Khalid Saeed
Jiří Dvorský (Eds.)

LNCS 12133

Computer Information Systems and Industrial Management

19th International Conference, CISIM 2020
Bialystok, Poland, October 16–18, 2020
Proceedings

Springer

Lecture Notes in Computer Science 12133

More information about this series at http://www.springer.com/series/7409

Khalid Saeed · Jiří Dvorský (Eds.)

Computer Information Systems and Industrial Management

19th International Conference, CISIM 2020
Bialystok, Poland, October 16–18, 2020
Proceedings

Springer

Editors
Khalid Saeed
Bialystok University of Technology
Bialystok, Poland

Jiří Dvorský
VSB - Technical University of Ostrava
Ostrava, Czech Republic

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-47678-6 ISBN 978-3-030-47679-3 (eBook)
https://doi.org/10.1007/978-3-030-47679-3

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

CISIM 2020 was the 19th in a series of conferences dedicated to computer information systems and industrial management applications. The conference was held during October 16–18, 2020, in Poland at the Białystok University of Technology.

62 papers were submitted to CISIM 2020 by researchers and scientists from a number of reputed universities around the world. These scientific and academic institutions belong to Bulgaria, Chile, Colombia, Czech Republic, India, Italy, Pakistan, Poland, Spain, and Tunisia. Most of the papers were of high quality, but only 55 of them were sent for peer review. Each paper was assigned to at least two referees initially, and the accept decision was taken after receiving two positive reviews. In case of conflicting decisions, another expert's review was sought for the respective papers. In total, about 130 reviews and comments were collected from the referees for the submitted papers. In order to maintain the guidelines of Springer's *Lecture Notes in Computer Science* series, the number of accepted papers was limited. Furthermore, a number of electronic discussions were held within the Program Committee (PC) chairs to decide on papers with conflicting reviews and to reach a consensus. After the discussions, the PC chairs decided to accept for publication in this proceedings book the best 40 of the total submitted papers. The main topics covered by the chapters in this book are biometrics, security systems, multimedia, classification and clustering, and industrial management. Besides these, the reader will find interesting papers on computer information systems as applied to wireless networks, computer graphics, and intelligent systems. We are grateful to the three esteemed speakers for their keynote addresses. The authors of the keynote talks were Prof. Rituparna Chaki, University of Calcutta, India; Prof. Marina Gavrilova, University of Calgary, Canada; Prof. Witold Pedrycz, University of Alberta, Canada; Prof. Danuta Rutkowska, University of Social Sciences in Lodz, Poland; and Prof. Michał Woźniak, Wroclaw University of Technology, Poland.

We would like to thank all the members of the PC and the external reviewers for their dedicated efforts in the paper selection process. Special thanks are extended to the members of the Organizing Committee both the international and the local ones and the Springer team for their great efforts to make the conference a success. We are also grateful to Andrei Voronkov, whose EasyChair system eased the submission and selection process and greatly supported the compilation of the proceedings.

We hope that the reader's expectations will be met and that the participants enjoyed their stay in the beautiful city of Białystok.

October 2020

Khalid Saeed
Jiří Dvorský

Organization

Conference Patron

Lech Dzienis	Białystok University of Technology, Poland

General Chair

Khalid Saeed	Białystok University of Technology, Poland

Conference Co-chairs

Marek Krętowski	Białystok University of Technology, Poland
Rituparna Chaki	University of Calcutta, India
Agostino Cortesi	Ca' Foscari University of Venice, Italy

International Organizing Committee

Zenon Sosnowski (Chair)	Białystok University of Technology, Poland
Pavel Moravec	VŠB-Technical University of Ostrava, Czech Republic
Jiří Dvorský	VŠB-Technical University of Ostrava, Czech Republic
Nabendu Chaki	University of Calcutta, India
Sławomir Wierzchoń	Polish Academy of Sciences, Poland

Local Organizing Committee

Pavel Moravec (Co-chair)	VŠB-Technical University of Ostrava, Czech Republic
Maciej Szymkowski (Co-chair)	Białystok University of Technology, Poland
Mirosław Omieljanowicz	Białystok University of Technology, Poland
Grzegorz Rubin	Lomza State University, Poland
Mariusz Rybnik	University of Białystok, Poland
Aleksander Sawicki	Białystok University of Technology, Poland
Marek Tabędzki	Białystok University of Technology, Poland

Program Committee

Chair

Khalid Saeed	Białystok University of Technology, Poland

Members

Waleed Abdulla	The University of Auckland, New Zealand
Adrian Atanasiu	Bucharest University, Romania
Aditya Bagchi	Indian Statistical Institute, India
Valentina Emilia Balas	University of Arad, Romania
Anna Bartkowiak	Wrocław University, Poland
Rahma Boucetta	National Engineering School of Gabes, Tunisia
Nabendu Chaki	University of Calcutta, India
Rituparna Chaki	University of Calcutta, India
Agostino Cortesi	Ca' Foscari University of Venice, Italy
Dipankar Dasgupta	University of Memphis, USA
Pierpaolo Degano	University of Pisa, Italy
Riccardo Focardi	Ca' Foscari University of Venice, Italy
Marina Gavrilova	University of Calgary, Canada
Jan Devos	Ghent University, Belgium
Andrzej Dobrucki	Wrocław University of Technology, Poland
Jiří Dvorský	VŠB-Technical University of Ostrava, Czech Republic
David Dagan Feng	The University of Sydney, Australia
Pietro Ferrara	IBM T. J. Watson Research Center, USA
Raju Halder	Ca' Foscari University of Venice, Italy
Christopher Harris	State University of New York, USA
Kauru Hirota	Tokyo Institute of Technology, Japan
Khalide Jbilou	Université du Littoral Côte d'Opale, France
Ryszard Kozera	The University of Western Australia, Australia
Tomáš Kozubek	VŠB-Technical University of Ostrava, Czech Republic
Marek Lampart	VŠB-Technical University of Ostrava, Czech Republic
Christoph Lange	Fraunhofer IAIS, Germany
Jens Lehmann	University of Bonn, Germany
Flaminia Luccio	Ca' Foscari University of Venice, Italy
Pavel Moravec	VŠB-Technical University of Ostrava, Czech Republic
Romuald Mosdorf	Białystok University of Technology, Poland
Debajyoti Mukhopadhyay	Maharashtra Institute of Technology, India
Yuko Murayama	Iwate University, Japan
Nobuyuki Nishiuchi	Tokyo Metropolitan University, Japan
Andrzej Pacut	Warsaw University of Technology, Poland
Jerzy Pejaś	WPUT in Szczecin, Poland
Marco Pistoia	IBM T. J. Watson Research Center, USA
Piotr Porwik	University of Silesia, Poland
Jan Pries-Heje	IT University of Copenhagen, Denmark
S. P. Raja	Vel Tech Institution of Science and Technology, India
Isabel Ramos	University of Minho, Portugal
Anirban Sarkar	National Institute of Technology Durgapor, India
Ewa Skubalska-Rafajłowicz	Wrocław University of Technology, Poland
Kateřina Slaninová	VŠB-Technical University of Ostrava, Czech Republic
Krzysztof Ślot	Lodz University of Technology, Poland

Václav Snášel	VŠB-Technical University of Ostrava, Czech Republic
Zenon Sosnowski	Białystok University of Technology, Poland
Jarosław Stepaniuk	Białystok University of Technology, Poland
Marcin Szpyrka	AGH Kraków, Poland
Andrea Torsello	Ca' Foscari University of Venice, Italy
Qiang Wei	Tsinghua University, China
Sławomir Wierzchoń	Polish Academy of Sciences, Poland
Michał Woźniak	Wrocław University of Technology, Poland
Sławomir Zadrożny	Polish Academy of Sciences, Poland

Additional Reviewers

Marcin Adamski	Białystok University of Technology, Poland
Paola Patricia Ariza Colpas	Universidad de la Costa, Colombia
Giancarlo Bigi	University of Pisa, Italy
Rituparna Chaki	University of Calcutta, India
Tomasz Grzes	Białystok University of Technology, Poland
Wiktor Jakowluk	Białystok University of Technology, Poland
Miguel A. Jimenez-Barros	Universidad de la Costa, Colombia
Wojciech Kwedlo	Białystok University of Technology, Poland
Marek Lampart	VŠB-Technical University of Ostrava, Czech Republic
Tomáš Martinovič	VŠB-Technical University of Ostrava, Czech Republic
Ireneusz Mrozek	Białystok University of Technology, Poland
Dionicio Neira Rodado	Universidad de la Costa, Colombia
Sabina Nowak	University of Gdansk, Poland
Miroslaw Omieljanowicz	Białystok University of Technology, Poland
Hugo Hernandez Palma	Universidad del Atlántico, Colombia
Nadia Pisanti	University of Pisa, Italy
Grzegorz Rubin	Lomza State University, Poland
Mariusz Rybnik	University of Białystok, Poland
Soharab Hossain Shaikh	BML Munjal University, India
Maciej Szymkowski	Białystok University of Technology, Poland
Marek Tabędzki	Białystok University of Technology, Poland
Amelec Viloria Silva	Universidad de la Costa, Colombia
Lukáš Vojáček	VŠB-Technical University of Ostrava, Czech Republic

Keynotes

NSP and Its Application Towards Increasing Patient Satisfaction in Assisted Living

Rituparna Chaki

University of Calcutta, India
rituchaki@gmail.com

Abstract. The domain of nurse scheduling is a well-researched one. Researchers have been focusing mainly to increase the efficiency of nurses' allocation while maintaining the nurse satisfaction at an acceptable level. However, most of the existing works focus on optimizing the utilization of nursing staff. In the age of IoT, as more and more researches are carried on in the domain of assisted living, it becomes more important to use NSP for maximizing patient recovery. In our bid to understand the challenging issues of adapting the convention solutions in the assisted living scenario, a thorough state-of-the-art study has been done. This study led us to note the following issues that need to be addressed:

- Nurse scheduling solutions mainly aim at assigning the nurses so as to maximize the utilization of available nurses. The solutions mostly lacked consideration for the perspective of patients, viz, their preferences and types of ailments.
- Matching nurses' expertise to patients' requirements, as well as patients' preferences with respect to nurse availability is an issue that needs to be better investigated.
- The problem of increased size of search space regarding increasing number of patients' requirements need to be focused in more specific way.

The problem is defined as the assignment of a set of available nurses N to a set of patients P, depending on a number of criteria. The identification of parameters and the definition of constraints (soft and hard) is to be considered in such a way so as to maximize patient satisfaction. The cost function also needs to be formulated in terms of these parameters and the goal is to keep the cost at an optimum level. In this talk, the focus will be on discussing some of the relevant techniques used for solving the nurse scheduling problem, including a novel solution specifically aimed to increase patient satisfaction.

Adaptive and Reliable Decision Making for Multi-modal Biometric Systems

Marina Gavrilova

University of Calgary, Canada
marina@cpsc.ucalgary.ca

Abstract. The area of biometrics, without a doubt, has advanced to the forefront of an international effort to secure societies from both physical and cyber threats. This keynote provides an overview of the state of the art in multi-modal data fusion and biometric system design, linking those advancements with real-world applications.

The rapid development of massive databases and image processing techniques has led over the past decade to the significant advancements in both fundamental biometric research and in a relevant commercial product development. Typical biometric applications include banking, border control, law enforcement, medicine, e-commerce, smart sensors, and consumer electronics. A variety of issues related to biometric system performance and analysis has been addressed previously. A high number of biometric samples, data variability, data quality, data acquisition, types of fusion and system architectures have been shown to affect an individual biometric system's performance. Addition of new types of behavioral data, based on social interactions, presents unique challenges and opportunities. This keynote reviews current trends related to design of adaptive and reliable multi-modal biometric systems, with the focus on issues of security and privacy of person data. It supports the theoretical developments with the practical examples on the use of multi-modal biometrics in industrial applications, including city planning, finance, medicine, and situation awareness systems.

Explainable AI: From Data to Symbols and Information Granules

Witold Pedrycz

University of Alberta, Canada
pedrycz@ee.ualberta.ca

Abstract. With the progress and omnipresence of Artificial Intelligence (AI), two aspects of this discipline become more and more apparent. When tackling with some important societal underpinnings, especially those encountered in strategic areas, AI constructs call for higher explainability capabilities. Some of the recent advancements in AI fall under the umbrella of industrial developments (which are predominantly driven by numeric data). With the vast amounts of data, one needs to resort themself to engaging abstract entities in order to cope with complexity of the real-world problems and delivers transparency of the required solutions. All of those factors give rise to a recently pursued discipline of *explainable* AI (XAI). From the dawn of AI, symbols and ensuing symbolic process have assumed a central position and ways of symbol grounding become of interest. We advocate that in the realization of the two timely pursuits of XAI, information granules and Granular Computing (embracing fuzzy sets, rough sets, intervals, among others) play a significant role. The two profound features that facilitate explanation and interpretation are about an accommodation of the logic fabric of constructs and a selection of a suitable level of abstraction. They go hand-in-hand with the information granules. First, it is shown that information granularity is of paramount relevance in building linkages between real-world data and symbols encountered in AI processing. Second, we stress that a suitable level of abstraction (specificity of information granularity) becomes essential to support user-oriented framework of design and functioning AI artifacts. In both cases, central to all pursuits is a process of formation of information granules and their prudent characterization. We discuss a comprehensive approach to the development of information granules by means of the principle of justifiable granularity. Here various construction scenarios are discussed including those engaging conditioning and collaborative mechanisms incorporated in the design of information granules. The mechanisms of assessing the quality of granules are presented. In the sequel, we look at the generative and discriminative aspects of information granules supporting their further usage in the AI constructs. A symbolic manifestation of information granules is put forward and analyzed from the perspective of semantically sound descriptors of data and relationships among data. With this regard, selected aspects of stability and summarization of symbol-oriented information are discussed.

Artificial Intelligence and Image Understanding

Danuta Rutkowska

University of Social Sciences in Lodz, Poland
darutko@gmail.com

Abstract. Applications of Artificial Intelligence (AI) have increased rapidly in recent years. A lot of interest is focused on Deep Learning and Big Data. We are witnesses of spectacular results presented by Google with regard to Image Recognition. Deep Learning has also been successfully applied in speech processing and much more.

However, this part of AI – that is data driven – reflects the process of learning from examples. In this case, an intelligent system, e.g. a Deep Learning network, achieves a result, e.g. an image recognition, by the learning ability that increases along with larger amount of data examples. In this way, the system solves the problem without an explanation concerning the result.

Although with regard to Image Recognition, the explanation is not always necessary (we see what we get), in other AI applications a user wants to know how and why the system come up with the result. This is very important, e.g., in recommender systems and medical applications (also referring to medical images).

On another side of AI are expert systems – that are knowledge based – and realize an inference process, with explanation facilities. The knowledge is usually represented by logical rules. Therefore, it is possible to explain how a result has been obtained by use of the rules.

Hybrid intelligent systems can reflect both aspects of intelligence: an inference based on the knowledge represented by rules and the learning ability. This means that the inference can be realized – when the rules are known, otherwise the knowledge of the form of the rules can be acquired from data (examples) during the learning process.

In contrary to the Deep Learning approach – that is viewed as a "black box" because of the lack of the explanation – we propose a rule based system to solve an Image Understanding problem. The rules considered with regard to the knowledge of this system are fuzzy rules or the rules generated within the rough set theory. The linguistic description of images are produced by the system, and then analyzed within the framework of databases and AI. The goal is to describe an image based on the color segmentation, and location of particular color granules, as well as their size and shape. Mutual relationships between the color granules are taken into account in order to explain the understandable description of an image.

Chosen Challenges of Imbalanced Data Classification

Michał Woźniak

Wrocław University of Science and Technology, Poland
michal.wozniak@pwr.edu.pl

Abstract. Imbalanced data classification is still a focus of intense research because most of the learning methods can work with a reasonably balanced data set. Still, many real-world applications have to face imbalanced data sets. A data set is said to be imbalanced when several classes are under-represented (minority classes) in comparison with others (majority classes). Learning from imbalanced data is among the contemporary challenges in Machine Learning, and multi-class imbalance, as well as an imbalanced data stream, stand out as the most challenging scenarios.

In binary imbalanced learning, the relationships between classes are easily defined: one class is the majority one, while the other is the minority one. However, in multi-class scenarios, this is no longer obvious, as the correlations among classes may vary, e.g., one class can be at the same time minority and majority, or one of the different classes. Therefore canonical methods designed for binary cases cannot be directly applied in such scenarios.

Another topic which we will discuss during the talk is imbalanced data stream classification because only a few of the authors distinguish the differences between the imbalanced data stream classification problem and a scenario where the prior knowledge about the entire data set is given. This discrepancy is a result of the lack of knowledge about the class distribution, and this issue is notably present in the initial stages of data stream classification. Another difficulty is the presence of the phenomenon called *concept drift*, which can usually lead to classifier quality deterioration. The concept drift may have different nature, but it causes the change of the probability characteristics of the decision task, e.g., it could lead to a shift in the prior probabilities, i.e., the frequency at which the objects appear in the examined classes. A typical example of such a case is the technical diagnosis in which the fault probability increases with utilization time, and it may be a result of material fatigue. Sometimes the relationship between the minority and majority classes changes in a way that the former becomes the majority class.

This talk will discuss the main problems of imbalanced data classification, as multi-class imbalanced data analysis or imbalanced data stream classification, with particular attention to the methods developed by the Machine Learning team from the Department of Systems and Computer Networks from Wroclaw University of Science and Technology.

Contents

Industrial Management and other Applications

Machine Learning and High Performance Computing

Modelling and Optimization

A Machine Learning Approach for Severe Maternal Morbidity Prediction at Rafael Calvo Clinic in Cartagena-Colombia

Eugenia Arrieta Rodríguez[1(✉)], Fernando López-Martínez[2], and Juan Carlos Martínez Santos[3]

[1] Universidad del Sinú Cartagena Elias Bechara Zainum, Cartagena, Colombia
investigacionsistemas@unisinucartagena.edu.co
[2] Department of Computer Science, Oviedo University, Oviedo, Spain
felmco@gmail.com
[3] Universidad Tecnológica de Bolívar, Cartagena, Colombia
jcmartinezs@utb.edu.co
http://www.unisinucartagena.edu.co, http://www.uniovi.es/,
http://www.utb.edu.co

Abstract. There is a huge problem in public health around the world called severe maternal morbidity (SMM). It occurs during pregnancy, delivery, or puerperium. This condition establishes risk for babies and women lives since it's earlier detection isn't easy [8]. In order to respond to such a situation, the current study suggests the use of logistic regression, and supports vector machine to construct a predicting model of risk level of maternal morbidity during pregnancy. Patients for the current study was the pregnant women who received prenatal care at Rafael Calvo Clinic in Cartagena, Colombia and final attention in the same clinic. This study presents the results of two machine learning algorithms, logistic regression and support vector machine. We validated the datasets from the first, second and third quarter of pregnancy with both techniques. The study shows that logistic regression achieves the best results with the prenatal control dataset from the first and second quarter and the support vector machine algorithm achieves the best prediction results with the data set from the third quarter. We generated two datasets using the information of medical records on pregnancy patients at Maternidad Rafael Calvo Clinic. The first dataset contains the six initial months of pregnancy data and the second dataset contains the last quarter of pregnancy data. We trained the first model with logistic regression and the datasets corresponding to the first semester of pregnancy. We obtained a classification of 97% sensibility, 51.8% positive predictive value and F1 score of 67.7%. The support vector machine model was implemented with the datasets obtained from the third quarter of pregnancy. We obtained a classifier with 100% of sensibility, 27.0% of precision.

Keywords: Severe maternal morbidity · Machine learning · Logistic regression · Support Vector Machine

K. Saeed and J. Dvorský (Eds.): CISIM 2020, LNCS 12133, pp. 208–219, 2020.
https://doi.org/10.1007/978-3-030-47679-3_18

1 Introduction

Nowadays more people are using artificial intelligence techniques in different industry areas, especially in the medical field. The most common uses are prediction and classification of diseases, as in the case of this work, which attempts to apply supervised learning techniques to reduce morbidity rates in maternal complication scenarios, more commonly called Severe Maternal Morbidity (SMM) [3,8,11,15,17].

Maternal morbidity summarizes a set of complications that can have severe adverse effects on a woman's health, and that occurs during pregnancy, childbirth, or puerperium; putting at risk the lives of the women and their babies. When it appears, it is necessary to provide immediate attention to the patient to avoid death [14].

Despite advances in maternal health, complications related to pregnancy remain a significant public health problem in the world. Approximately 500,000 women die every year during pregnancy, delivery, or puerperium [6]. There are about 50 million problems in maternal health annually, and approximately 300 million women suffer in the short and long-term from diseases and injuries related to pregnancy, childbirth, and puerperium [14,19].

This condition is complicated to detect at an early stage. In response to the above, this article proposes the use of one of the techniques of machine learning that are considered more relevant in biomedical studies, such as Logistic Regression [7,9,13]. This technique performs a training and learning process to then predicts the level of risk for severe maternal morbidity in patients during pregnancy. The studied population corresponds to pregnant women who received prenatal care and final care at Clínica Rafael Calvo (CMRC) in Cartagena, Colombia.

Several studies worldwide for classification of diseases have shown excellent results when implementing logistic regression and support vector machine [2,10], being the logistic regression the most used and accepted regarding sensitivity and specificity. For this reason, in this study, logistic regression and support vector machine were implemented.

In a previous study called "Early Prediction of Severe Maternal Morbidity Using Machine Learning Techniques" [18], was presented the feature selection by classical statistical techniques. Fairly results with these techniques were obtained. Thus, for this study, we decided to test with a more significant sample data set along with machine learning feature selection techniques [5].

In the previous study, we applied an ANOVA analysis for features elimination. In this study, we used features selection of machine learning to reduce the number of features. Although the selected variables were not the same, the results obtained showed no improvement. We concluded that the problem was not the filtering techniques used, but the limited data set used for the training of the model.

For this work, logistic regression and support vector machine were considered the machine learning techniques [20], implemented with sklearn libraries (Python) [16].

2 Dataset Construction

For the dataset construction, a population of about 1300 patients who fulfilled the criteria of inclusion was selected, and it is shown in Table 1.

Table 1. Inclusion criteria

Criteria	Description
Age	13 to 45 years-old
Prenatal visits	Yes
Controls number	2 or more

Dataset construction was performed with the same inclusions criteria of the previous study as well as the variables collected by each patient.

A sample population of 479 patients was obtained, a retrospective analysis was performed to data contained in the clinical histories of pregnant patients, where between 6,000 to 8,000 cesarean sections are attended per year.

This list of variables collected for the study was provided by the area of surveillance in Public Health and Obstetrics and Gynecology of the research center at CMRC.

Two data sets were created with the collected information. The first data set corresponds to the prenatal controls of the first and second quarter. The second data set corresponds to the prenatal controls of the third quarter. It was necessary to include filtering techniques to the variables due to the enormous amount of cases. This type of technique is being used to reduce the number of variables in the data set to improve the outcome and performance of the classifier. The function RFECV (Recursive Features Elimination with Cross Validation) of the python sklearn library was used for variable elimination *Sklearn* [16].

3 Logistic Regression

Logistic regression notwithstanding its name is a linear classification model instead of a regression model [12].

Logistic regression is a widely linear classification algorithm used in medicine where a sigmoidal function is coupled with a linear regression model [22].

In this work, the logistic regression was applied with cross validation and a regularization component was introduced to control the complexity of the model. In order to evaluate the results of the statistical analysis, it uses the cross validation technique to ensure that the results of each layer or segment are independent between the test data and the training data, consists of analyzing the data in a subset called “training set” and validating the analysis in the other subset called “test set”.

In order to solve this optimization problem, the cost function is minimized and an adjustment factor is added that allows controlling the complexity of the model. The regularization consists in reducing the importance of the θ parameters being modified the function of costs by the addition of the sum of all the θ parameters with a factor called parameter of regularization, λ. Getting Eq. 1 as a result [1].

$$min \ -\frac{1}{N}\sum_{i=1}^{N}[y_n \log h_\theta(x) + (1-y_n)\log(1-h_\theta(x))] + \lambda||\theta||^2 \quad (1)$$

Knowing that N is the number of variables, θ are the parameters of each variable, y is the vector of response (only manages binary values (0,1)), and λ is the parameter of regularization.

3.1 Variable Filtering

we used a recursive feature elimination and cross-validated selection to select best number of features, using 80% of data to train and 20%. Dataset was corresponding to the first and second trimester data of pregnancy were used, called on SUBSET1. The outcome by using this method was that 11 of 60 variables were recommended as predictors or parameters for the classifier. The output of the algorithm is a matrix with false and true values that indicates the 11 selected variables. Table 2 shows the selected variables.

Table 2. Features filtered of 1^{st} y 2^{nd} trimester

Type	Feature
Personal information	Age
Ethnicity	Palequero
Health care regulation	Contributory Regime (CR)
Antecedent	Preeclampsia
Antecedent	Eclampsia
Antecedent	Diabetes
Antecedent	Urinary tract infection
Diagnosis	E20–E35: Disorders of other endocrine glands
Diagnosis	O20–O29: Respiratory tract infection
Diagnosis	N30–N39: Diseases that can affect the fetus
Diagnosis	Z30–Z39: Medical care for reproduction

The same algorithm was applied to the another dataset (third trimester), and this will be called SUBSET2. In this case, the algorithm selected 5 variables as predictors of 59. Which can be seen in Table 3.

Table 3. Features 3^{rd} trimester

Type	Feature
Personal information	Multiparity
Antecedent	Urinary tract infection
Diagnosis	O10–O16: Edema proteinuria and hypertension
Diagnosis	O30–O48: Complications of pregnancy that require attention to the mother
Diagnosis	O60–O75: Complications of pregnancy and delivery

3.2 Training and Validation

An analysis was necessary to be performed to validate the effectiveness of the created data set and verify if this dataset provides a good predictor to predict severe maternal morbidity in its early stages. A cross-validation analysis of the behavior of SUBSET1 and SUBSET2 was performed with logistic regression, and both showed 63% of accuracy. Based in this result we decided to use only SUBSET1 and the third trimester data will not longer be used, only will be use the 223 examples of data set of first and second trimester. In this work was implemented 5-fold cross-validation like see in Fig. 1.

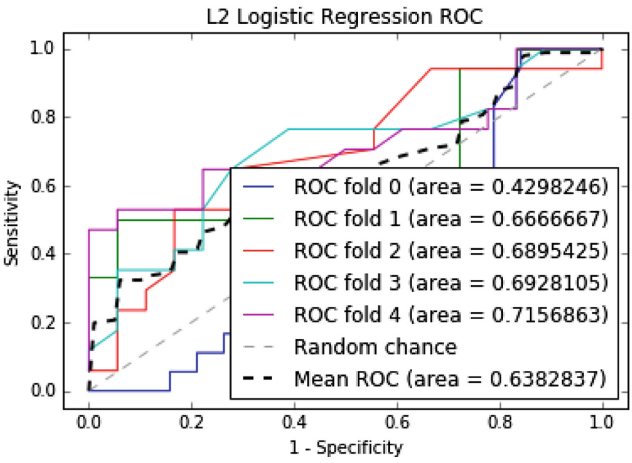

Fig. 1. ROC curve

The dotted line shows the average of the AUC for all folds. The prediction model should detect a mayor number of patients with SMM risk, and for this

reason, the selected metrics were sensitivity and precision. DATASET1 contain 223 examples and data analysis was performed only with the 80% of this data with a balanced distribution of the positive and negative class, and 11 variables were used as predictor variables for the classifier (View Table 3).

The result of the prediction can be seen in the confusion matrix shown in Fig. 2. When total of $TP = 85$ (True Positive), this means that 85 patients were classified as SMM and is true that they have SMM. $TN = 12$ (True Negatives), which indicates that 12 patients were classified as no with SMM condition and they were not with SMM condition. $FN = 2$ (False Negatives), this indicates classification errors of patients with SMM condition. Finally, $FP = 79$ (False Positive), patients that were classified with SMM but did not have the condition.

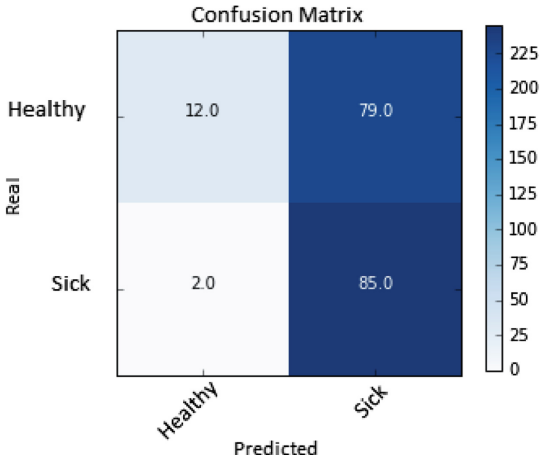

Fig. 2. Confusion matrix 1^{st} y 2^{nd} quarter

The confusion matrix results allow calculate the Precision, Recall and F1 Score. As seen in the Table 4.

Table 4. Results

Metrics	Result
Precision	51.8%
Recall	97.7%
F1 score	67.7%

Precision is defined as the proportion of positive predictions made correctly by the model and denoted by Eq. 2.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall or Sensitivity that is the ability of the model to predict positive cases that are really sick, like see in Eq. 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 score is a harmonic measure between recall and precision. It is a weighted average between these two measures as indicated in Eq. 4.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Another way to evaluate the classifier is by calculating the ROC curve as shown in Fig. 1 [4,21]. This is a graphical representation of the sensitivity vs. the specificity for a binary classifier. This can also be expressed as the reason between TPR (true positive rate) and FPR (false positive rate) which is equal to 1 – specificity. Comparing the result of the ROC curve of this work that is 63% and the previous one corresponds to 66% [18]. It is evident that there was no improvement in the terms of the ROC curve, which indicates that the problem is not related to the selection technique of variables. The reason is the small number of prenatal exams or examples that are counted in the first and second trimesters of pregnancy.

4 Support Vector Machine Algorithm

Support Vector Machines (SVM) are one of the methods of supervised learning for two types of classification problems. SVMs work with linearly non-separable problems, and seek to separate the data with a large gap or hyperplane. For the case of problems that are not linearly separable, it is recommended the inclusion of core functions or Kernel which has the effect of mapping the inputs to a high-dimensional space, where the data will be linearly separable. The core function objective is to separate the support vectors from the rest of the training data, this is a quadratic programming problem (QP). See Eqs. 5 and 6.

$$minw, \xi \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \tag{5}$$

subject to

$$y(\boldsymbol{x_i} \cdot w + b) \geq 1 - \xi \tag{6}$$

In this study the core function RBF (Radial Base Function) is used. It is also known as the "exponential" core. The RBF core functions take the form $\exp(-\gamma|x - x'|^2) \cdot \gamma$. Where γ eis a constant of proportionality whose range of useful values must be estimated for each particular application.

4.1 Variable Filtering

We used the same algorithm RFECV, but with linear support vector machine as classifier and the second dataset SUBSET1. As result was eliminated 45 features and finally it is obtain 15 features, as shown Table 5.

Table 5. Features for third and second quarter con technique SVM

Classification	Feature
Personal date	Age
Ethnicity	Raizal
Socio economic	Strata
Marital status	Consensual union couples
Marital status	Widowhood
Gynecological data	Parity
Multiple pregnancy	Multiple
History of	Preeclampsia
History of	Eclampsia
History of	Urinary Tract Infection
ICD-10	I11–I15: Hypertensive diseases
ICD-10	J00–J06: Acute upper respiratory infections
ICD-10	N30–N39: Other diseases of urinary system
ICD-10	O10–O16: Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
ICD-10	O20–O29: Other maternal disorders predominantly related to pregnancy

Followed by this the feature, the selection algorithm was applied again, but taking 80% of the data set of the SUBSET2. For a total of 59 variables, the algorithm eliminates 45 variables and only 13 are recommended, as shown in Table 6.

4.2 Training and Validation

The result of the SVM classifier is shown in Fig. 3 which is the confusion matrix. In this one, we can see a total of $VP = 95$ (True Positives), this means that 95 patients who really had it were classified with risk for SMM. $VN = 8\ 8$ (True Negatives), which indicates that 8 patients were classified as healthy or without risk for SMM and that in reality they did not have SMM. On the other hand, $FN = 0$ (False Negatives) which means that no patient who ended up in SMM was classified without risk for SMM. Finally, false positives $FP = 245$ False Positive), this indicates that 245 patients without risk were classified with risk for SMM.

Table 6. SVM third quarter predictor

Classification	Feature
Ethnicity	Black Race
Insurance	Commercial
Insurance	Self-pay
Insurance	Government
Location	Urban
Location	Rural
Marital status	Preeclampsia
History of	Eclampsia
History of	Urinary Tract Infection
ICD-10	E00–E07: Disorders of Thyroid Gland
ICD-10	O10–O16: Edema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
ICD-10	O60–O75: Complication of labor and delivery, unspecified
ICD-10	Z30–Z39: Persons encountering health services in circumstances related to reproduction

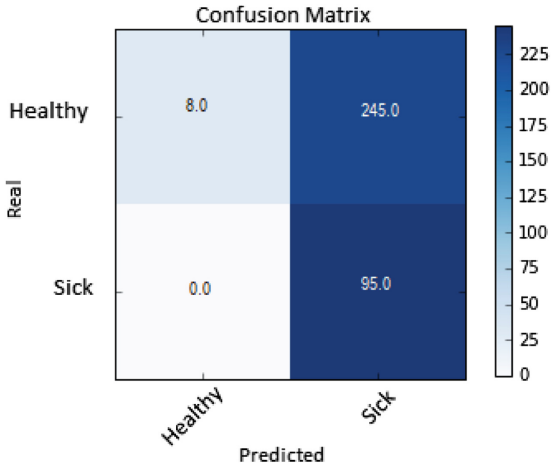

Fig. 3. SVM third quarter confusion matrix

Based on the confusion matrix, the measures that are of interest to evaluate the performance of the SMM classifier, which are the Precision, Recall and measure F1, are calculated. Table 7 shows the results of each one.

The result of 100% sensitivity or recall indicates that all patients with potential SMM risk will be detected by the classifier. A precision of 27% which indicates that there will be many patients who may not develop SMM but the

Table 7. Performance metrics of support vector machine

Metric	Result
Precision	27.9%
Recall	100%
F1	43.6%

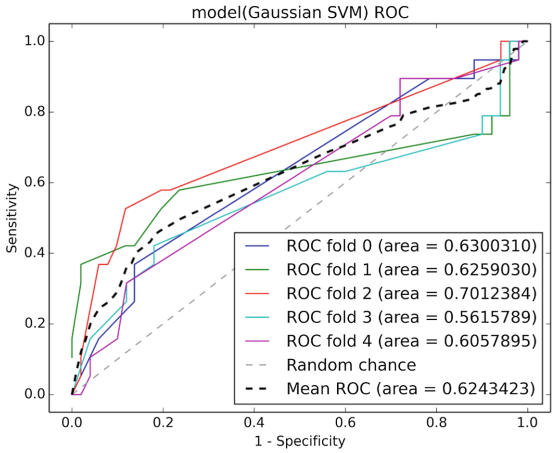

Fig. 4. ROC graph using SVM for the third quarter

predictor will tend to classify them as sick. In the same way as the ROC curve is used in the logistic regression to evaluate the performance, it is also used for support vector machine, the ROC graphic represents sensitivity vs specificity, as shown in the Fig. 4.

5 Conclusion and Future Work

In this research work, we present the use of features selection, logistic regression, and support vector machine applied to the data set obtained from the prenatal controls of the CMRC.

Cross validation and test error were applied to determine the supervised learning technique best suited to the early prediction problem of severe maternal morbidity. The logistics regression technique was selected for the first and second quarter dataset. The result was a classifier with 97% recall, 51.8% precision and 67.7% measurement F. For the third quarter dataset, support vector machine were selected with results of 100% recall and 27% precision.

In the previous study we made an analysis of variance (ANOVA) while in the current one we use recursive features elimination with cross-validation (RFECV) technique. The results in the first work are similar than second, 65% and 63% in ROC graphic respectively. Whereby it is concluded that filtering technique

not improvement the performance of the model, but this may be being affected by the no quality of data, then, required apply other techniques of datamining and data cleaning. I Nevertheless, in current study we show logistic regression technique presents more realistic results than the SVM technique in terms of precision and recall.

Recommendations for future work is aimed at improving the collection of information by CMRC, guiding clinical histories and the prevention of variables recommended by the bibliography.

As a result of observation of the data, it is identified that there are few data in the first and second quarters of pregnancy, which shows a culture of not attending prenatal check-ups, so it is recommended that health institutions implement the campaigns to obtain information about the early stages of pregnancy that facilitate the use of artificial intelligence techniques for the early detection of SMM. Finally, this work builds a baseline for future projects in SMM prediction using machine learning techniques.

References

1. Caicedo-Torres, W., Paternina, Á., Pinzón, H.: Machine learning models for early dengue severity prediction. In: Montes-y-Gómez, M., Escalante, H.J., Segura, A., Murillo, J.D. (eds.) IBERAMIA 2016. LNCS (LNAI), vol. 10022, pp. 247–258. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47955-2_21
2. Calvert, J.S., et al.: A computational approach to early sepsis detection. Comput. Biol. Med. **74**, 69–73 (2016)
3. Farran, B., Channanath, A.M., Behbehani, K., Thanaraj, T.A.: Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. BMJ Open **3**(5), e002457 (2013)
4. Fawcett, T.: ROC graphs: notes and practical considerations for researchers. Mach. Learn. **31**(1), 1–38 (2004)
5. Feizi-Derakhshi, M.R., Ghaemi, M.: Classifying different feature selection algorithms based on the search strategies. In: International Conference on Machine Learning, Electrical and Mechanical Engineering (2014)
6. Haaga, J.G., Wasserheit, J.N., Tsui, A.O., et al.: Reproductive Health in Developing Countries: Expanding Dimensions, Building Solutions. National Academies Press, Washington, D.C. (1997)
7. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
8. Jahan, S., Begum, K., Shaheen, N., Khandokar, M.: Near-miss/severe acute maternal morbidity (SAMM): a new concept in maternal care. J. Bangladesh Coll. Phys. Surg. **24**(1), 29–33 (2006)
9. Lorduy Gómez, J., Carrillo González, S., Muñoz Baldiris, R.E., Díaz-Pérez, A., Perez, I.: Prognostic factors of early neonatal sepsis in the city of Cartagena Colombia (2018)
10. Mani, S., et al.: Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J. Am. Med. Inform. Assoc. **21**(2), 326–336 (2014)
11. Nanda, S., Savvidou, M., Syngelaki, A., Akolekar, R., Nicolaides, K.H.: Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. Prenat. Diagn. **31**(2), 135–141 (2011)

12. Ng, A.: Machine learning: Stanford machine learning course materials
13. Nilashi, M., bin Ibrahim, O., Ahmadi, H., Shahmoradi, L.: An analytical method for diseases prediction using machine learning techniques. Comput. Chem. Eng. **106**, 212–223 (2017)
14. World Health Organization, UNICEF: Revised 1990 estimates of maternal mortality: a new approach. World Health Organization (1996)
15. Park, F.J., Leung, C.H., Poon, L.C., Williams, P.F., Rothwell, S.J., Hyett, J.A.: Clinical evaluation of a first trimester algorithm predicting the risk of hypertensive disease of pregnancy. Aust. N. Z. J. Obstet. Gynaecol. **53**(6), 532–539 (2013)
16. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
17. Poon, L.C., Kametas, N.A., Maiz, N., Akolekar, R., Nicolaides, K.H.: First-trimester prediction of hypertensive disorders in pregnancy. Hypertension **53**(5), 812–818 (2009)
18. Rodríguez, E.A., Estrada, F.E., Torres, W.C., Santos, J.C.M.: Early prediction of severe maternal morbidity using machine learning techniques. In: Montes-y-Gómez, M., Escalante, H.J., Segura, A., Murillo, J.D. (eds.) IBERAMIA 2016. LNCS (LNAI), vol. 10022, pp. 259–270. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47955-2_22
19. Tsui, A.O., Wasserheit, J.N., Haaga, J.G., et al.: Healthy pregnancy and childbearing (1997)
20. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)
21. Yang, Z., Zhang, T., Lu, J., Zhang, D., Kalui, D.: Optimizing area under the ROC curve via extreme learning machines. Knowl.-Based Syst. **130**, 74–89 (2017)
22. Zheng, Z., Li, Y., Cai, Y.: The logistic regression analysis on risk factors of hypertension among peasants in east china & its results validating. Int. J. Comput. Sci. Issues (IJCSI) **10**(2 Part 1), 416 (2013)