International Journal of Medical Engineering and Informatics

ISSN online: 1755-0661 - ISSN print: 1755-0653
https://www.inderscience.com/ijmei

A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification

Himani S. Deshpande, Leena Ragha

DOI: 10.1504/IJMEI.2023.10051207

Article History:

Received:	15 September 2020
Last revised:	14 February 2021
Accepted:	17 February 2021
Published online:	30 November 2022

A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification

Himani S. Deshpande* and Leena Ragha

Department of Computer Engineering,
Ramrao Adik Institute of Technology,
Mumbai, Maharashtra, India
Email: himaniuphigh@gmail.com
Email: leena.ragha@gmail.com
*Corresponding author

Abstract: Every woman's body is unique and will have some features playing a vital role contributing towards a healthy pregnancy and manually it is difficult to decide the important features to be observed to prevent the pregnancy complications. In this proposal we have consider 21 physical features of 903 women of varied age groups, economy status and health conditions. Variation and information-based random forest (VIBRF) hybrid model using mutual information and F-score is applied to evaluate each feature looking into the variation within the feature and mutual information across the features. We experimented using various classifiers, and it is observed that Gaussian NB has shown most significant improvement in terms of prediction accuracy, from 31% with all features to 80% with our feature selection process. Though SVM prediction accuracy is 84% it is observed AUC drastically improved for GNB by 10%. As it is a medical application, it is important to achieve higher AUC and so through this experiment it is concluded that GNB performs better with proposed model.

Keywords: features selection; F-score; decision tree; random forest; hybrid model; preterm birth; classification.

Reference to this paper should be made as follows: Deshpande, H.S. and Ragha, L. (2023) 'A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification', *Int. J. Medical Engineering and Informatics*, Vol. 15, No. 1, pp.84–96.

Biographical notes: Himani S. Deshpande is a research scholar and she is pursuing her PhD in Computer Science from the Mumbai University. Her areas of interest are data science, machine learning and logic design. She has completed her Master's in Engineering from the Mumbai University and Bachelor of Technology from the Uttar Pradesh technological University. She is serving as an Assistant Professor, with an around six years of teaching experience. Various students have been actively involved in different research activities under her.

Leena Ragha is a Professor and the Head of the Department of Computer Engineering Department, Ramrao Institute of Technology. She holds a Doctorate in Engineering and Technology. She is a member of the Board of Studies in Computer Engineering at the University of Mumbai, India. She has more than 65 research papers published in reputed journals, national and international conferences. She has around 30 years of teaching experience and has immensely contributed to the growth of research in various domains.

1 Introduction

Pregnancy is considered as a beautiful phase of a woman's life, but with today's lifestyle they are facing various complications. Preterm delivery is one such complication that can affect the health of both mother and child. The babies born before 37 weeks or 259 days of gestation are considered premature babies and such early deliveries are termed as preterm birth (PTB) (Pari et al., 2017). Preterm childbirth is the leading cause of mortality among children below the age of five years (https://www.who.int/newsroom/fact-sheets/detail/preterm-birth). As the technologies are advancing there is need of providing smart solutions to realise the important women's physical health features that can be monitored and controlled by the individuals to prevent complications and PTB. The ultimate goal is to realise the health issues ahead of time using our proposed solution to ensure a happy and healthy family.

Through the intensive survey it is observed that most of the researches on maternal issues are based on clinical and obstetric parameters (Catley et al., 2006; Idowu et al., 2014; Son et al., 2017), which can be monitored only with the help of medical personnel using proper equipments. Work done towards pregnancy outcomes have focused only on designing prognostic models, using existing statistical or machine learning techniques (Collins et al., 2015; Robinson et al., 2010; Vogel et al., 2005; Von Dadelszen et al., 2011). There is a need to design a cost effective prognostic model that can help the mother to do the self testing based on the variations in physical parameters observed and seek the medical help if the PTB complications are predicted.

The paper is organised with a detailed technical literature survey in Section 2. Section 3 talks about the dataset and methodologies used in the proposed research. Section 4 gives a detailed view of the proposed system. The results are discussed in Section 5 and concluded in Section 6 along with the future scope.

2 Literature survey

We researched for the current work happening with respect to preterm birth and realised that standard datasets are not available in PTB and researchers have created local datasets majorly with obstetric features and test results. For PTB domain researchers have focused on using existing statistical and machine learning methodologies. It is also observed that research work concentrates upon the methodology of prediction and not on the methods for analysing importance of mother's features. Therefore, we aimed on developing our own dataset and have researched various domains of technologies that may suit our application, towards feature selection. In the following paragraphs, we focused on the work done in PTB, then we looked into features selection researches.

Allen et al. (2016), worked with an integrated qualitative and quantitative analysis to realise antenatal care suggestions to avoid PTB among young girls. Thomas and Kulanthaivel (2016), with 5 features and 1,052 records, worked to minimise numbers of rules used for competitive co-evolution PTB prediction. Morken et al. (2014), worked to predict the risk of spontaneous preterm birth, they used multiple logistic regression. Grzymala-Busse and Woolery (1994) came up with bucket brigade genetic algorithm for

predicting preterm delivery, they used attributes like infant sex, risk factors and age of mother for setting classification rules using LERS. Lee et al. (2011) predicted spontaneous preterm birth using demographic, clinical, and genetic factors for 522 deliveries, using chi-test, t-test, Bayesian filtering, statistical technique. Pari et al. (2017) predicted preterm birth with 2,600 samples and used ensemble learning for fine-tuning. Catley et al. (2006) and Idowu et al. (2014) implemented multiple artificial neural networks to predict preterm birth using obstetric, electromyography features for preterm birth prediction. Son et al. (2017) analysed the importance of cervical length and foetal fibronectin using ultrasound images, with the help of statistical methods. It is important to identify the strong features so that woman can observe the variations to prevent the PTB complications. Towards this we surveyed feature selection methodologies applied in various domains.

Feature selection methods are broadly divided into three categories, namely filter methods, wrapper methods and embedded methods (Chandrashekar and Sahin, 2014; Khalid et al., 2014; Sheikhpour et al., 2017). It is suggested that selecting relevant features by combining methods, i.e., hybrid model, shows improvement in terms of classification accuracy, speeds up the process, and reduces error rate (Lee, 2009; Jiang et al., 2017; Zhang et al., 2018). We further discuss researches on hybrid models. Lu et al. (2017), proposed MIMAGA-selection algorithm, they evaluated features based on mutual information and then an adaptive genetic algorithm was used. Uğuz (2011) used two-stage process, information gain is used to find important features which were further given as input to PCA and genetic algorithm for evaluation. Hsu et al. (2011) used a hybrid model, where two feature sets are generated using F-score and Information gain separately, which were further given as input to SVM-based wrapper method. Peng et al. (2005), used minimum redundancy maximal relevance criterion for evaluating features, followed by backward and forward wrapper selection. Zheng et al. (2018), used Information gain as the first criteria to filter features followed by support vector machine sequential search algorithm. Lee and Leu (2011), first used genetic algorithm with dynamic parameter (GADP) followed by X2 test for homogeneity analysis. Atallah et al. (2019), weighted features using information gain-based and probabilistic Naive Bayes wrapper feature selector. Modified version of F-score has been embedded in hybrid models by many researchers (Chen and Lin, 2006; Lee, 2009; Xie and Wang, 2011; Jaganathan et al., 2012; Lin et al., 2016; Zhang et al., 2018) to suit feature selection for classifications based on one, two and more classes.

Based on the above analysis, we propose to create our own dataset consisting of easily measurable physical features that reflect the health issues and to apply hybrid model to extract relevant features which can be monitored to prevent PTB complications.

3 Methodology

In this section we will be looking into the different methods which are part of proposed hybrid model along with the details of dataset used for experiments.

3.1 Mother's physical feature dataset

This work focuses on the physical features of mother to identify the promising features that help to predict pregnancy outcome being full-term or preterm birth. To have a better understanding of the domain, various doctors with considerable experience in gynaecological and obstetric fields were contacted, who suggested mother's features with their experience and observations. Based on the doctor's advice and the literature, physical features of mothers, selected for the study are age, height, weight before pregnancy, haemoglobin levels, menstrual cycle post-marriage, menstrual cycle before marriage, time taken to conceive, birth parity, father's age, time taken to conceive, infertility treatment, BMI, IVF, polycystic ovary syndrome (PCOS), thyroid, hypertension, gestational diabetes, viral infection, low amniotic fluid, high amniotic fluid, and no health complication during pregnancy. Thus, mother's physical feature (MPF) dataset consists of a total of 21 features out of which age, weight, height, haemoglobin and BMI are continuous features, whereas others are categorical. Records of 903 women were taken into consideration, out of which 146 had preterm delivery while 757 had full-term delivery (Deshpande and Ragha, 2021). Records were taken for the duration of 21 months, from February 2018 to September 2019. Women who delivered babies during this time at D.Y. Patil Hospital, located in Mumbai Metropolitan Region, were interviewed by medical personnel.

3.2 Pre-processing

Dataset has been normalised before being used for preterm prediction using classification algorithm. All the features were scaled down to a common range between 0 to 1, to avoid redundant data and inconsistency within the database tables (Kumar and Azad, 2017). Normalisation is done, such that for each record, the sum of the square of normalised values is equal to one. For the feature selection process, continuous features are converted into categorical data. Decision tree is found to be biased towards features with more categories (Fang et al., 2017). To avoid this issue, all the features under consideration are categorised into four or five levels to maintain symmetry in terms of unique values under each feature. Table 1, shows the values falling under each category (level) of continuous features. Range for each level is decided as per discussion with doctors. We have tested the performance of these categories empirically, using the proposed model.

Table 1 Levels defined for continuous features

Level	*Age of mother (in yrs.)*	*Height (in cm)*	*BMI*	*Haemoglobin (in gm/dl)*	*Age of father (in yrs.)*	*Weight mother (in kg)*
1	<=20	<=152	<=18.5	<=8	<=25	<=45
2	21–25	153–157	18.51–24.9	8.1–9.9	26–30	46–55
3	26–30	158–162	25–29.9	10–12.5	31–35	56–65
4	31–35	163–167	>30	>12.5	36–40	>65
5	>35	>167	-	-	>40	-

3.3 Feature selection methods

Through the survey we observed that the classification outcome will be more relevant if a feature exhibits higher variation with respect to its values falling under different dataset outcomes and higher mutual information across the other features. Based on this we aim to exploit the advantages of both filter and embedded methods. This work proposes a hybrid method combining F-score, decision tree, and random forest for evaluating the relevance of each feature. Selected methodologies are combined so as to provide a complete analysis of each feature towards efficiency of classification.

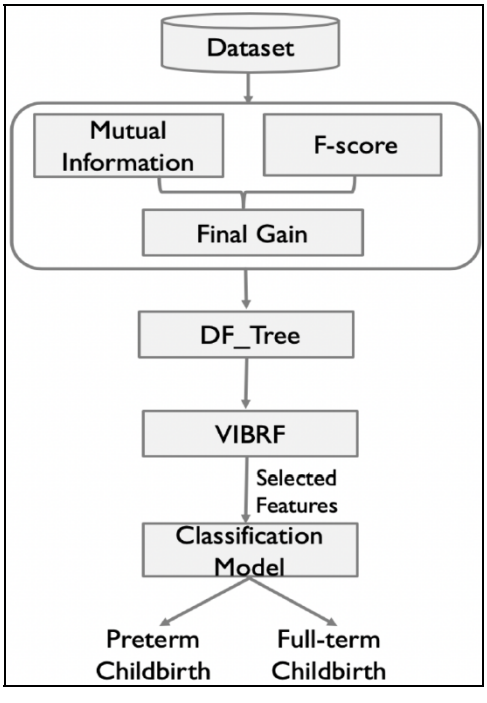
F-score is a simple and effective filter method, which works by finding the variation within the sets individually and the discrimination between the sets from each other. Features having less imbalance within the same class and more contrast across the classes are considered to be more relevant in predicting outcomes. The larger the F-score, the more discriminant the feature will be. However, F-score lacks to find the mutual information between the features (Chen and Lin, 2006), as it evaluates variation within a feature but does not help to identify the variation of the selected feature from the outcome of the dataset. As decision tree is based on mutual information which measures the relationship between the features (Learned-Miller, 2013), the ID3 variant of decision tree is merged with F-score to overcome this lacking. With the combined approach, each feature will be looked upon in terms of gain provided, the more the gain, the more information feature will provide towards the dataset and the feature will be having less distortion (Fang et al., 2017; Hsu et al., 2011). It is further observed that decision tree suffers the over fitting issue (Song and Ying, 2015), solution for the same could be provided using random forest. Random forest works on the principle of ensemble learning. With random forest, randomness while selecting the training data for each decision tree results in training using different combinations from the dataset thus helps to resolve the over-fitting issue of decision tree.

The proposed model merges the F-score, decision tree and random forest methodologies, in such a way that a complete solution could be provided to evaluate each feature looking from different perspectives. We refer this solution as variation and information-based random forest (VIBRF) hybrid feature selection model.

4 Proposed model

Proposed hybrid model as shown in Figure 1, utilises the capabilities of both filter and embedded methods in one framework while evaluating the potential of each feature towards the prediction ability. The algorithm for creating a hybrid decision tree named DF_Tree is explained in Algorithm 1. VIBRF which is formed using multiple DF_Tree is explained in Algorithm 2. VIBRF enables evaluation of each feature using filter and embedded method simultaneously, thus implementation is done without increasing the time complexity of algorithm.

Figure 1 Flowchart of relevant features-based classification model



4.1 DF_Tree

DF_Tree is a modified version of the standard ID3 decision tree that uses F-score and Mutual information to evaluate features. DF_Tree modifies the final gain (FG) equation of decision tree, a fraction (x%) of F-score and a fraction (y%) of information gain is used to find the FG for each feature. Algorithm 1, shows the steps involved to create DF_Tree. After experimenting with various possibilities of x and y on our dataset, we have obtained the best results with x as 0.5 and y as 1, which indicates that the best results are obtained with 50% of F-score weightage and 100% of Information gain weightage. In the proposed DF_Tree algorithm, GN is information gain of N^{th} feature, FGN represents the FG. Sorted vector FG_Vector holds the FG of each of the contributing features. Feature with maximum FG is selected as the next node of the DF_Tree. The process is continued until all features are included in the tree. To implement proposed DF_Tree, the whole dataset is divided into training (70%) and testing(30%) dataset. While designing each DF_Tree, 60% of the training dataset is randomly selected to create DF_Tree. Out of the 21 available features, randomly any 'M' (we have taken 'M' as 15) features are selected to construct a decision tree.

Algorithm 1 Proposed DF_Tree

- **Input:** Dataset with 'X' features
- **Output:** DF_Tree

1 Create sub-dataset 'S' having a set (FSet) of randomly selected 'M' features, where M < X.

2 If dataset 'S' is not empty:

- Calculate the entropy of dataset (S) for split, with two classes namely preterm (PT) and full-term (FT).

$$E_s = -p(PT)\log_2 p(PT) - p(FT)\log_2 p(FT) \tag{1}$$

- For N ← 1 to M

a Calculate entropy of N^{th} feature, with 't' unique values.

$$E_N = \sum_{a=1}^{t} p(a)E(a) \tag{2}$$

b Calculate F-score value of N^{th} feature.

$$F_N = \frac{\left(x_i - x_i\right)^{-(PT)-2} + \left(x_i - x_i\right)^{-(FT)-2}}{\frac{1}{\left(n_{(PT)}-1\right)}\sum_{k=1}^{n_{(PT)}}\left(x_{k,i} - x_i\right)^{-(PT)-(PT)^2} + \frac{1}{\left(n_{(FT)}-1\right)}\sum_{k=1}^{n_{(FT)}}\left(x_{k,i} - x_i\right)^{-(FT)-(FT)^2}} \tag{3}$$

c Calculate information gain of N^{th} feature.

$$G_N = E_s - EN \tag{4}$$

d Calculate final gain of N^{th} feature.

$$FG_N = (x\%)F_N + (y\%)G_N \tag{5}$$

- Prepare a vector (FG_Vector), having FG_1, FG_2, FG_3, … FG_N in descending order.
- Add feature node to DF_Tree considering FG_Vector.

3 Recursive calls to step 2, with the remaining features.

4.2 Variation and information-based random forest

VIBRF has used a novel approach for ranking features through a random forest created out of DF_Tree. Algorithm 2, depicts VIBRF which consists of 600 DF_Trees, each of which is formed using a randomly selected subset of 15 features, across randomly selected data records. Thus each DF_Tree is unique, in terms of training sets and features. VIBRF model considers recall as the measure to decide the efficiency of DF_Tree. Each of the 600 DF_Trees are arranged in descending order by recall value. Top 300 trees are taken into consideration, while the remaining 300 with least recall value are neglected. This way VIBRF looks into the DF_Trees which are formed using a subset of the feature, which helps reach better results. Recall value is selected as evaluation criteria because, being a medical dataset, it is expected that the model should give right results whenever there are chances of preterm complication, i.e., it is important to have right prediction in case of positive (preterm birth) outcome. Feature with maximum FG is considered to be most informative and while forming the DF_Tree it is made the root node, thus root node could be rated as the most relevant feature. Once the best 300 DF_Trees are found, root

node of each of these DF_Tree is taken into consideration and features are voted for being the root node across selected DF_Trees. For getting the relevant features, ten-fold cross-validation process is performed. Average of the votes for each feature being selected as root node across ten executions is taken as weightage of feature. Features with average above a selected threshold are selected as relevant and are given as input to the classification algorithm.

Algorithm 2 VIBRF – feature selection model

- **Input:** 'T' DF_Tree, each with 'M' features formed using, Algorithm 1
- **Output:** Relevant features

1 Create a sorted array 'R_Desc[ ]' holding the recall value for all 'T' DF_Tree in descending order.

2 Select the first 'S' DF_Trees from the array 'R_Desc[ ]'.

3 Extract the root nodes of selected DF_Tree in an array 'Root_Tree[ ]'.

4 Create an array 'A[ ]', of size X, initialise all the values to 0.

5 For N ← 1 to S

```
if (Root_Tree[N] = = X)
   A[X] += 1
```

6 Considering the values of A[ ], all the features above a selected threshold are selected as being relevant towards the classification of dataset.

Based on this model we performed good number of experiments by varying number of DF_Tree contributing towards forest, the results are quite promising and are discussed in the next section.

5 Results and observations

This section presents the classification results with relevant features selected using hybrid model. We experimented using five different classifiers namely, random forest, decision tree, Gaussian NB, KNN (neighbour = 3), and SVM (kernel = rbf).

5.1 Testing proposed model on standard datasets

As the PTB standard dataset is not available we validated the performance of proposed hybrid model on two different standard datasets. Experimentation was done on Pima India Diabetes dataset (https://www.kaggle.com/uciml/pima-indians-diabetesdatabase) and Monk's problem dataset (Thrun, 1992). On both the datasets, classification results using selected relevant features, resulted in better results as compared to considering all the available features. Tables 2(a), 2(b), 2(c) and 2(d), shows the results on Monk's dataset and Pima India diabetes, it has been observed that there has been considerable improvement in terms of prediction accuracy as well as AUC value using different classification algorithms which show that proposed hybrid model work well in selecting relevant features for classification.

Table 2 Classification results using all features versus selected relevant features using VIBRF,

(a) Prediction accuracy results on Pima India Diabetes dataset				*(b) AUC results on Pima India Diabetes dataset*			
Classifier	*Prediction accuracy*			*Classifier*	*AUC*		
	All features	*Relevant features*	*Improvement %*		*All features*	*Relevant features*	*Improvement %*
GNB	63.94	71.48	11%	GNB	0.64	0.74	10%
DT	61.78	63.82	3%	DT	0.58	0.61	3%
RF	68.21	70.5	3%	RF	0.70	0.73	3%
KNN	66.16	67.45	2%	KNN	0.65	0.69	4%
SVM	63.46	65.54	3%	SVM	0.69	0.71	2%
(c) Prediction accuracy results on Monk's 1 dataset				*(d) AUC results on Monk's 1 dataset*			
Classifier	*Prediction accuracy*			*Classifier*	*AUC*		
	All features	*Relevant features*	*Improvement %*		*All features*	*Relevant features*	*Improvement %*
GNB	68.46	72.56	6%	GNB	0.77	0.79	2%
DT	91.32	100	9%	DT	0.91	1.00	9%
RF	90.55	100	9%	RF	0.97	1.00	3%
KNN	85.16	100	15%	KNN	0.92	1.00	8%
SVM	85.27	94.74	10%	SVM	0.95	0.99	5%

5.2 Testing on MPF dataset

The MPF dataset with 21 features is given as input to the proposed hybrid model to select relevant features for classification. Figure 2, shows the weightage given to each feature, considering the average of the votes given across ten executions of VIBRF. Out of all the features, 'no health issue during pregnancy', is found to be the most relevant feature, 'high amniotic fluid' and 'IVF', being second and third most relevant. Among all the features, age of mother, age of father, menstrual cycle after marriage, health conditions like gestational diabetes, gastric issue and hyper-tension are the six features that are found to be least relevant and are not given as input to the classification algorithm.

Experiments are performed with two sets of features, firstly with all the available features and secondly with the selected relevant features. Six-fold cross-validation is used to obtain prediction accuracy, and AUC for analysis and comparisons, as shown in Table 3. Tables 3(a) and 3(b), depicts the results for prediction accuracy and AUC respectively across all the five classification algorithms. The last column of Tables 3(a) and 3(b) exhibits comparative analysis, it shows improvement, in terms of percentage for results while using only relevant features against using all available features.

Figure 2 weights given to each feature by taking the average across 10 executions of VIBRF

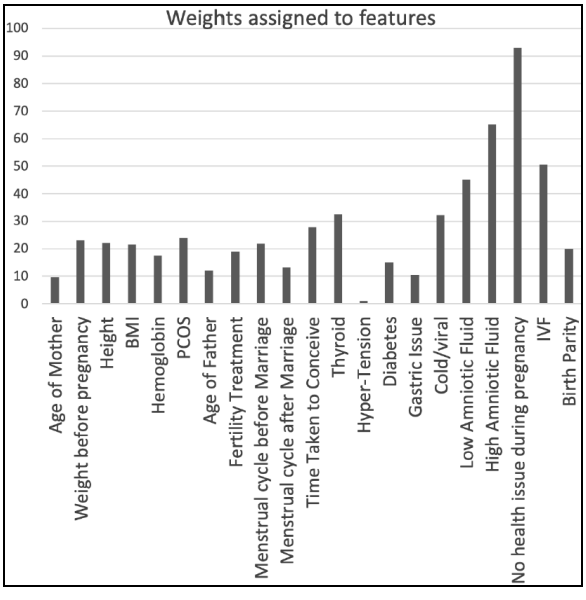


Table 3 Result analysis on pregnancy dataset of various classification algorithms using all features verses relevant features extracted using the proposed model

(a) Prediction accuracy				*(b) AUC value*			
Classifier	*Prediction accuracy*			*Classifier*	*AUC*		
	All features	*Relevant features*	*Improvement %*		*All features*	*Relevant features*	*Improvement %*
GNB	31.77	80.00	60%	GNB	0.58	0.65	10%
DT	72.34	75.40	4%	DT	0.50	0.54	7%
RF	82.33	83.16	1%	RF	0.57	0.58	1%
KNN	80.28	82.66	3%	KNN	0.52	0.55	5%
SVM	81.70	84.19	3%	SVM	0.51	0.54	6%

The results of Table 3, shows that relevant features performed better in terms of AUC value and prediction accuracy. There has been an improvement in results with all the five classification algorithms when considering relevant 15 features. Though there has not been much improvement while using random forest as classifier, but all other classifiers show good improvement in prediction results. Prediction accuracy has shown improvement between 1% to 60 %, using different classifiers. In terms of AUC, there has been increase in value, ranging between 1% to10% using different classifiers. Among all the classifiers used, Gaussian NB has shown a great improvement in prediction results when using relevant features against using all the available features. Pregnancy dataset (WPF) used in the experiment is imbalanced, with majority of the records being full-term. Looking at the execution results, it has been observed that the classifiers are biased towards majority class, which causes low AUC value, but there has been considerable improvement while using only relevant features. We propose to continue the experiments further to work on mentioned issues.

6 Conclusions

The hybrid model based on two novel algorithms DF_Tree and VIBRF is efficient in selecting relevant features that improved the classification accuracy. By removing irrelevant features and by evaluating features using embedded and filter methods simultaneously, the proposed hybrid model helps in achieving better time and space complexity. Out of 21 women's physical features in dataset, 15 features, are selected as being relevant for the prediction of childbirth outcome and observing these features can create consciousness among mothers towards a healthy pregnancy outcome. Based on the results we conclude that the three most contributing features are, 'no health issue during pregnancy', 'high amniotic fluid' and 'IVF'. SVM has given highest prediction accuracy, whereas GNB has given best AUC value. In the imbalanced dataset, AUC is given higher importance over prediction accuracy since we want algorithm to behave rightly with the skewed dataset, for the same reason we would say that GNB has given the best results on MPF dataset using selected relevant features.

7 Future work

In the proposed model, we have worked on the relevance of features, we plan to work further on identifying redundant features using tree-structured hybrid model. Better results could be obtained by working towards the imbalance aspect of the dataset. Only physical features are considered in this research. In the future, we suggest exploring features like stress levels, lifestyle, social features on pregnancy outcomes.

References

Allen, J., Kildea, S. and Stapleton, H. (2016) 'How optimal caseload midwifery can modify predictors for preterm birth in young women: integrated findings from a mixed methods study', *Midwifery*, Vol. 41, pp.30–38.

Atallah, D.M., Badawy, M., El-Sayed, A. and Ghoneim, M.A. (2019) 'Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier', *Multimedia Tools and Applications*, Vol. 78, No. 14, pp.20383–20407.

Catley, C., Frize, M., Walker, R.C. and Petriu, D.C. (2006) 'Predicting high-risk preterm birth using artificial neural networks', *IEEE Transactions on Information Technology in Biomedicine*, Vol. 10, No. 3, pp.540–549.

Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers & Electrical Engineering*, Vol. 40, No. 1, pp.16–28.

Chen, Y.W. and Lin, C.J. (2006) 'Combining SVMs with various feature selection strategies', in *Feature Extraction*, pp.315–324, Springer, Berlin, Heidelberg.

Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G. (2015) 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)', *Annals of Internal Medicine*, Vol. 162, No. 10, pp.735–736.

Deshpande, H. and Ragha, L. (2021) 'Mother's significant feature (MSF) dataset', *IEEE Dataport*, 22 April 22, doi: https://dx.doi.org/10.21227/kq5k-b784.

Fang, L., Jiang, H. and Cui, S. (2017) 'An improved decision tree algorithm based on mutual information', in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNCFSKD)*, IEEE, July, pp.1615–1620.

Grzymala-Busse, J.W. and Woolery, L.K. (1994) 'Improving prediction of preterm birth using a new classification scheme and rule induction', in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, p.730.

Hsu, H.H., Hsieh, C.W. and Lu, M.D. (2011) 'Hybrid feature selection by combining filters and wrappers', *Expert Systems with Applications*, Vol. 38, No. 7, pp.8144–8150.

Idowu, I.O., Fergus, P., Hussain, A., Dobbins, C. and Askar, H. (2014) 'Advance artificial neural network classification techniques using EHG for detecting preterm births', in *2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems*, IEEE, July, pp.95–100.

Jaganathan, P., Rajkumar, N. and Kuppuchamy, R. (2012) 'A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification', *International Journal of Machine Learning and Computing*, Vol. 2, No. 6, p.741.

Jiang, Y., Liu, X., Yan, G. and Xiao, J. (2017) 'Modified binary cuckoo search for feature selection: a hybrid filter-wrapper approach', in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, IEEE, December, pp.488–491.

Khalid, S., Khalil, T. and Nasreen, S. (2014) 'A survey of feature selection and feature extraction techniques in machine learning', in *2014 Science and Information Conference*, IEEE, August, pp.372–378.

Kumar, K. and Azad, S.K. (2017) 'Database normalization design pattern', in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, IEEE, October, pp.318–322.

Learned-Miller, E.G. (2013) *Entropy and Mutual Information*, Department of Computer Science, University of Massachusetts, Amherst.

Lee, C.P. and Leu, Y. (2011) 'A novel hybrid feature selection method for microarray data analysis', *Applied Soft Computing*, Vol. 11, No. 1, pp.208–213.

Lee, K.A., Chang, M.H., Park, M.H., Park, H., Ha, E.H., Park, E.A. and Kim, Y.J. (2011) 'A model for prediction of spontaneous preterm birth in asymptomatic women', *Journal of Women's Health*, Vol. 20, No. 12, pp.1825–1831.

Lee, M.C. (2009) 'Using support vector machine with a hybrid feature selection method to the stock trend prediction', *Expert Systems with Applications*, Vol. 36, No. 8, pp.10896–10904.

Lin, X., Huangfu, W., Wang, F., Liu, L. and Long, K. (2016) 'A breast cancer risk classification model based on the features selected by novel F-score index for the imbalanced multi-feature dataset', in *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, IEEE, October, pp.198–203.

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y. and Gao, Z. (2017) 'A hybrid feature selection algorithm for gene expression data classification', *Neurocomputing*, Vol. 256, pp.56–62.

Morken, N.H., Källen, K. and Jacobsson, B. (2014) 'Predicting risk of spontaneous preterm delivery in women with a singleton pregnancy', *Paediatric and Perinatal Epidemiology*, Vol. 28, No. 1, pp.11–22.

Pari, R., Sandhya, M. and Sankar, S. (2017) 'Risk factors based classification for accurate prediction of the preterm birth', in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, November, pp.394–399.

Peng, H., Long, F. and Ding, C. (2005) 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp.1226–1238.

Robinson, C.J., Hill, E.G., Alanis, M.C., Chang, E.Y., Johnson, D.D. and Almeida, J.S. (2010) 'Examining the effect of maternal obesity on outcome of labor induction in patients with preeclampsia', *Hypertension in Pregnancy*, Vol. 29, No. 4, pp.446–456.

Sheikhpour, R., Sarram, M.A., Gharaghani, S. and Chahooki, M.A.Z. (2017) 'A survey on semi-supervised feature selection methods', *Pattern Recognition*, Vol. 64, pp.141–158.

Son, M. and Miller, E.S. (2017) 'Predicting preterm birth: cervical length and fetal fibronectin', in *Seminars in Perinatology*, WB Saunders, December, Vol. 41, No. 8, pp.445–451.

Song, Y.Y. and Ying, L.U. (2015) 'Decision tree methods: applications for classification and prediction', *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, p.130.

Thomas, J. and Kulanthaivel, G. (2016) 'Rule minimization in predicting the preterm birth classification using competitive co evolution', *Indian Journal of Science and Technology*, Vol. 9, p.10.

Thrun, S. (1992) *UCI Machine Learning Repository*, School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213, USA [online] https://archive.ics.uci.edu/ml/datasets/MONK's+Problems (accessed August 2020).

Uğuz, H. (2011) 'A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm', *Knowledge-Based Systems*, Vol. 24, No. 7, pp.1024–1032.

Vogel, I., Grove, J., Thorsen, P., Moestrup, S.K., Uldbjerg, N. and Møller, H.J. (2005) 'Preterm delivery predicted by soluble CD163 and CRP in women with symptoms of preterm delivery', *BJOG: An International Journal of Obstetrics & Gynaecology*, Vol. 112, No. 6, pp.737–742.

Von Dadelszen, P., Payne, B., Li, J., Ansermino, J.M., Pipkin, F.B., Côté, A.M., Douglas, M.J., Gruslin, A., Hutcheon, J.A., Joseph, K.S. and Kyle, P.M. (2011) 'Prediction of adverse maternal outcomes in pre-eclampsia: development and validation of the full PIERS model', *The Lancet*, Vol. 377, No. 9761, pp.219–227.

Xie, J. and Wang, C. (2011) 'Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases', *Expert Systems with Applications*, Vol. 38, No. 5, pp.5809–5815.

Zhang, X., Shi, Z., Liu, X. and Li, X. (2018) 'A hybrid feature selection algorithm for classification unbalanced data processsing', in *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, IEEE, August, pp.269–275.

Zheng, W., Zhu, X., Wen, G., Zhu, Y., Yu, H. and Gan, J. (2018) 'Unsupervised feature selection by self-paced learning regularization', *Pattern Recognition Letters*, Vol. 132, pp.4–11

Websites

https://www.kaggle.com/uciml/pima-indians-diabetes-database.

https://www.who.int/news-room/fact-sheets/detail/preterm-birth.