Machine Learning Approach for Pre-Eclampsia Risk Factors Association

Antonieta Martínez-Velasco†
Faculty of Engineering
Universidad Panamericana Campus México
Ciudad de México, México
amartinezv@up.edu.mx

Lourdes Martínez-Villaseñor
Faculty of Engineering
Universidad Panamericana Campus México
Ciudad de México, México
lmartine@up.edu.mx

Luis Miralles-Pechuán
Centre for Applied Data Analytics Research (CeADAR)
University College Dublin
Belfield, Dublin, 4 Ireland
luis.miralles@ucd.ie

ABSTRACT

The preeclampsia/eclampsia syndrome is a multisystem disorder that usually includes cardiovascular changes, hematologic abnormalities, hepatic and renal impairment, and neurologic or cerebral manifestations. Preeclampsia (PE) is a clinical syndrome that afflicts 3–5% of pregnancies and it is a leading cause of maternal mortality, especially in developing countries. To understand in greater depth the preeclampsia/eclampsia syndrome, we applied some well-known Machine Learning (ML) techniques. ML has been successfully applied to medical research to improve the diagnosis and the prevention of complex diseases and syndromes. In our contribution, we have created a supervised model to predict if a patient suffers the disease. This model has been optimized by selecting the best features and by optimizing the threshold when predicting a class. We used these techniques to point out the most related features of the patients to the disease. Finally, we used interpretability techniques to extract and visualize through a decision tree the most relevant associations of the disease with the patients' features.

CCS CONCEPTS

• Applied computing • Life and medical sciences • Healthcare information systems.

KEYWORDS

Preeclampsia, Risk Factors, Genetic Variants, Machine Learning.

ACM Reference format:

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11…$15.00
https://doi.org/10.1145/3284869.3284912

1 Introduction

Pregnancy is associated with a group of physiologic and pathologic changes. One of the most important pathologies accompanying pregnancy is the preeclampsia/eclampsia syndrome. The syndrome is a multisystem disorder that can include cardiovascular changes, hematologic abnormalities, hepatic and renal impairment, and neurologic or cerebral manifestations [1]. Preeclampsia (PE) is a clinical syndrome that afflicts 3–5% of pregnancies and is a leading cause of maternal mortality, especially in developing countries. It is a multisystem hypertensive disorder [2].

PE diagnosis can be improved with the use of e-health methods. The current focus of health care researchers is to promote the use of e-health technology in developing countries to support medical decisions [3]. The low Socio-economic factors act as multiple risk factors for pre-eclampsia. In Mexico, the low socio-economic status of women doubled the risk of pre-eclampsia and eclampsia [4]. A study in Australia found working women compared to non-working ones had a higher risk of developing pre-eclampsia and eclampsia [5]. This may be linked to the stress that women get during work. ML techniques have been used to support the health expert in the prevention of preeclampsia [11-21].

Nevertheless, some decision support systems use so-called “black-box” ML techniques [6]. If the health expert does not understand how the system makes decisions, he/she usually stop using the system. Trust is vital for the system adoption. Rule-based ML techniques are commonly used to be more interpretable. Interpretability, according to [7], is “the degree to which a human can understand the cause of decision”.

This paper seeks to analyze the socioeconomic and health factors that represent a risk factor for pregnant women to suffer preeclampsia, and present some ways in which the decisions made by the system are understandable to specialists. To gain the trust of health experts, we used interpretable techniques to extract and visualize through a decision tree the most relevant associations of the disease with the patients' features. Our study shows a significant association between maternal education, income, and pre-eclampsia. We propose to analyze the dataset obtained from Public Private Ventures. Evaluation of Children’s Futures: Improving Health and Development Outcomes for Children in Trenton, New Jersey, 2001-2005 to develop a procedure to study other data from different ethnic groups to do extensive these methods.

The rest of the paper is as follows. The definition of preeclampsia syndrome and the risk factors associated with the disease is presented in section 2. A brief state of the art of the analysis of some approaches to study preeclampsia is presented in section 3. In section 4, we proposed a ML-based diagnosis and prediction for Preeclampsia. In section 5, experiments and results are shown and discussed. Section 6 concludes the paper and highlights future work in this context.

2 Preeclampsia

Preeclampsia (PE) is defined as the new onset of hypertension and proteinuria during the second half of pregnancy that typically appears around 20 weeks of gestation with symptoms of hypertension and proteinuria [7]. A few clinical circumstances increase the risk of preeclampsia: type I diabetes mellitus or type II diabetes mellitus, obesity, systemic lupus erythematosus, advanced maternal age (older than 40 years). The increased prevalence of chronic hypertension and other comorbid medical illnesses in women older than 35 years may explain the increased frequency of preeclampsia among older women [8]. Major efforts have been directed at the identification of demographic factors, biochemical analyses, or biophysical findings, alone or in combination, to predict early in pregnancy the later development of preeclampsia [2]. The utility of a predictive test will depend on the overall prevalence of the disease. It is necessary to consider that the prediction in each case is important, for which reason a methodology should be devised that maximizes the number of cases predicted adequately [9]. Results from mechanistic studies have provided insights into the pathogenesis of the disease; also have created opportunities to study circulating and urinary biomarkers to predict the disease [10].

3 Machine Learning Based Diagnosis and Prediction for Preeclampsia

ML techniques have been applied in some approaches that include metabolites, images analyses, and risk factors datasets, among others to diagnose and to predict PE. In Table 1, we present a list of some relevant works that studied PE under different approaches. It includes the risk factors determination approach and the molecular biology approach. All of them using ML methods.

Kenny [11] proposed a tree-based genetic programming to diagnose PE analyzing three metabolites in blood plasma. The authors consider metabolites in plasma obtained from 87 cases and controls. They present some demographic data and analyzes the concentration of some metabolites in both groups. They present results in a chromatogram and some rules to obtain conclusions. Neocleous [12] proposed Neural Networks to classify a database containing 15 risk factors with the best results obtained were with a multi-slab neural structure to estimates of the risk of occurrence of PE at an early stage. This study considers the medical history as well as the consumption of drugs or alcohol in pregnant women. Graphs about the performance of the neural networks used are presented. The results are reported in tables and conclusions by means of texts. Thus, Espinilla [13] classified a risk factors dataset using decision trees without pruning and linguistic fuzzy transformation using a genetic approach. This work presents a methodology, which allows a linguistic monitoring in real time. They presented some rules generated by the decision tree. Velikova [14] proposed a database classification by means Bayesian Networks model that is manually built using expert knowledge. Relevant input data – risk factors and measurements of signs are provided in the mobile app. The dataset contains medical history and some external data. The author presents user interfaces showing the risk to suffer the disease based on the temporal Bayesian network. They do not present any elements to support interpretability for medical experts. Tejera [15] classified a clinical history dataset including maternal heart rate variability (HRV) indexes and risk factors to characterize PE. Results are presented in terms of ROC curves, sensitivity a specificity variations and Normalized importance of the independent variables in the obtained Artificial Neural Networks. There are not elements to improve the results interpretability. Villa [16] proposed Bayesian clustering to compute the risk ratio of each disease outcome. The analysis indicates an exponential increase in the risk of preeclampsia as the number of risk factors increased. The analysis is based in case and controls medical records. This work displays the results of the cluster analysis in the heat-map that presents the risk factors in the different clusters. The way to take the decisions by the system is not explicit. Moreira [17] proposed the classification of risk factors, physiological mechanisms, and symptoms dataset to identify high-risk pregnancy. The main contribution of this work includes the presentation of a Bayesian network built to help decision makers in moments of uncertainty in the care of pregnant women. This work focused on the construction of a smart system designed to support a medical decision for pregnant healthcare. Fergus [18] classified a genetic variants dataset based on Genomic Wide Association Study (GWAS) using Deep learning stacked autoencoders to allow early detection of preeclampsia. They proposed to include in future works use structured logic rules to reduce un-interpretability of neural networking models. Cox [19] proposed Bayes Net as the better algorithm to classify a plasma membrane proteins dataset. They present exhaustively the procedures and datasets used. Mehta [20] presented a survey and analysis of data mining methods applied to maternal care. In terms of interpretability, the author concludes that graphical representation of Decision Trees and Naïve Bayes models are easier to understand, unlike Neural Networks and Vector Machines.

In summary, we can say that although some related works use interpretable techniques, the authors do not care to present results so that the user understand how the decisions were made.

4 Machine Learning Based Diagnosis and Prediction for Preeclampsia

In this section, we describe the use of ML methods to discover relations between medical and socioeconomic variables and PE in women living in New Jersey. In the same way, we present interpretable techniques to present results so that the health- expert understand how the decisions were made. We proposed to develop a ML model to diagnose PE in order to support the prognosis made by the health expert.

Table 1. Preeclampsia studies by means of Machine Learning methods.

	DATA	TASK	METHODS
Kenny[11], 2005	Data matrix obtained from a 3D plot. It contains three metabolites peak in each sample.	Classification	Tree-based Genetic Programming.
Neocleous, 2009 [12]	The database includes 15 parameters.	Classification	Neural Networks with a multi-slab neural structure.
Espinilla, 2017 [13]	Dataset containing risk factors.	Classification.	Decision trees. Linguistic Fuzzy transformation.
Mackenzie, 2016 [21]	Dataset of Exosomes extracted from plasma.	The contribution of different tissues to gene expression.	Deconvolution (Negative Matrix Factorization).
Cox, 2011 [19]	Plasma membrane proteins.	Classification.	Multiple ML algorithms were used.
Velikova, 2013 [14]	Risk factors, signs and clinical history.	Classification.	Bayesian Networks.
Tejera, 2011 [15]	Clinical history dataset.	Classification.	Artificial Neural Network
Villa, 2017 [16]	Risk factors database.	Clustering.	Bayesian clustering.
Moreira, 2016 [17]	Risk factors, physiological mechanisms, symptoms.	Classification.	Bayesian Networks.
Fergus, 2018 [18]	SNP database	Classification.	Deep learning stacked autoencoders.
Mehta, 2016 [20]	Survey and analysis of data mining methods applied to Maternal care domain.		

The aim to use ML techniques for these study case is to detect patterns and relations than a human, or traditional statistic cannot. ML methods are able to learn automatically without explicit programming, effectively, and with less cost and effort [22]. Models are evaluated by metrics commonly used by the scientific community such as accuracy, the root mean square error (RMSE), sensitivity or specificity [23]. These metrics allow to evaluate model precision and to compare the results with the research community.

4.1 Dataset

The dataset was obtained from the public study Evaluation of Children's Futures: Improving Health and Development Outcomes for Children in Trenton, New Jersey, 2001-2005[24]. These data were collected for the initial phase of the evaluation of the Children's Futures initiative, a comprehensive set of interventions aimed at improving child health and development outcomes from prenatal to age three in Trenton, New Jersey. The dataset includes 1634 records and 25 features

4.2 Experimentation and Results

In the present work, we predict if the patient will suffer the disease. We use supervised ML classification methods to determine the probability of this happening. We applied the following methodology: First, we preprocessed de dataset that had missing values and is imbalanced. Second, we faced the imbalance applying leave one out cross validation and analyzing de ROC graph. Third, we ordered features according to their importance to suffer the PE syndrome.

We processed a dataset, which includes 1634 records, and 25 features. The database is imbalanced; there were 269 PE cases and 1365 healthy subjects. The dataset was examined to identify the empty values. In the columns with categorical values, we have replaced the NaN values by the mode. In those that have numerical values, we have replaced the missing values by the average of the columns.

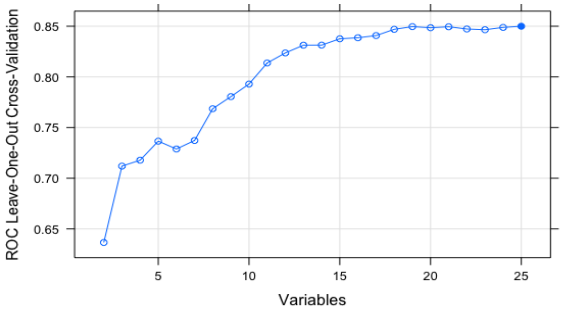

Figure 1: ROC curve for preeclampsia dataset

Once the database was preprocessed, we made feature selection by means a function, which performs a leave one out cross validation (loocv) experiment of a learning system on a given data set. The function is completely generic. The generality comes from the fact that the function that the user provides as the system to evaluate, needs in effect to be a user-defined function that takes care of the learning, testing, and calculation of the statistics that the user wants

to estimate through loocv [25]. Then, for the variables in the selection, we used a methodology based on the Leave One Out with cross-validation of 5 repetitions and 10 parts. We faced this problem using Receiver Operating Characteristic curve (ROC) for Leave One Cross Validation method as is shown in Fig.1 (The ROC curve penalizes such condition by limiting the maximum possible value along the axis). We build ML models with all the variables and select the best ROC (Table 2) obtained from the Random Forest model.

Table 2. Models constructed with 25 features

Method name	ROC	Acc	AvgAcc	Sens	Spec	Prec	F1	Time(Sec)
Random Forest	0.8499	0.8530	0.7730	0.6846	0.8614	0.1985	0.3078	208.572
AdaBoost Classification Trees	0.8252	0.8471	0.7184	0.5524	0.8843	0.3762	0.4476	2561.04
Stochastic Gradient Boosting	0.8244	0.8543	0.7540	0.6377	0.8704	0.2669	0.3763	27.294
Glmnet	0.8186	0.8521	0.7782	0.6974	0.8590	0.1799	0.2860	13.462
MARSplines	0.8161	0.8513	0.7350	0.5956	0.8745	0.3011	0.4000	7.554
Linear Discriminant Analysis	0.8106	0.8513	0.7294	0.5742	0.8846	0.3740	0.4530	2.577
Bayesian GLM	0.8092	0.8548	0.7479	0.6203	0.8755	0.3048	0.4088	3.132
NN with Feature Extraction	0.8038	0.8586	0.7455	0.5965	0.8945	0.4364	0.5040	61.636
SVM Radial Basis Funct Kernel	0.8009	0.8479	0.7298	0.5934	0.8661	0.2409	0.3427	38.424
SVM with Linear Kernel	0.7891	0.8460	0.7192	0.5694	0.8690	0.2654	0.3620	13.658
k-Nearest Neighbors	0.7875	0.8379	0.6922	0.5142	0.8701	0.2825	0.3647	44.644
Single C5.0 Tree	0.7731	0.8392	0.6978	0.5176	0.8781	0.3398	0.4103	5.471
Boosted Logistic Regression	0.7658	0.8397	0.6979	0.5222	0.8735	0.3056	0.3856	5.937
C4.5-like Trees	0.7492	0.8430	0.7090	0.5304	0.8876	0.4022	0.4575	38.509

The model shows that most samples are predicted with values close to 1 as can be seen in a histogram in Figure 2. The histogram shows that above 0.85 the individual is healthy.

4.3 Interpretability

An important challenge to overcome in the studies made with ML is that health professionals do not trust them because it is not clear to them how they were obtained. The higher the interpretability of a model, the easier it is for someone to comprehend why certain decisions were made.

To improve the interpretability, we present in a graph PE diagnosis attributes ordered by their importance; a decision tree using 25 variables according to the Random Forest method. MDGI metric measures the proportion of incorrectly classified samples when the evaluated feature is removed from the dataset used to build the model. The MDGI value of each variable is scaled in the range [0, 100]. The most important variables are Duration of Completed Pregnancy in Weeks (PRGLNGTH), Poverty, Water retention/edema in pregnancy (SWLNANKL), Toxemia, Education (Completed Years of Schooling) (EDUCAT), Highest Completed Year School or Degree (HIEDUC), Pregnancy Outcome (OUTCOME), Labor Force Status (LABORFOR).

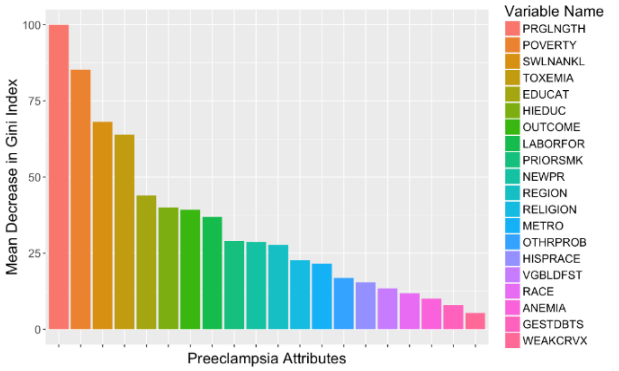

Figure 2: Variable Importance from Random Forest

We present in Figure 3 a decision tree based on three variables of the dataset. Nodes are labeled with unique numbers. The root node is 1. The tree diagram shows in color the node numbers for the tree. Only the terminal node numbers are displayed. The number in the first row of each node indicates the group: "2" represents that the patient suffering the disease. The two numbers in the second row of each node indicate the probability to be affected by the variable. Finally, the number in the third row represents the sample percentage covered by the node. Therefore, in the decision tree, we see that the feature "Water retention/edema in pregnancy" (SWLNANKL) divides the dataset into two sets (16% to present water retention). On a second level, TOXEMIA is used to make the second classification. Node 3 represents the individuals that do not have PE. Elements from Node 10 have a pregnancy length lesser to 36 weeks and greater or equal to 18 weeks. Individuals in Node 20 represents the woman that works in this period (Labor Force Status). If women have greater or equal to 6.5 she has the 58% of like hood. In the last level, is represented by the like hood to have the conditions represented in predecessors Nodes.

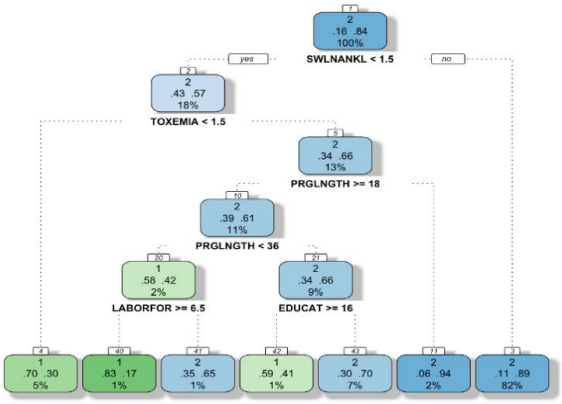

Figure 3: Decision Tree based on three variables of the dataset

In brief, feature importance provides a highly compressed, global insight into the model's behavior. The tree structure is perfectly suited to cover interactions between features in the data and the interpretation is simple. The tree structure also has a natural visualization, with its nodes and edges.

5 Conclusions and Future Work

In this work, we used ML methods to develop a classification model to determine if an individual is likely to suffer PE syndrome. We also proved the relevance of Duration of Completed Pregnancy in Weeks (PRGLNGTH), Poverty, Water retention/edema in pregnancy (SWLNANKL), Toxemia, Education (EDUCAT), Highest Completed Year School or Degree (HIEDUC), Pregnancy Outcome (OUTCOME), Labor Force Status (LABORFOR) to the classification process and hence to predict if an individual will have the disease. For this purpose, we preprocessed the dataset. We had an imbalanced database, consequently, we used ROC graph from Leave One Cross Validation method to find the like hood to suffer the disease.

Regarding interpretability, we present graphic interpretations obtained from the dataset. First, the variable importance according to the MDGI metric, using Random Forest Classification Method. Random Forest method is acceptable because it has the highest accuracy, specificity, and sensitivity. Second, the decision tree based in four variables to guide the decision way to determine the probability to suffer the disease in case that women present "Water retention/edema in pregnancy", TOXEMIA, pregnancy length is greater or equal to 18 and lesser to 36 weeks, works in pregnancy period (Labor Force Status).

We conclude that the PE study based on socio-demographic and health information obtained in a local population can be valid and extensive to other communities. To the extent that the population studied includes more individuals of different ethnicities, this will provide more information to prevent the condition of PE among pregnant women.

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11…$15.00
https://doi.org/10.1145/3284869.3284912

REFERENCES

1. Duckitt, K., Harrington, D.: Risk factors for pre-eclampsia at antenatal booking: Systematic review of controlled studies. Br. Med. J. 330, 565–567 (2005).
2. Roberts, J.M., Druzin, M., August, P.A., Gaiser, R.R., Bakris, G., Granger, J.P., Barton, J.R., Jeyabalan, A., Bernstein, I. a, Johnson, D.D., Karamanchi, S.A., Spong, C.Y., Lindheiner, M.D., Tsingas, E., Owens, M.Y., Martin Jr, J.N., Saade, G.R., Sibai, B.M.: ACOG Guidelines: Hypertension in pregnancy. (2012).
3. Zayyad, M.A., Toycan, M.: Factors affecting sustainable adoption of e-health technology in developing countries: an exploratory survey of Nigerian hospitals from the perspective of healthcare professionals. PeerJ. 6, e4436 (2018).
4. Cerón-Mireles, P., Harlow, S.D., Sánchez-Carrillo, C.I., Núñez, R.M.: Risk factors for pre-eclampsia/eclampsia among working women in Mexico City. Paediatr. Perinat. Epidemiol. 15, 40–6 (2001).
5. Najman, J.M., Morrison, J., Williams, G.M., Keeping, J.D., Andersen, M.J.: Unemployment and reproductive outcome. An Australian study. Br. J. Obstet. Gynaecol. 96, 308–13 (1989).
6. Vellido, A., Martin-Guerrero, J.D., Lisboa, P.: Making machine learning models interpretable. 20th Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn. 163–172 (2012).
7. Sircar, M., Thadhani, R., Karumanchi, S.A.: Pathogenesis of preeclampsia. Curr. Opin. Nephrol. Hypertens. 24, 131–138 (2015).
8. ACOG Committee on Practice Bulletins-Obstetrics: ACOG practice bulletin. Diagnosis and management of preeclampsia and eclampsia. Number 33, January 2002. Obstet. Gynecol. 99, 159–67 (2002).
9. Kohn, M.A., Carpenter, C.R., Newman, T.B.: Understanding the direction of bias in studies of diagnostic test accuracy. Acad. Emerg. Med. 20, 1194–1206 (2013).
10. Bossuyt, P.M.M.: Clinical validity: Defining biomarker performance. Scand. J. Clin. Lab. Invest. 70, 46–52 (2010).
11. Kenny, L.C., Dunn, W.B., Ellis, D.I., Myers, J., Baker, P.N., Kell, D.B.: Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. Metabolomics. 1, 227–234 (2005).
12. Neocleous, C.K., Anastasopoulos, P., Nikolaides, K.H., Schizas, C.N., Neokleous, K.C.: Neural networks to estimate the risk for preeclampsia occurrence. In: Proceedings of the International Joint Conference on Neural Networks. pp. 2221–2225 (2009).
13. Espinilla, M., Medina, J., García-Fernández, Á.-L., Campaña, S., Londoño, J.: Fuzzy Intelligent System for Patients with Preeclampsia in Wearable Devices. Mob. Inf. Syst. 2017, 1–10 (2017).

14. Velikova, M., Van Scheltinga, J.T., Lucas, P.J.F., Spaanderman, M.: Exploiting causal functional relationships in Bayesian network modeling for personalized healthcare. Int. J. Approx. Reason. 55, 59–73 (2014).
15. Tejera, E., Jose Areias, M., Rodrigues, A., Rama, A., Manuel Nieto-Villar, J., Rebelo, I.: Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes. J. Matern. Neonatal Med. 24, 1147–1151 (2011).
16. Villa, P.M., Marttinen, P., Gillberg, J., Inkeri Lokki, A., Majander, K., Ordén, M.R., Taipale, P., Pesonen, A., Räikkönen, K., Hämäläinen, E., Kajantie, E., Laivuori, H.: Cluster analysis to estimate the risk of preeclampsia in the high-risk Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study. PLoS One. 12, 1–14 (2017).
17. Moreira, M.W.L., Rodrigues, J.J.P.C., Oliveira, A.M.B., Ramos, R.F., Saleem, K.: A preeclampsia diagnosis approach using Bayesian networks. In: 2016 IEEE International Conference on Communications (ICC). pp. 1–5 (2016).
18. Fergus, P., Montanez, C.C., Abdulaimma, B., Lisboa, P., Chalmers, C.: Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women. 1–11 (2018).
19. Cox, B., Sharma, P., Evangelou, A.I., Whiteley, K., Ignatchenko, V., Ignatchenko, A., Baczyk, D., Czikk, M., Kingdom, J., Rossant, J., Gramolini, A.O., Adamson, S.L., Kislinger, T.: Translational Analysis of Mouse and Human Placental Protein and mRNA Reveals Distinct Molecular Pathologies in Human Preeclampsia. Mol. Cell. Proteomics. 10, M111.012526 (2011).
20. Mehta, R., Tech, M., Bhatt, N., Ganatra, A.: A Survey on Data Mining Technologies for Decision Support System of Maternal Care Domain. Int. J. Comput. Appl. 138, 975–8887 (2016).
21. MacKenzie, B.: Blebs of Hope: Searching for markers of preeclampsia in placental exosomes. (2016).
22. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M. V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17 (2015).
23. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437 (2009).
24. Walker, K.E.: Evaluation of Children's Futures: Improving Health and Development Outcomes for Children in Trenton, New Jersey, 2001-2005. Inter-university Consort. Polit. Soc. Res. v1, (2008).
25. Torgo, L.: Data mining with R : learning with case studies. 289 (2011).
26. Friedman, J.H.: 999 Reitz lecture greedy function approximation: a gradient boosting machine 1. Ann. Stat. 29, 1189–1232 (2001).