Early Prognosis of Preeclampsia Using Machine Learning

E. Sivaram, G. Vadivu, K. Sangeetha, Vijayan Sugumaran

Early prognosis of preeclampsia using machine learning.

Sivaram. E[1][0000-0002-7278-839X], Dr.G. vadivu [2][0000-0003-2982-4145], Ms. Sangeetha k [3][0000-0002-3393-8965] and Vijayan Sugumaran, Ph.D.[4] [0000-0003-2557-3182]

[1] SRMIST, Kattankulathur, Chengalpet-603203, Tamilnadu, India
se3033@srmist.edu.in
[2] SRMIST, Kattankulathur, Chengalpet-603203, Tamilnadu, India
vadivug@srmist.edu.in
[3]SRMIST, Kattankulathur, Chengalpet-603203, Tamilnadu, India
sangeetk@srmist.edu.in
[4]Oakland University, Rochester, MI 48309, USA.
sugumara@oakland.edu

Abstract. Preeclampsia is a type of hypertension condition that can be induced by a variety of circumstances during pregnancy. Typically, a diagnosis is made after 20 weeks of gestation. Several investigations employing machine learning techniques have been undertaken to diagnose preeclampsia. SVM, KNN, random forest, gradient boosting methods, and deep learning approaches are examples of these. These techniques can be implemented to detect preeclampsia earlier in an efficient way for preventing the complications caused. This paper demonstrates how hyperparameter tuning of Support vector classification of the various factors involved in the classification of preeclampsia helps in efficiently separating the patients who are prone to have preeclampsia. The selection of the hyperparameter is done through the Grid Search CV algorithm by iterative trialing of the different hyperparameters.

Keywords: Preeclampsia, SVM, Grid Search, Hyperparameter.

1 Introduction

One of the hypertensive outcomes of pregnancy is preeclampsia. Preeclampsia influences 3-5 percent of pregnant ladies around the world, remembering 5.4 percent for India. This can likewise be lethal for both the mother and the embryo [1]. Preeclampsia is related with vasospasm, pathologic vascular sores in various organ frameworks, expanded platelet initiation, and ensuing enactment of the coagulation framework in the miniature vasculature. After the twentieth seven day stretch of pregnancy, preeclampsia is progressively logical. Several studies have shown the ratio of the PlGF to sFLT-1 and Placental protein 13, soluble Endoglin, Triglycerides and Cystatin C as major biomarkers for the detection of the preeclampsia [1-4]. Also, many physiological and clinical parameters are also used in detection of the preeclampsia [5]. The proposed system is an application based on the hyperparameter tuning of which takes the above-mentioned data as input. This will help us in predicting and protecting the patient who

are prone to preeclampsia earlier. Initially the patients are recruited in the first trimester of the pregnancy. The physiological data, blood count values, Thyroid level, medical history of the particular patient and family along with regular activity data will be collected. With these data collected a statistical analysis will be conducted using and a comprehensive review on the factors for the preeclampsia will be discussed. The angiogenic factor that facilitates the growth of the placenta during the gestation is placental growth factor (PlGF) and its receptor Soluble fms-like tyrosine kinase 1 (s-FLT1) [6]. Other than these two biomarkers there are other biomarkers that will be taken as parameters, which are the Soluble endoglin(sENG) [7], placental protein 13 [2], Triglycerides and Cystatin C [8]. These markers are also expressed during the first trimester of the gestation, and the fluctuation in their values are caused in preeclampsia. The PlGF/ sFLT-1 and sENG/PlGF are more accurate and sensitive. One of the most used algorithms is Support vector classifier with hyperparameter tuning will lead to have optimum selection of the parameter using Grid Search cross validation. This helps is to know how various kernels and their hyperparameter involves in the setting of boundaries with coverage of the data point. The Gamma, C, degree are the various hyperparameters. This hypermeter varies to different kernel that is considered for the SVC. This approach leads to define what kernel along with their hyperparameter that needs to tune to have good accuracy.

2 Related literature work

The studies that are intended to detect preeclampsia are mostly done after 20 weeks of the gestation and studies that use biomarkers like cystatin C and Placental protein are earlier in detection of preeclampsia patients had significantly greater levels of sFlt-1/PlGF than control women. Data suggest that the sFlt-1/PlGF ratio has a higher accuracy for distinguishing PE patients from non-PEs than it does for distinguishing severe or early onset forms of the disease [1]. Women with preeclampsia who experienced complications had a considerably greater PlGF/sFLT1 ratio than women with preeclampsia who did not develop issues [6]. Various serum markers were used to detect the preeclampsia, amongst them sFLT-1 /PlGF ratio was most promising in detecting the preeclampsia [4]. Associations between many first-trimester maternal factors and placental Doppler investigations that are connected to placental performance and serum PlGF levels to uncover significant relationships that should be considered in screening procedures When trophoblast cells were compared, they produced considerably more sEng, sFlt-1, and PlGF. compared to those from normal TCs without preeclampsia, which is more important and addresses the problem's core cause [9]. Placental Protein 13 (PP13) is ensnared in the pathophysiology of hindered placentation and the resulting improvement of early PE but estimating this placental protein at 11-13 weeks is unlikely to be viable in sickness screening. Maternal PP13 levels separating the principal trimester is a promising symptomatic procedure for foreseeing preeclampsia with great awareness and importance in the primary trimester. [3]. The different variables of hypertension with Artificial neural network and other machine learning

algorithms like SVM, decision tree and application for the equivalent [10]. The expectation of the preeclampsia by the random forest algorithm with 17 variables. It had the best AUROC in external validation. This minimal expense algorithm upgraded primer expectation to decide if pregnant ladies would be anticipated by models with high particularity and progressed indicators [11]. As a modelling procedure, the hidden Markov model was utilized. Knowledge is used to gain a better grasp of how an illness develops. The training method is hampered by prior knowledge of preeclampsia. Preeclampsia was classified, and the observations were categorized [12]. The usage of the elastic net algorithm, containing the informative model with most features for the prediction of preeclampsia [13]. The biomarkers are considered for the for producing the insights for preeclampsia and ANN used depicts varies features in preeclampsia [5].

3 Proposed Work

Studies have revealed that demographics and clinical indicators such as the total blood count are stronger predictors of preeclampsia. Placental growth factor, soluble fms like tyrosine kinase, soluble Endoglin, and other biomarkers were included in our study. Placental protein 13 showcases the novelty of the study [1-4,7]. Once all the data are collected, the data will be pre-processed. After the pre-processing of the data, selection of the model is done in our case its Support vector classifier and this algorithm will be executed using scikit learn package. An incremental evaluation of all hyperparameters must then be performed to determine the ideal hyperparameter. That hyper parameter must be used to train the model. The trained model is then compared to the testing data to evaluate accuracy before the model is deployed. The support vector classifier is one of the most widely used machine learning techniques (SVC). This also applies to classification and regression problems. It also uses the kernel approach to adjust the data, and based on these modifications, the model finds the best fitting boundary or hyperplane for segregating preeclamptic patients from non-preeclamptic patients in our example.

The hyperplane can be circle, line, or sigmoidal plane, but it there with their margins to separate or to classify the different classes of data points. This transformation is called Kernel. This normally involves in every Support vector classification, but this produces generic results or classification. This model can be tuned by the introducing the hyperparameter to the model.

4 Tuning of Kernel Hyperparameter

The model first goes through a training phase where the system is trained with collected data which has undergone the pre-processing. This is where the system learns various patterns in the data that is fed into the system. In this supervised phase of the model, the output produced is compared with the actual or desired output for the pattern of the data that is given. The difference in the output value will be compared and iterative

tuning of the hyperparameter will be done for the model to choose the best hyperparameters. This includes the tuning of C and Gamma function and the kernel of the Support vector classifier until the output produces lowest error possible with higher accuracy, more of the true positive and true false values compared to false positive and false negative values. Since different factors are used in the detection of the preeclampsia and various machine learning algorithm has parameters that can be tuned, called hyperparameter. It is necessary that it is to be notified before training the model. This approach helps in creating a more robust and accurate model and also creates the balance between the bias and variance preventing the model from the underfitting or overfitting.

The hyperparameters that are actually considered for the support vector classifier are C and Gamma. The SVC's major operation is to create a decision boundary which segregates different classes, binary in our study, need not be definitely a straight line. Since the real-world data is noisy and dirty, this might lead to the overfitting or underfitting of the model. The major concern to have great model are,

- To have wider decision boundary for classes.
- To minimize the false positive and false negative prediction.

to achieve this an obvious trade-off must be taken care, as a matter of fact that decision boundaries are sensitive to minor changes.

The trade-off could be controlled by C parameter, which implements the cost for each misclassified data point. Considering that C is small and the cost for the misclassified is low so that, the decision boundary has higher margin which is included. If the C is large, the tries to reduce the number of the misclassified values or the datapoints along with a narrowed margin. The cost involved is actually directly proportional to the distance between the decision boundaries.

In many cases, the dataset will not be linearly separable; however, employing the kernel function will make it so. This is also true in our case. As a result, the kernel was chosen, and it is far better suited for the classification process. It should also be more accurate in data point segregation than other kernels such as poly, linear, rbf, and so on. [14] The kernel function converts the input data points into the required output data points. In this case, we're interested in the polynomial kernel for the SVC.

The polynomial kernel doesn't only take in account the similarity of the vector in a particular dimension but also the cross dimension. This enables us to have the interaction between the features. The polynomial function is computed by the degree (d) of the polynomial kernel between two vectors in-turn describing the relationship between vectors [15,16].

Formally, the Polynomial kernel is defined by the equation:

$$K(x, y) = (\gamma . x.y + C)^{d} \quad (1)$$

The x and y are input vectors and d is degree of polynomial function.

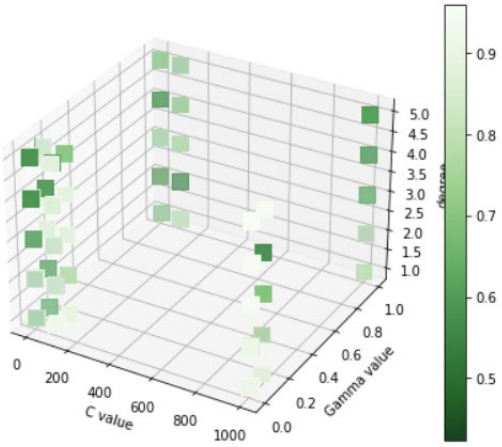

Fig. 1. Visualisation of the accuracy produced by various iteration

The accuracy produced by the different hyperparameters is shown visuall. With the polynomial kernel, C value at 0.5 and gamma value 1, it will produce the good classification the data points.

This value is achieved by various iteration of the C, gamma, kernel and degree value, this is carried by the algorithm grid search cv. This works as follows:-

- The selection of the machine learning that is to be tuned here, SVC.
- Deciding the parameters such as Kernel (poly, linear, rbf, sigmoid), gamma, C, (10^{-3} to 10^{3}) and the degree(1,2,3,4,5..).
- From these parameters iterative application of the parameters is done.
- The scores produced from the iterations are compared to choose the best C and gamma value along with accuracy score

After the training and test is done, the system is ready for the prediction, once the patients register and the describe demographical and clinical data are collected initially. With data model is created to predict what is the probability a particular patient is prone to have preeclampsia. The cut-off values will be set through the statistical analysis that will be done for the data collected. Once the prediction for the patient is above cut-off value then, the patient is marked as the preeclampsia prone patient.

5 Discussion and Results

Based on the above discussion, the prediction of the preeclampsia in the earlier stages can be done by incorporating various biomarkers which has their significance in the first trimester during the development of the placenta. This approach helps to address

the root cause and the pathology of the preeclampsia. Since, considering the facts that Placental growth factor and soluble fms like trysine kinase 1, and other early biomarkers like the cystatin C and serum lipids has a significant weightage in the early prediction of the preeclampsia and the fluctuation of these biomarkers in first trimester helps determine the preeclampsia condition in the earlier stage. The hyperparameter tuning will provide us the insights in various data pattern and distinguishes minor difference between the preeclamptic people. Also, the C value is only one parameter that needs to be tuned for the linear model. However, considering the polynomial model, both the C and gamma values must be tuned to achieve good accuracy.

Table 1. comparison of the accuracy for various kernel

C value	Degree	Gamma	Kernel
0.5	1	1	Poly
0.5	1	1	Rbf
0.5	1	1	Sigmoid

Table 1 shows the various values of the hyperparameters that produce good results. The polynomial kernel seems to produce the most accurate results compared to others. In future even if more data points are added for the further training purposes the polynomial kernel will be able produce more result compared to the linear kernel. Since the linear kernel includes only the C value and not the gamma and might be easily influence the margin compared to the Polynomial kernel. Also, the polynomial kernel is more likely to handle data points of larger dimensionalities, and hence better suited for our study.

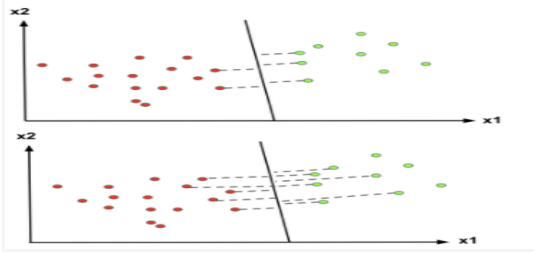

Fig. 2. Comparison of low and high gamma values

The lower gamma value has higher decision boundary taking many data points into account. Also, from the table1 parameter the degree is 1 and the gamma is 1 and change is seen only in the C value which makes both the linear and polynomial kernel to same.

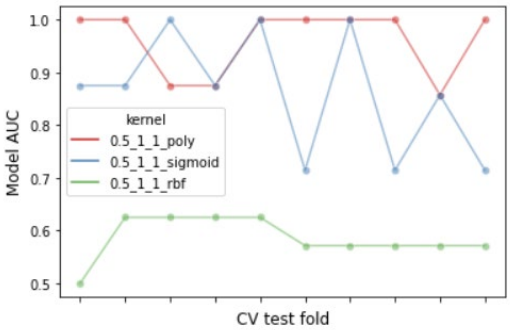

Fig. 3. Cross validation results of various SVM kernel

(See Fig. 4) depicts the 10-fold cross validation of the various kernels considered with stipulated amount of the data point at each fold and proves that the polynomial kernel is most suited for our dataset and produces a good classification of the datapoint this produces us the very lesser false positive and false negative datapoint.

Table 2. correlation between the various SVM kernel

Kernel	**Poly**	**Sigmoid**	**Rbf**
Poly	1.000	-0.269	-0.331
Sigmoid	-0.269	1.000	0.358
Rbf	-0.331	0.358	1.000

The table 2, explains that the correlation between the both the kernel poly and linear are more lesser considering our dataset and dimensionalities.
The tuning of the hyperparameters have given us a good accuracy compared to the SVC with without the hyperparameter tuning and previous study quoted in the related work.

Table 3. Comparison of the accuracy with previous study

Accuracy of SVC w/o hyperparameter tuning	Accuracy of SVM in previous study	Accuracy of SVC with hyperparameter tuning
84.6 %	89.2%	96.07%

From the above table is seen that the tuning of the hyperparameter helps in correctly predicting patients prone to preeclampsia. This model is also efficient even if the data points expand in the future.

6 Future Work

Further, the product of this study has to be Once after the model is ready for the deployment, application that is lightweight with minimalist interface is planned to develop so that the end user it more useful.

7 Reference

1. P. Nikuei, M. Rajaei and N. e. a. Roozbeh, "Diagnostic accuracy of sFlt1/PlGF ratio as a marker for preeclampsia," BMC Pregnancy Childbirth, vol. 20, p. 80, 2020.
2. A. R., S. A., J. Beta, K. R. and N. K.H., "Maternal serum placental protein 13 at 11–13 weeks of gestation in preeclampsia.," vol. 29, pp. 1103-1108, 2009.
3. I. Chafetz, I. Kuhnreich, M. Sammar, Y. Tal, Y. Gibor, H. Meiri, H. Cuckle and M. Wolf., " First-trimester placental protein 13 screening for preeclampsia and intrauterine growth restriction.," American journal of Obstetrics and Gynecology., vol. 197, no. 1, 2007.
4. J. Wang, H. Hu and X. L. e. al., " Predictive values of multiple serum biomarkers in women with suspected preeclampsia: a prospective study," available at Research Square, vol. 3, no. PREPRINT, 2020.
5. T. M. Nair., "Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia," Computational Biology and Chemistry, vol. 75, 2018.
6. V. Pant, Y. B.K. and J. Sharma, " A cross sectional study to assess the sFlt-1:PlGF ratio in pregnant women with and without preeclampsia.," BMC Pregnancy Childbirth, vol. 19, p. 266, 2019.
7. A. Leaños-Miranda, C. S. Navarro-Romero, L. J. Sillas-Pardo, K. L. Ramírez-Valenzuela, I. Isordia-Salas and L. M. Jiménez-Trejo, "Soluble Endoglin As a Marker for Preeclampsia, Its Severity, and the Occurrence of Adverse Outcomes.," Hypertension, vol. 74, no. 4, pp. 991-997, 2019.
8. B. Mukherjee and G. Sarangi., "Predictive significance of serum Cystatin-C and serum lipid in preeclampsia," International Journal of Clinical Obstetrics and Gynaecology., vol. 2, pp. 24-28, 2018.
9. Y. Gu, D. F. Lewis and Y. Wang., " Placental Productions and Expressions of Soluble Endoglin, Soluble fms-Like Tyrosine Kinase Receptor-1, and Placental Growth Factor in Normal and Preeclamptic Pregnancies," The jounal of Clinical Endocrinology & Metabolism., vol. 93, no. 1, pp. 260-266, 2008.
0. M. A. J. Tengnah, R. Sooklall and S. D. Nagowaha., "A Predictive Model for Hypertension Diagnosis Using Machine Learning Techniques.," Telemedicine Technologies Big Data, Deep Learning, Robotics, Mobile and Remote Application for Global Healthcare, vol. 9, pp. 139-152, 2019.

1. H. S. MD, Y.-W. W. P. and P. Emily Chia-Yu Su, " Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia," EBioMedicine, vol. 54, 2020.
2. I. Marin, B. Pavaloiu, C. Marian, V. Racovita and N. Goga., "Early Detection of Preeclampsia based on a Machine Learning Approach," E-Health and Bioengineering Conference (EHB),Iasi, Romania, pp. 1-4, 2019.
3. A. T. N. A. A. M. D. K. S. G. M. S. V. D. W. Ivana Marić, "Early prediction of preeclampsia via machine learning," American Journal of Obstetrics & Gynecology MFM, vol. 2, no. 2, 2020.
4. v. Rijn, J. N and F. Hutter, "Hyperparameter Importance Across Datasets," Association for Computing Machinery, p. 2367–2376, 2018.
5. A. Rojas-Domínguez, L. C. Padierna, J. M. C. Valadez, H. J. Puga-Soberanes and H. J. Fraire, "Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis," IEEE Access, vol. 6, pp. 7164-7176, 2018.
6. E. Duarte and J. Wainer, "Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters," vol. 88, pp. 6-11, 2017.