97.4% of correctly classified instances compared to 2.6% incorrectly classified, with a 0.93 concordance level between the observed and the randomly expected ones, and 92 constructed rules. The accuracy was 0.974, Precision 0.975, Recall 0.974 and F-Measure 0.974.

With the algorithm C4.5, with parameters: confidence factor for the pruning 0.25, minimum number of instances per paper 2. As a result we have 98.4% of correctly classified instances compared to 1.6% incorrectly classified, with a level of concordance 0.96 between the observed and the randomly expected ones. Number of papers 275, size of the tree 549, Accuracy 0.984, Precision 0.983, Recall 0.984 and F-Measure 0.984.

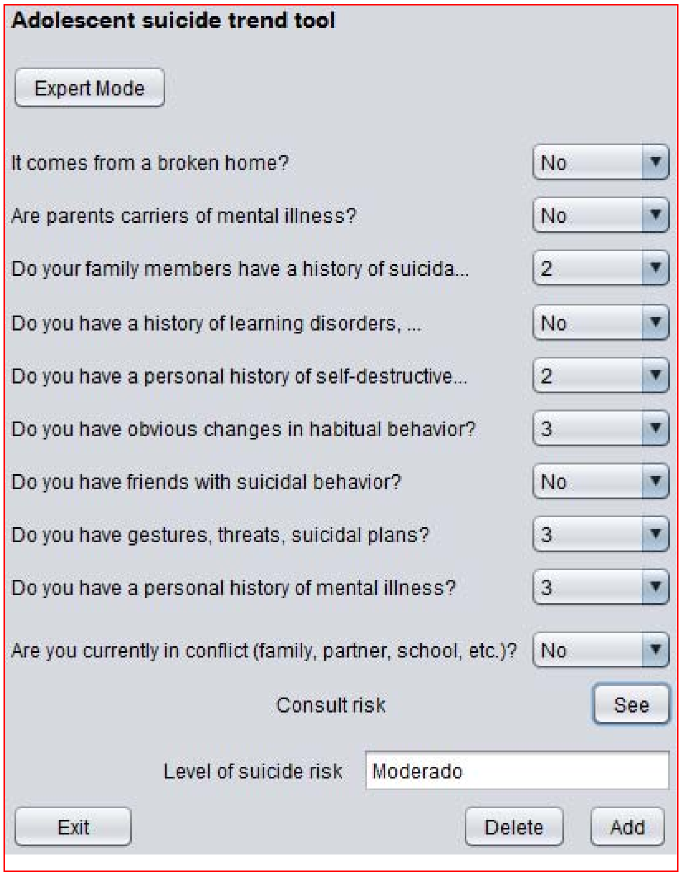
Testing with the Naive Bayes algorithm the results are: 98.65% of correctly classified instances versus 10.35% incorrectly classified, with a level of concordance 0.71 between the observed and the randomly expected ones Accuracy 0.987, Precision 0.885, Recall 0.897 and F-Measure 0.888.

When comparing the results of the evaluated algorithms, the C4.5 algorithm obtains the best measure in F-measure with 0.984. Other indicators indicate that the best classifier model for this investigation is C4.5; also it obtains 98.4% of correctly classified instances versus 1.6% incorrectly classified instances. Therefore it is recommended using this algorithm for the model of suicide tendency training for adolescents.

V. Results and Discussion

In this work, data have been generated by simulating the population behavior with a tendency towards suicide in adolescents. Taking into account the attributes proposed by [14] and comparing it to another research [16], they construct the dataset with variables that are related to suicide such as stress, self-esteem, depression among others. These variables are studied in risk factors for suicide in adolescents.

In this work, three algorithms have been evaluated in order to obtain the best model as a suicide tendency classifier: the JRIP rule-based algorithm, the C5.4 algorithm of the decision tree family, and the Naive Bayes algorithm based on probabilities. In the research [4] results using Naive Bayes, C5.5 and SVM are shown. It may be conclude that the combination of algorithms improve their performance.



Adolescent suicide trend tool

It comes from a broken home? No

Are parents carriers of mental illness? No

Do your family members have a history of suicida... 2

Do you have a history of learning disorders, ... No

Do you have a personal history of self-destructive... 2

Do you have obvious changes in habitual behavior? 3

Do you have friends with suicidal behavior? No

Do you have gestures, threats, suicidal plans? 3

Do you have a personal history of mental illness? 3

Are you currently in conflict (family, partner, school, etc.)? No

Level of suicide risk: Moderado

Fig. 5. Desktop aplication of tendency to suicide

In our work we present a dataset with 10000 instances and we propose a desktop application that allows to determine if an adolescent has a suicidal tendency or not by using the algorithm C5.4. As for [15], its data base is used with 880 real records of suicide attempts, in [6] the data is obtained from the smart phones, and a mobiles application is also proposed.

VI. Conclussions and Future Work

As a result we have the dataset generated and balanced instances which simulate the behavior of adolescents with a tendency towards suicide; this can be used to classify the adolescent suicidal tendency. To use the classifier model, the adolescent must provide interview-type information, where the questions are based on point assessments that the system will finally classify with the labels minor, moderate, serious and extreme. The families and clinics that use this tool will be able to prevent suicide.

In future work, it is necessary to construct a model taking into account other variables that lead suicide, this should be done by using psychological studies [8]. In addition suicide should be divided into suicide in children, suicide in young people and suicide in elders.

References

[1] Heath' world organization - OMS. "Suicide" http://www.who.int/mediacentre/factsheets/fs398/es/ consulted August 7th, 2017.

[2] National library of medine U:S "Suicide and suicidal behavior" https://medlineplus.gov/spanish/ency/article/001554.htm consulted August 7th, 2017.

[3] D'monte, S., & Panchal, D. (2015, May). Data mining approach for diagnose of anxiety disorder. In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 124-127). IEEE.

[4] Burnap, P., Colombo, W., & Scourfield, J. (2015, August). Machine classification and analysis of suicide-related communication on twitter. In Proceedings of the 26th ACM Conference on Hypertext & Social Media (pp. 75-84). ACM.

[5] Tran, T., Phung, D., Luo, W., Harvey, R., Berk, M., & Venkatesh, S. (2013, August). An integrated framework for suicide risk prediction. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1410-1418). ACM.

[6] Larsen, M. E., Cummins, N., Boonstra, T. W., O'Dea, B., Tighe, J., Nicholas, J., & Christensen, H. (2015, August). The use of technology in suicide prevention. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (pp. 7316-7319). IEEE.

[7] Alam, M. G. R., Cho, E. J., Huh, E. N., & Hong, C. S. (2014, January). Cloud based mental state monitoring system for suicide risk reconnaissance using wearable bio-sensors. In Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (p. 56). ACM.

[8] Health Ministery of Chile "National program of suicidal prevention" 2013, http://web.minsal.cl/sites/default/files/Programa_Nacional_Prevencion.pdf visited August 7th, 2017.

[9] Zhang, D., & Tsai, J. J. (Eds.). (2005). Machine learning applications in software engineering (Vol. 16). World Scientific.

[10] Reese, R. M. (2015). Natural language processing with Java. Packt Publishing Ltd.

[11] Sharma, S., & Bhagat, A. (2016, December). Data preprocessing algorithm for Web Structure Mining. In Eco-friendly Computing and Communication Systems (ICECCS), 2016 Fifth International Conference on (pp. 94-98). IEEE.

[12] Empresa Editora El Comercio, Suicidios, http://elcomercio.pe/lima/suicidios-3-6-adolescentes-intentaron-lima-callao-276648 consulted 07 de agosto de 2017.

[13] National Institute of mental health - Perú, Press realease,

http://exitosanoticias.pe/depresion-causa-el-80-de-suicidios-en-el-peru/ consulted August 7th, 2017

[14] Pérez, B,S. (1999), Suicide, behavior and preventions, Cuban magazine of General Comprehensive Medicine, v.15 n.2.

[15] J. Van-Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In ICML '07: Proceedings of the 21±th International Conference on Machine Learning, pages 935-942, New York, NY, USA, 2007. ACM Press