Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.13, No.4, 2022

An Appropriate Feature Selection Technique for Use on Socio-Demographic Predictor Variables to Enable Early Detection of Preeclampsia: A Review of Literature

Arina A. Jamwa[1*] Mgala Mvurya[1] Antony Luvanda[2] Pamela Kimetto [3]
1.Institute of Computing & Informatics, Technical University Mombasa, PO Box 90420 Mombasa
2.Department of Computing, Alupe University, P.O Box 845-50400 Busia
3.School of Medicine & Health Science, Kabarak University, P.O Box Private Bag 20157 Kabarak
* E-mail of the corresponding author: arinajamwa@gmail.com

Abstract
Preeclampsia is categorized by the World Health Organization as one of the leading causes of high morbidity and mortality in infant and mothers around the world. It accounts for between 3% to 5% of all pregnancy related complications reported worldwide. This condition is much higher among women aged between 30 and 40 years in developing nations especially those in the sub-Saharan region, where the figures range between 5.6% to 6.5% of all reported pregnancies. Preeclampsia is a condition normally detected in the third trimester of pregnancy that is characterized by high risk factors such as sudden High Blood Pressure, High levels of protein in Urine, Chronic kidney disease and Type 1 or 2 diabetes. If preeclampsia is not detected early, it can advance to eclampsia or result to maternal and fetal death. This study sought to identify the optimal features as predictors to enable early detection of preeclampsia through a systematic review of relevant literature. The predictors under consideration were; Maternal age, Occupation, Education, ANC Attendance, BMI, Blood Pressure, Medical History, Urine dipstick, Gravida, Ethnicity, Gestation weeks as identified from literature.
Keywords: Ante natal care service, Preeclampsia, feature engineering, socio-demographic features, machine learning
DOI: 10.7176/CEIS/13-4-02
Publication date: August 31st 2022

1.0 Introduction

The Sustainable Development Goals (SDG) has prioritized universal healthcare as one of the core pillars of any developed and developing country in the world (Marmot & Bell, 2018; Morton et al., 2017). The Millennium Development Goals (MDG) which is a precursor of the SDG, advocated for reduced child mortality and improved maternal health by 2015 (Ibrahim et al., 2019). However, developing nations especially in the sub Saharan region are at a constant risk of high maternal and child mortality rates. Some of the challenges stem from socio-economic to socio-demographic challenges (Vandemoortele, 2018).

As the developed world continues to meet their targets, the developing countries continue to struggle in achieving the set goals (Leal Filho et al., 2019). As the countries continue to stray “off track”, the mother and child healthcare also continues to face more challenges (D’Alessandro & Zulu, 2017). Some of the challenges pregnant mothers experience related to health include; gestational diabetes mellitus, hypertension, preeclampsia, caesarean birth, and post-delivery weight retention (Gayatri Devi Ramalingam, 2021). Other challenges such as access to antenatal healthcare facilities, malnutrition, access to healthcare information, socio-economic and demographic obstacles are some of the many issues affecting pregnant women (World Health Organization. Regional Office for Europe, 2022).

Big data and artificial intelligence technologies have played a big role in the development of machine learning algorithms that predict in the healthcare domain (Bhardwaj et al., 2017). The adoption of machine learning in research and other procedures has assisted immensely in accurate diagnosis and decision-making process (Sidey-Gibbons & Sidey-Gibbons, 2019). The use of machine learning has improved the accuracy of outcomes by identifying patterns, classifying them and providing the best probable outcomes (J. H. Chen & Asch, 2017).

Such independent tools could deliver distinct predictions with a confident degree of assurance based on information that can be collected about the subject, so that researchers and clinicians may be supported by these predictions in order to take better and more effective decisions (Wiens & Shenoy, 2018). Machine learning is being utilized in a variety of medical fields such as cancer diagnosis, drug discovery, personalize treatment, disease prediction and robot surgery (Davenport & Kalakota, 2019).

Data mining is instrumental in extracting useful maternal healthcare information from large sets of data (Shastri & Mansotra, 2019). The use of data mining is particular of use in both structured and unstructured data to discover data patterns that would have been using manual methods. It is through this procedure that maternal healthcare metadata can be interpreted and useful knowledge to predict pregnancy related conditions. To

correctly and correctly extract and analyze maternal health data for ideal prediction outcomes, the maternal health data characteristics or features need to be tested for their relevance and importance as suitable independent variables.

Feature selection involves the preprocessing of data in machine learning to identify suitable predictors that can relate with the dependent variable to build an appropriate model to avoid overfitting (Li et al., 2018). This guarantees that appropriate fields are carefully chosen and improves on the precision and performance of the machine learning predictive result. Feature selection ensures complexity reduction in the dataset, improves the learning efficiency and increase predictive power by noise reduction (Remeseiro & Bolon-Canedo, 2019).

The adoption of feature selection in maternal healthcare data will extract and use the most appropriate socio-demographic feature that will be able to predict preeclampsia. There four types of feature selection methods; *Filter method* are more focused on the data generalization hence computationally effective in feature selection modeling. *Wrapper method* use a learning method that evaluates all candidates in a dataset which computationally costly but offers better performance. *Embedded method* which use the strength of both filter and wrapper methods to offer less computational expense and improving predictive strength of the method(C.-W. Chen et al., 2020).

2.0 Methods and materials

The quantitative research conducted relied on reviewed journal articles to establish independent (Socio-demographic features) and intervening variables (maternal risk factors) to the dependent variable (Preeclampsia). To establish a comparative analysis of the intervening and independent variables to the dependent variable of the study (Bilge et al., 2020).

Carreño & Qiu. conducted a study using imperialist competitive algorithm (ICA) for dimension reduction and sample progression discovery (SPD) algorithm for clustering in the use of machine learning in predicting preeclampsia. The outcome showed appropriate feature selection of variables improved the prediction rate from 85% to 93%. The features selected include; maternal age, education level, occupation, previous medication conditions (Carreño & Qiu, 2020).

Tahir et al. in their study adopted Particle Swarm Optimization (PSO) algorithm to reduce the features from 17 to 7 attributes. The reduction greatly improved the accuracy of the deep learning model from 95.12% to 95.68% in predicting preeclampsia model. Which led to faster execution time by reducing the dataset size. The reduced features are as follows; Age of the mother, access to ANC facility, occupation, education, preexisting conditions, and residence (Tahir et al., 2018).

Irfan et al. used correlation-based feature selection (CBFS) and C5.0 algorithms in pregnancy risk monitoring by selecting the appropriate variables for their machine learning model. The CFS is a proven feature selection method that offers satisfactory results in medical diagnosis. C5.0 algorithm produced better accuracy with less memory consumption. The features used; maternal age, gestation age, previous pregnancy complications, hypertension, number of previous pregnancies, gap between pregnancies, and access to Ante Natal Care (Irfan et al., 2021).

Kurniawan et al. used correlation-based feature selection (CBFS) algorithm to reduce the features from 8 to 4 in predicting a new born baby's health. The outcome of the test showed an increase in Precision, recall, and accuracy in the Naïve Bayes classification when implemented CBFS. The accuracy after preprocessing was 67% an increase of 2% before preprocessing in predicting preeclampsia. The features selected were; blood pressure, number of babies delivered previously, congenital diseases before pregnancy and problems during pregnancy (Kurniawan et al., 2020).

Kumar et al. conducted a study on exposure of polycyclic aromatic hydrocarbons on pregnant women through inhalation and diet consumed. The study investigated the association of maternal socio-demography and blood polycyclic aromatic hydrocarbons on low birth weight. A total of One hundred and seventy-five pregnant women from Dibrugarh, Assam in India. The features extracted were; maternal age, occupation, residence, education, Body Mass Index (BMI) etc. The model was build using SVM light and Waikato Environment for Knowledge Analysis (Weka). Multiple Logistic regression depicted improved probability of low birth weight due blood PAH. The model predicted LBW offspring with 84.35% sensitivity and 74% specificity. Occupation, BMI, nutritional habits were better predictors and offered the best machine learning precision (Kumar et al., 2020).

Desyani et al. conducted an experimental research using python to implement a feature selection and classification of caesarian section prediction which was extracted from the University of California, Irvine data repository. The classification algorithm used was Naïve Bayes. Feature selection methods are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), Sequential Forward Floating Selection (SBFS), Sequential Backward Floating Selection (SBFS), and selectKBest. The Heart Problem feature in the dataset offered better predictor in the study. Naïve Bayes with SelectKBest feature selection technique offered better performance and accurate predictions on C-section

(Desyani et al., 2020).

Pavlova et al. did a study on anxiety level determinant among males in military conflict threat experience in Ukraine using socio-demographic features. A questionnaire was administered consisting of socio-demographic data, military service information and quality of life. The framework consisted of five phases; data selection, data preprocessing, number of clusters and their interpretation, classifier learning and validation for optimal accuracy and lastly trained model for prediction of test data. Quality of life was the best feature used in the model, as it accurately predicted the anxiety level in young military males in threat experiences during active or inactive duties (Pavlova et al., 2020).

Singh et al. in their research worked on a software-based liver disease prediction model using feature selection and various classification algorithms such as Logistic regression, Random Forests, Naïve Bayes, J48 etc. Some of the attributes extracted include; age, gender, total proteins albumin and Globulin ratio etc. and used as features during the training and test modeling phase. Data selection, feature selection, pre-processing and transformation, classifier implementation, performance evaluation, disease precision and software development were implemented on the data collected from University of California, Irvine using WEKA tool with 10-fold cross validation. Logistic regression performed best with highest accuracy and execution time compared to other classifiers (Singh et al., 2020).

Khan et al. in their research study used supervised machine learning algorithms in predicting childhood anemia in Bangladesh using a retrospective cross-sectional study with a sample size of 2013 children under the age of five. Data was extracted from the Bangladesh Demographic and Health survey for fitting predictors based on maternal and paternal variables. Some of the independent variables considered for the study for both parents; age, education, anemia family history, place of residence, wealth index, and child size at birth etc. Random Forests offered the best performance, accuracy in predicting anemia in Bangladesh with socio-demographic features and health characteristics being important features (Khan et al., 2019).

M. S. Bin Alam et al., used ensemble bagging classification to predict birth mode i.e. caesarean section or normal delivery with the aim of finding hidden patterns. A cross sectional study was used in the creation of the dataset collected from the Bangladesh Demographic and Health survey that collected 4493 samples. The target population was women who had children in the last five years of between 15 to 49 years of age. Data cleaning and feature selection was applied to the collected data which were used on the test and training. Commonly used algorithms such Naïve Bayes (NB), logistic regression, k-nearest neighbor (KNN), support vector machine (SVM) and Decision Tree (DT) used for prediction. Using bagging ensemble classifiers which is a novel technique against socio-demographic factors improved birth mode prediction and offered precautionary plan for Bangladesh pregnant women (M. S. Bin Alam et al., 2021).

Espinosa et al., used machine learning and feature selection in predicting various pregnancy related complications. The dataset was collected was subjected to feature engineering in extracting the best and optimal socio-demographic predictors. The maternal health challenges used in the study included; preeclampsia, prematurity, and stillbirth. Their approach was based on modeling data maternal health data against socio-demographic factors such as maternal age, medical complications, access to ANC facilities, previous pregnancy complications, educational status, income availability etc. The outcome of their study showed that women in low- and middle-income areas are more vulnerable to facing pregnancy complications based on socio-demographic factors. The features used in the dataset on the machine learning model gave accurate outcomes and better performance (Espinosa et al., 2021).

Prediction of future progression of type 2 diabetes was conducted by M. S. Islam et al. whereby they used feature selection to optimal variables as predictors for type 2 diabetes study. The study used risk factors as part of the independent variables in developing a machine learning prediction model. The target population were 5,158 participants of Mexican American and non-Hispanic whites between 25-64 years of age. Glucose and insulin index were used as part of the predictors in addition to age, nutrition type, diabetes family history, and other clinical importance features. The model had a accuracy of 95.94%, sensitivity of 100% and specificity of 91.5% which outperformed previous models (M. S. Islam et al., 2020).

Previous studies from the review of relevant literature have focused on a single risk factor against clinical factors. The previous studies have not focused deeply on socio-demographics as an integral factor in maternal health. This has limited the accuracy and performance of the prediction models in the past studies leading to late detection of maternal health conditions such as preeclampsia. This paper proposes a model that will use a combination of preexisting preeclampsia risk factors and socio-demographic features to enhance the model accuracy and performance in predicting preeclampsia in the early onset stage of the second trimester of the pregnancy.

3.0 Findings

Maternal and Child Health continues to be an important pillar in any developed or developing country. Challenges related to maternal health such as preeclampsia, continues to be a major concern in relation to

maternal health challenges around the globe. If preeclampsia is not detected early enough, it progresses to eclampsia, premature births, organ damages, preterm birth, and placental abruptions.

Correlation-based feature selection which is a filter-based method, ensures that all the selected variables are relevant and optimal for building the prediction model. Correlation-based feature selection compares between the available variables and selects those that are in correlation. Correlation-based feature selection ensures that the relevant and suitable variables are selected and the those that irrelevant are ignored. This will reduce the curse of dimensionality, improve on the precision of the classification model and prevent overfitting challenges due to inappropriate variables.

Bibliography

Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). A Study of Machine Learning in Healthcare. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 236–241. https://doi.org/10.1109/COMPSAC.2017.164

Bilge, C., Mecdi Kaydirak, M., Gür Avci, D., & Hotun Sahin, N. (2020). Effect of Shift Working on Depression Prevalence and Sexual Life of Female Nurses: A Correlational Study in Turkey. *International Journal of Sexual Health*, *32*(4), 357–364. https://doi.org/10.1080/19317611.2020.1819502

Carreño, J. F., & Qiu, P. (2020). Feature selection algorithms for predicting preeclampsia: A comparative approach. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2626–2631. https://doi.org/10.1109/BIBM49941.2020.9313108

Chen, C.-W., Tsai, Y.-H., Chang, F.-R., & Lin, W.-C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, *37*(5), e12553. https://doi.org/10.1111/exsy.12553

Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *The New England Journal of Medicine*, *376*(26), 2507–2509. PubMed. https://doi.org/10.1056/NEJMp1702071

D'Alessandro, C., & Zulu, L. C. (2017). From the Millennium Development Goals (MDGs) to the Sustainable Development Goals (SDGs): Africa in the post-2015 development Agenda. A geographical perspective. *African Geographical Review*, *36*(1), 1–18. https://doi.org/10.1080/19376812.2016.1253490

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. PubMed. https://doi.org/10.7861/futurehosp.6-2-94

Desyani, T., Saifudin, A., & Yulianti, Y. (2020). Feature Selection Based on Naive Bayes for Caesarean Section Prediction. *IOP Conference Series: Materials Science and Engineering*, *879*(1), 012091. https://doi.org/10.1088/1757-899X/879/1/012091

Espinosa, C., Becker, M., Marić, I., Wong, R. J., Shaw, G. M., Gaudilliere, B., Aghaeepour, N., Stevenson, D. K., Stelzer, I. A., Peterson, L. S., Chang, A. L., Xenochristou, M., Phongpreecha, T., De Francesco, D., Katz, M., Blumenfeld, Y. J., & Angst, M. S. (2021). Data-Driven Modeling of Pregnancy-Related Complications. *Reproductive and Sexual Health*, *27*(8), 762–776. https://doi.org/10.1016/j.molmed.2021.01.007

Ibrahim, M. D., Daneshvar, S., Hocaoğlu, M. B., & Oluseye, O.-W. G. (2019). An Estimation of the Efficiency and Productivity of Healthcare Systems in Sub-Saharan Africa: Health-Centred Millennium Development Goal-Based Evidence. *Social Indicators Research*, *143*(1), 371–389. https://doi.org/10.1007/s11205-018-1969-1

Irfan, M., Basuki, S., & Azhar, Y. (2021). Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning. *Bulletin of Electrical Engineering and Informatics*, *10*(3), 1621–1633. https://doi.org/10.11591/eei.v10i3.2344

Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2019). *MACHINE LEARNING ALGORITHMS TO PREDICT THE CHILDHOOD ANEMIA IN BANGLADESH*. 24.

Kumar, S. N., Saxena, P., Patel, R., Sharma, A., Pradhan, D., Singh, H., Deval, R., Bhardwaj, S. K., Borgohain, D., Akhtar, N., Raisuddin, S., & Jain, A. K. (2020). Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. *Reproductive Toxicology*, *94*, 92–100. https://doi.org/10.1016/j.reprotox.2020.03.009

Kurniawan, Y. I., Cahyono, T., Nofiyati, Maryanto, E., Fadli, A., & Indraswari, N. R. (2020). Preprocessing Using Correlation Based Features Selection on Naive Bayes Classification. *IOP Conference Series: Materials Science and Engineering*, *982*(1), 012012. https://doi.org/10.1088/1757-899X/982/1/012012

Leal Filho, W., Tripathi, S. K., Andrade Guerra, J. B. S. O. D., Giné-Garriga, R., Orlovic Lovren, V., & Willats, J. (2019). Using the sustainable development goals towards a better understanding of sustainability challenges. *International Journal of Sustainable Development & World Ecology*, *26*(2), 179–190. https://doi.org/10.1080/13504509.2018.1505674

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data

Perspective. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

M. S. Bin Alam, M. J. A. Patwary, & M. Hassan. (2021). Birth Mode Prediction Using Bagging Ensemble Classifier: A Case Study of Bangladesh. *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 95–99. https://doi.org/10.1109/ICICT4SD50815.2021.9396909

M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, & M. A. Abdul-Ghani. (2020). Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes. *IEEE Access*, *8*, 120537–120547. https://doi.org/10.1109/ACCESS.2020.3005540

Marmot, M., & Bell, R. (2018). The Sustainable Development Goals and Health Equity: *Epidemiology*, *29*(1), 5–7. https://doi.org/10.1097/EDE.0000000000000773

Morton, S., Pencheon, D., & Squires, N. (2017). Sustainable Development Goals (SDGs), and their implementation. *British Medical Bulletin*, 1–10. https://doi.org/10.1093/bmb/ldx031

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Bennie, E. (2017). MatLab vs. Python vs. R. *Journal of Data Science: JDS*, *15*, 355–372.

Pavlova, I., Zikrach, D., Mosler, D., Ortenburger, D., Góra, T., & Wąsik, J. (2020). Determinants of anxiety levels among young males in a threat of experiencing military conflict–Applying a machine-learning algorithm in a psychosociological study. *PLOS ONE*, *15*(10), 1–24. https://doi.org/10.1371/journal.pone.0239749

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Shastri, S., & Mansotra, V. (2019). Data Mining Probabilistic Classifiers for Extracting Knowledge from Maternal Health Datasets. *International Journal of Innovative Technology and Exploring Engineering*, *9*, 2769–2776. https://doi.org/10.35940/ijitee.B6633.129219

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. *Procedia Computer Science*, *167*, 1970–1980. https://doi.org/10.1016/j.procs.2020.03.226

Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 977–982. https://doi.org/10.23919/MIPRO.2019.8757088

Tahir, M., Badriyah, T., & Syarif, I. (2018). Classification Algorithms of Maternal Risk Detection For Preeclampsia With Hypertension During Pregnancy Using Particle Swarm Optimization. *EMITTER International Journal of Engineering Technology*, *6*(2), 236–253. https://doi.org/10.24003/emitter.v6i2.287

Vandemoortele, J. (2018). From simple-minded MDGs to muddle-headed SDGs. *Development Studies Research*, *5*(1), 83–89. https://doi.org/10.1080/21665095.2018.1479647

Wiens, J., & Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153. https://doi.org/10.1093/cid/cix731