Archives of Gynecology and Obstetrics (2019) 300:1565–1582
https://doi.org/10.1007/s00404-019-05325-3

MATERNAL-FETAL MEDICINE

Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features

Toktam Khatibi[1,2] · Naghme Kheyrikoochaksarayee[3] · Mohammad Mehdi Sepehri[1,2]

Received: 5 February 2019 / Accepted: 9 October 2019 / Published online: 24 October 2019

Abstract

Purpose High rate of preterm birth (birth before 37 weeks of gestation) in the world, its negative outcomes for pregnant women and newborns necessitate to predict preterm birth and identify its main risk factors. Premature deliveries have been divided into provider-initiated (with medical intervention for early terminating the pregnancy) and spontaneous preterm birth (without any intervention) categories in the previous studies. The main aim of this study is proposing methods for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features.

Methods Data from national databank of Maternal and neonatal records (IMAN registry) is used in the study. The collected data have information about more than 1,400,000 deliveries with 112 features. Among them, 116,080 preterm births have occurred (from which 11,799 and 104,281 cases belong to provider-initiated preterm birth and spontaneous premature delivery, respectively). The data can be considered as big data due to its large number of data records, large number of the features and unbalanced distribution of the data between three classes of term, provider-initiated and spontaneous preterm birth. Therefore, we need to analyze data based on big data algorithms. In this paper, Map Reduce-based machine learning algorithms named MR-PB-PFS are proposed for this purpose. Map phase use parallel feature selection and classification methods to score the features. Reduce phase aggregates the feature scores obtained in Map phase and assign final scores to the features. Moreover, the classifiers trained in Map phase are aggregated based on two different ensemble rules in Reduce phase.

Results Experimental results show that the best performance of the proposed models for preterm birth prediction is accuracy of 81% and the area under the receiver operating characteristic curve (AUC) of 68%. Top features for predicting term, provider-initiated preterm and spontaneous premature birth identified in this study are having pregnancy risk factors, having gestational diabetes, having cardiovascular disease, maternal underlying diseases, and mother age. Chronic blood pressure is a high rank feature for preterm birth prediction and father nationality is highly important for discriminating provider-initiated from spontaneous premature delivery.

Conclusions Identifying the pregnant women with high risk of spontaneous premature or therapeutic preterm delivery in our proposed model can help them to: (1) reduce the probability of premature birth with monitoring and management of the main risk factors and/or (2) educate them to care from the premature newborn. Management and monitoring top features discriminating term, provider-initiated preterm and spontaneous premature birth or their associated factors can reduce preterm labor or its negative outcomes.

Keywords Preterm birth prediction · Big data · Map-reduce · Feature selection · Ensemble classifier

Introduction

Preterm birth occurs before 37 weeks of gestation [1]. Rate of preterm labor ranges from 5 to 18% in different countries [2]. Every year, about 15 million premature delivery is occurred all over the world [1].

The preterm birth has substantial and increasing burden [3]. Premature delivery is the leading cause of infant

✉ Toktam Khatibi
toktamk.khatibi@modares.ac.ir; toktamk@gmail.com; khatibi.t@iums.ac.ir

Extended author information available on the last page of the article

mortality during the first 28 days of life [4] and the first cause of child death in their first 5 years of their life [4]. Premature infants have the higher risk of later-life morbidities such as neurological disability and learning difficulties [5].

Early detection of patients at risk of preterm birth is an avenue to reduce the incidence of prematurity [6] and the occurrence of its related perinatal morbidity and mortality [7]. Therefore, this study aims at early detection and prediction of preterm birth.

Some previous studies have considered the preterm labor prediction based on (1) different input variables and (2) different analytical methods.

Previous researches have investigated the association between preterm birth and different features such as demographic properties (e.g. BMI and mother age), maternal anatomical characteristics (e.g. cervical length) [8], social or environmental risk factors (e.g. air pollutants) [3, 4], behavioral factors (mother's lifestyle and nutrition habits), clinical history (e.g. maternal underlying diseases) [9], fetal describing features [10] and some other feature categories. In this study, we have considered a combination of demographic properties, clinical history, fetal abnormalities and other maternal and fetal describing features.

Most of the analytical methods have been used in the previous studies for premature delivery prediction can be divided into statistical tests and/or machine learning models. Machine learning models which are a subset of artificial intelligence techniques and sometimes have statistical foundations are trained on the data and can extract hidden patterns from data. They are very popular and common for data analytical purposes. Therefore, we focus on the previous studies having used machine learning models for premature delivery prediction.

Weber et al. have used different machine learning classifiers such as random forests, K-nearest neighbors, generalized additive models, lasso regression, ridge regression and elastic nets for predicting spontaneous preterm birth. They have conducted a cohort study considering demographic, race-ethnicity and maternal features. Fivefold cross-validated area under the receiver operating characteristic curve (AUC) has been used as evaluation metric for comparing the performance of different classifiers. Weber et al. have concluded that different classifiers have similar performance with no sensitivity to missing value imputation method and they have selected the regression models for classification (AUC = 0.67, sensitivity = 0.62 and specificity = 0.65) [11].

Mailath-Pokonry et al. have exploited multivariate logistic regression models to predict delivery within 48 h after admission and delivery before 32 weeks. Their data sample includes 617 Austrian and 737 Toulouse patients. They have considered the predictive features such as mother age, gestational age at the time of admission, history of preterm delivery or late miscarriage, clinical features, vaginal bleeding, functional cervical length and Preterm Premature Rupture of Membranes (PPROM) [12].

Son and Miller have analyzed cervical length and fetal fibronectin to predict preterm delivery. For this purpose, they have tried to find the best cut points of cervical length and fibronectin for achieving better prediction performance [8].

Elaveyini et al. have used feed-forward backpropagation neural networks for preterm birth prediction. Their studied sample was women with vaginal bleeding in the first trimester. The features have been considered in their study were mother age, gestational age when the bleeding occurred, duration of bleeding, amount of bleeding, number of episodes, presence of hematoma and placental position [13].

The features considered in the previous studies are different from the considered features in this paper. Moreover, to the best of our knowledge, machine learning methods for big data analytics have not been exploited for preterm delivery prediction in the previous studies. The first reason for it is that the datasets have not big data characteristics and the common data analytical methods can be used for mining patterns from them. Our collected data which is registered in IMAN registry has big data characteristics according to its large number of records (data rows) and features (data columns). Therefore, we need to propose algorithms based on big data analytical frameworks for predicting preterm labor in our collected data.

At least 43 different definitions of big data exist in the literature but generally, big data must have some of the big data characteristics such as large volume, variety of feature types, velocity of data generation and big value corresponding to analyze big data [14]. For processing big data, algorithms must be scaled up. For this purpose, different strategies have been proposed in the previous studies which are based on subsampling, divide and conquer methods, algorithm weakening and online updating [15]. Subsampling methods may remove some representative cases and lead to results not generalizable to all samples. Divide and conquer methods split the problem into smaller problems and finally aggregates the results. Algorithm weakening tries to reduce the computational time and may lead to worse accuracy. Online updating methods have sequential steps with low computational time.

Among divide and conquer methods, a popular approach is Map-Reduce programming paradigm [16]. It consists of two steps: Map phase and reduce phase. In Map phase, data is split into smaller data chunks. Each data chunk is restored and processed on a separate computing node. The results generated by independent Map jobs are aggregated in reduce phase to generate the final result [15]. For scaling up the feature selection and classification methods based on Map-Reduce, previous studies have proposed different methods [15–19].

Common and popular methods can be used for classification and feature selection are decision trees [20], random forests [21, 22] and support vector machines [23, 24]. More details about decision trees (DT), random forests (RF) and support vector machines (SVM) are stated in Appendix A. These models have been proposed and used for small enough datasets which can be restored in a single computation node. For training them on big data, they must be scaled up.

Previous studies have proposed some different methods for scaling up decision trees, SVMs and random forests. Collobert et al. have proposed a mixture model of SVMs combining SVMs with multi-layer perceptron neural network (MLP). This model has been presented for classifying large scale datasets. Each SVM has been trained on a small subset of data [25]. But, for training MLP, it is required to feed all training data to MLP which cannot be restored in a single node memory and the training time is increased. Some researchers have proposed a parallel SVM method based on Map-Reduce. SVM has been trained for each data chunk in Map phase and non-support vector records are removed in reduce phase. Map and reduce phases have been executed iteratively. The remaining data has been split again and Map and reduce processes have been repeated till the volume of the remaining data will be small enough that could be processed in one computational node. Therefore, one SVM model has been trained on the remaining data [17–19]. Sun and Fox proposed method is very slow because of its repetitive nature.

Genuer et al. have proposed serial RF and parallel RF models for big data feature selection. In serial RF, an individual decision tree has been trained on a separate data chunk. The trained decision trees have been considered as the base classifiers for the final RF ensemble method. Parallel RF have trained separate RFs on different data chunks and aggregate their scores [15].

In this paper, three different methods for parallel feature selection and classification based on Map-Reduce is used for preterm prediction. We have used parallel RF the same as parallel RF being proposed by Genuer et al. and parallel decision trees which have similarities with serial RF being proposed by Genuer et al. Moreover, we have used parallel SVM which is different from the model proposed by Sun and Fox. In our parallel SVM, several SVMs are trained on data chunks. Moreover, our ensemble classifiers use different combinations of base classifiers and two different ensemble rules. It leads to 14 different ensemble classifiers which will be compared based on their Accuracy and AUC in this study.

The findings of this study can be used in two different categories:

- A model is proposed for preterm birth prediction based on top selected features: the model can be used as a decision support system to help the doctor for predicting premature delivery.
- The features are ranked based on their predictive ability for discriminating preterm birth from term birth in three different ways: rate of the preterm birth can be reduced by monitoring and management of top ranked features.

The rest of the paper is organized as follows: Sect. 2 states data description with some useful figures and exploratory analytics. In Sect. 3 the methodology of this research is described and then the results and findings are reported in Sect. 4. Section 5 discusses about the main findings of this study. Concluding remarks are presented in Sect. 6.

Data description

Iranian Maternal and neonatal records (IMAN registry) is one of the greatest data centers for monitoring maternal and neonatal health. Nearly all live and dead births are included in this registry in and out of hospitals across the country. Child delivery information for births occurring in childbirth facilities, homes and other places are recorded in IMAN [26].

In this paper, birth occurring at the gestational age of 28 weeks or more is considered. Data is collected from Iranian Maternal and neonatal records (IMAN registry) for childbirth from 2016/04/01 to 2017/01/01 including 1,431,597 birth deliveries. There were 116,080 preterm birth and 1,315,517 term labor. Premature delivery is about 8.11% of all childbirth data records in our studied dataset. From which 11,799 and 104,281 cases belong to provider-initiated preterm birth and spontaneous premature delivery, respectively. 112 different features are registered for each case as listed in Table 1.

As shown in Table 1, feature code, describing maternal or prenatal, feature name and feature type are listed for all features considered in this study.

Some elementary exploratory data analysis is performed on the features describing data. Geographical distribution of childbirth among different cities is listed in Table 2.

Childbirth is registered for about 430 cities. The maximum number of childbirth is occurred in Tehran. Second top city is Mashhad and the third and the fourth top cities based on childbirth number are Isfahan and Karaj. These cities are the most populated cities in our country, too. Therefore, it is expected that the number of childbirth in these cities is more than the other cities in the country.

The maximum number of childbirth is associated with mothers having Diploma but not having a college degree. The rate of preterm birth has near values for mothers with different education levels.

Table 1 List of the features registered in our maternal and neonatal dataset and analyzed in this study

F. code	M/I	F.name	F. type	F. code	M/I	F.name	F. type
F1	M	Having pregnancy risk factors	Binary	F2	M	Having gestational diabetes	Binary
F3	M	Having cardiovascular diseases	Binary	F4	M	Maternal underlying disease	Binary
F5	M	Chronic blood pressure	Binary	F6	I	Prenatal abnormalities	Binary
F7	M	HIV[a]+	Binary	F8	M	Preeclampsia risk factors	Binary
F9	I	IUGR[b]	Binary	F10	M	Infant mortality in previous pregnancies	Binary
F11	M	Still birth in previous pregnancies	Binary	F12	M	Type-1 or type-2 diabetes	Binary
F13	M	Hepatitis B	Binary	F14	I	Cleft lip with or without cleft palate	Binary
F15	I	Placental abruption	Binary	F16	I	Meconium-stained amniotic fluid	Binary
F17	I	Irregular fetal heartbeat	Binary	F18	I	Early rupture of the amniotic sac	Binary
F19	I	Placenta accreta	Binary	F20	M	IVF[c]	Binary
F21	M	Number of pregnancies	Numeric	F22	M	Number of previous deliveries	Numeric
F23	M	Miscarriage number	Numeric	F24	M	Prenatal gender	Binary
F25	M	Education level	Ordinal	F26	M	Consanguinity with spouse	Binary
F27	M	Address	Nominal	F28	M	Mother age	Numeric
F29	M	Chorioamnionitis	Binary	F30	M	Alcohol or drug addiction	Binary
F31	M	Smoking	Binary	F32	M	Province	Nominal
F33	M	City	Nominal	F34	I	Prenatal other abnormalities	Binary
F35	M	Father nationality	Nominal	F36	M	Mother nationality	Nominal
F37	M	VDRL[d]+	Binary	F38	I	Pulmonary stenosis	Binary
F39	I	Transposition large vessels	Binary	F40	I	Truncus arteriosus	Binary
F41	I	Aortic stenosis	Binary	F42	I	ASD[e] abnormality	Binary
F43	I	VSD[f] abnormality	Binary	F44	I	Hypoplastic left heart syndrome	Binary
F45	I	Tetralogy of fallot	Binary	F46	I	Single ventricle	Binary
F47	I	Coarctation of the aorta	Binary	F48	I	Multiple cardiac anomalies	Binary
F49	I	Endocardial cushion defect	Binary	F50	I	Other cardiovascular abnormalities	Binary
F51	I	Skin tag abnormality	Binary	F52	I	Epidermolysis bullosa	Binary
F53	I	Hyperkeratosis	Binary	F54	I	Anophthalmos	Binary
F55	I	Other skin abnormalities	Binary	F56	I	Unspecified anophthalmos/microphthalmos	Binary
F57	I	Microphthalmos	Binary	F58	I	Microtia	Binary
F59	I	Anotia	Binary	F60	I	Hyperthelorism	Binary
F61	I	Unspecified anotia microtia	Binary	F62	I	Glaucoma	Binary
F63	I	Cataract	Binary	F64	I	Polydactyly	Binary
F65	I	Reducing limb	Binary	F66	I	Club Foot	Binary
F67	I	Syndactyly	Binary	F68	I	Achondroplasia	Binary
F69	I	Osteogeneses imperfectal	Binary	F70	I	Renalectopy	Binary
F71	I	Renal agenesis disgenesis	Binary	F72	I	Ambiguous genitalia	Binary
F73	I	Hydronephrosis	Binary	F74	I	Multicystic kidney	Binary
F75	I	Posterior uretral valve	Binary	F76	I	Epispadias	Binary
F77	I	Hypospadias	Binary	F78	I	Omphalocele	Binary
F79	I	Bladder exstrophy	Binary	F80	I	Esophageal atresia/stenosis with or without fistula	Binary
F81	I	Gastroschisis	Binary	F82	I	Anorectal atresia/stenosis	Binary
F83	I	Small intestine atresia/stenosis	Binary	F84	I	Hirschsprung	Binary
F85	I	Malrotation of gut	Binary	F86	I	Eventration of diaphragm	Binary
F87	I	Congenital diaphragmatic hernia	Binary	F88	I	Cystic adenomatous malformation	Binary
F89	I	Choanal atresia bilateral	Binary	F90	I	Bronchogenic cysts	Binary
F91	I	Sequestration	Binary	F92	I	Anencephaly	Binary
F93	I	Laryngeal atresia, stenosis, web	Binary	F94	I	Encephalocele	Binary
F95	I	Spina bifida	Binary	F96	I	Microcephaly	Binary
F97	I	Meningomyelocele	Binary	F98	I	Hydrocephaly	Binary
F99	I	Holoprosencephaly	Binary	F100	I	Arnold–Chiari syndrome	Binary

Table 1 (continued)

F. code	M/I	F.name	F. type	F. code	M/I	F.name	F. type
F101	I	Dandy–Walker syndrome	Binary	F102	I	Trisomy 18	Binary
F103	I	Trisomy 21 (Down's syndrome)	Binary	F104	I	Prune belly sequence	Binary
F105	I	Trisomy 13	Binary	F106	I	CHARGE association	Binary
F107	I	VACTERAL syndrome	Binary	F108	I	Turner syndrome	Binary
F109	I	Pier Robin syndrome	Binary	F110	M	Number of pregnancies with IVF	Binary
F111	M	Preterm birth (target variable 1)	Binary	F112	M	Provider-initiated birth (target variable 2)	Binary

F. code feature code, *M/I* describing mother or prenatal, *F. name* feature name, *F. type* feature type

[a] Human immunodeficiency virus

[b] Intrauterine growth restriction

[c] In vitro fertilization

[d] Venereal Disease Research Laboratory test

[e] Atrial septal defect

[f] Ventricular septal defect

Table 2 Data distribution among different cities

Criteria	Number of cities
Having less than 1000 birth in our collected data	179 cities
Having less than 2000 birth and more than 1000 birth in our collected data	96 cities
Having less than 3000 birth and more than 2000 birth	45 cities
Having less than 4000 birth and more than 3000 birth	30 cities
Having birth ranges from 4000 to 5000 in our collected data	22 cities
Having birth ranges from 5000 to 6000 in our collected data	12 cities
Having birth ranges from 6000 to 7000 in our collected data	8 cities
Having birth ranges from 7000 to 8000 in our collected data	9 cities
Having birth ranges from 8000 to 9000 in our collected data	3 cities
Having birth more than 9000 in our collected data	26 cities

Highest rate of preterm birth is associated with mothers having doctor of medicine degree (about 10.89% of total childbirth in this group). Mothers can only read and write have the second-highest rate of preterm birth (near 10.60% of total childbirth for them). Total childbirth, term birth and preterm birth for the first group (second group) are, respectively 1976, 1961 and 215 (6492, 5804 and 688). Other educational groups have preterm birth rate of less than 10% of total child birth.

Most of childbirth cases have occurred in Tehran province, Khorasan Razavi and Khuzestan provinces, respectively. But, the rate of preterm birth per province does not show significant differences among many of the studied provinces.

According to the elementary exploratory data analysis, it can be concluded that:

- The most number of childbirth has occurred in Tehran, Mashhad, Isfahan and Karaj. These cities are the most populated cities in the studied country. Therefore, it is obvious that the number of childbirth could be the most.
- Mothers having Doctor of Medicine degree constitute the top group of preterm birth rate. After that, mothers can only read and write has the second rank of preterm birth rate.

Materials and methods

The data classification in this study is performed in two steps including:

- The first step: all cases are classified into term deliveries named as preterm birth named as class C1 (before 37 weeks of gestation) and class C2 (occurred in 37 weeks or more of maternity).
- The second step: the cases classified into class C1 in the first step are categorized into provider-initiated preterm birth named as class PI and spontaneous premature delivery named as class SP.

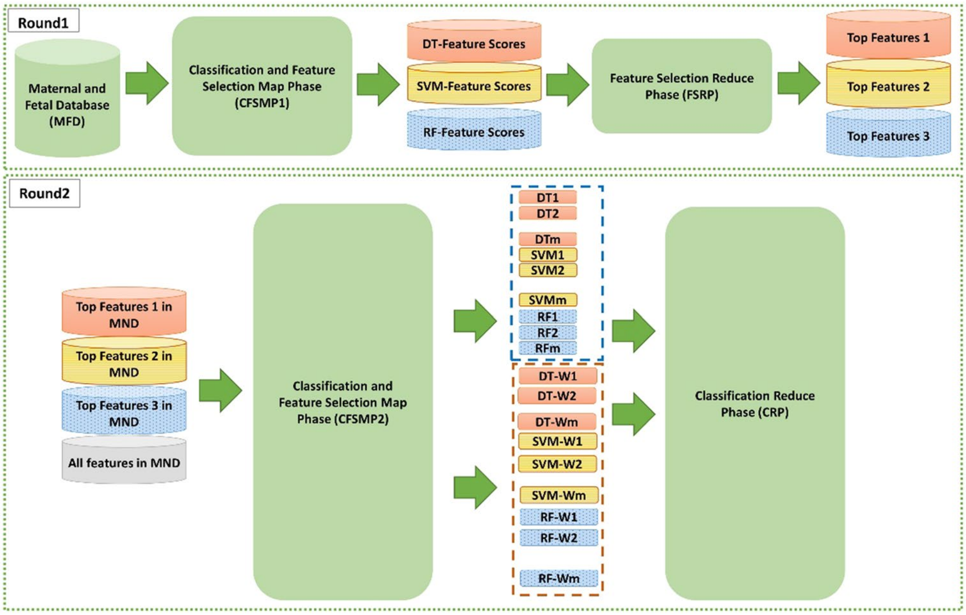


Fig. 1 The schema of Map-reduce based algorithm for data classification and feature selection (MR-PB-PFS)

In this paper, a method is proposed for data classification (into preterm/term birth at first step and provider-initiated and spontaneous premature labor in the second step) and feature selection based on Map-reduce named MR-PB-PFS as shown in Fig. 1.

As illustrated in Fig. 1, the module of classification and feature selection map phase (CFSMP1) is applied to our collected data to score the features based on their ability to discriminate preterm birth from term birth. The outputs of this module are three different scores assigned to each feature by three different ranking methods.

The feature scores generated in the Map phase are fed to the module of feature selection reduce phase (FSRP) as input variables. FSRP aggregates the feature scores and rank the features according to the final feature scores obtained in this phase. The output of FSRP is three lists of top features for predicting preterm birth. Each list includes the top features ranked based on one scoring method (Top Features 1, Top Features 2 and Top Features 3 are high-ranked features based on DT-feature scores, SVM-feature scores and RF-features scores, respectively).

In the second box (round 2), 4 datasets are fed to the module of classification and feature selection Map Phase (CFSMP2) as its input datasets and the module is executed once per each input dataset. Input datasets have different feature sets. The module of CFSMP2 trains the classifiers including decision trees, SVMs and random forests based on its input dataset to predict preterm birth. The classifiers are weighted according to the average value of their accuracy and AUC on the local testing data. Then, the module of classification reduce phase (CRP) is applied to the datasets.

More details about the modules of CFSMP$_i$ ($i=1, 2$), FSRP and CRP are described in the following subsections.

Module of Map phase

The framework of Map phase module (CFSMP$_i$, $i=1, 2$) is depicted in Fig. 2.

As illustrated in Fig. 2, input dataset is partitioned into non-overlapping data chunks. Each data chunk is processed separately in Map phase. The main steps of processing each data chunk include preprocessing data [including (1) missing value imputation with mode for categorical features and median for ordinal and/or numeric features and (2) min–max normalization for numeric features], balanced-sampling from data chunk to make local training set and local testing set, training the feature selection and classification models (including decision trees, support vector machines and random forests) based on the training data, scoring the features using the trained classification and feature selection models, evaluating the trained classifier by applying them to the local testing data and scoring the classifiers based on the average of their Accuracy and AUC when applying them to the local testing data.

The outputs of the module are the feature scores obtained by decision trees (DT-FS$_i$), SVM (SVM-FS$_i$) and RF (RF-FS$_i$) and the classifier weights (DT-W$_i$, SVM-W$_i$ and RF-W$_i$) for each data chunk (Data Chunki) and $1 \leq i \leq$ m.

Modules of reduce phase

Figure 3 denotes the framework of the reduce phase modules.

As illustrated by Fig. 3, the main framework of reduce phase modules consists of two separate modules:

- Module for feature selection reduce phase (FSRP).

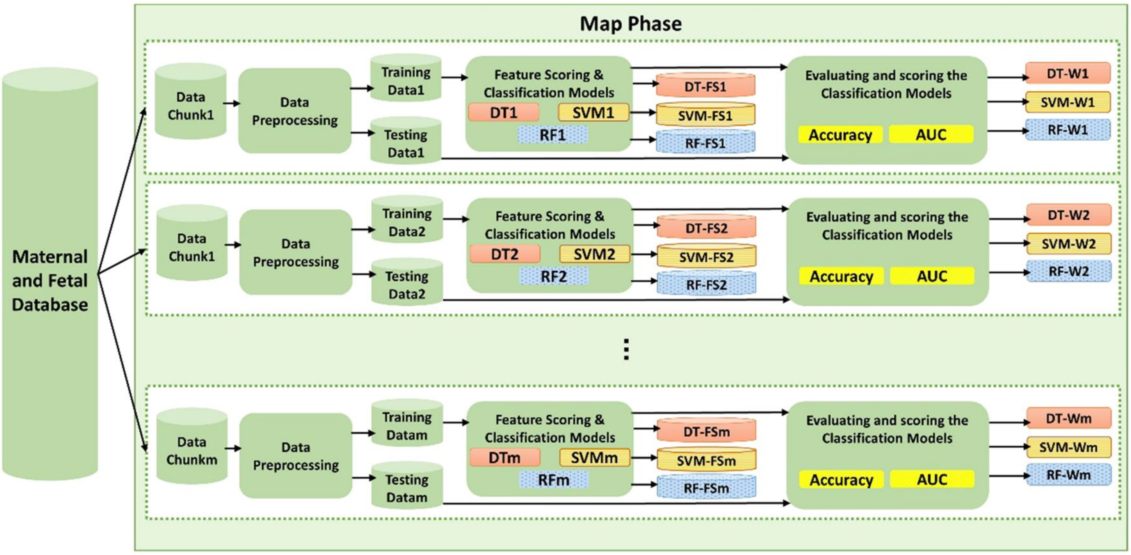


Fig. 2 The schema of the Map phase module (CFSMP$_i$, $i = 1$, 2) for data classification and finding the most important features for it

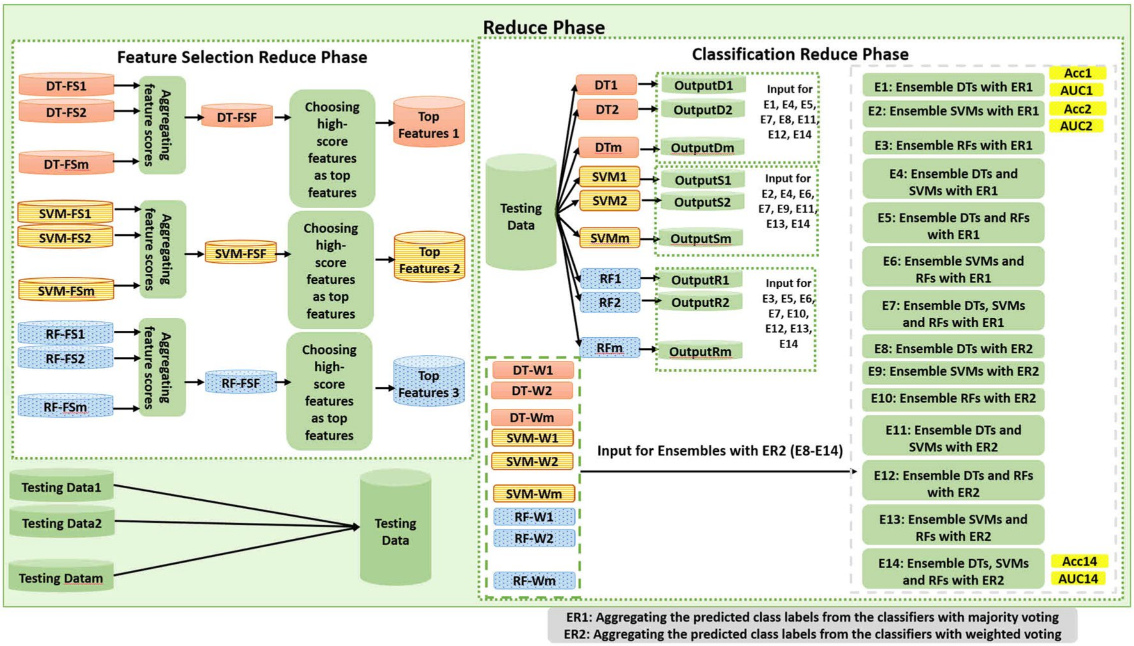

Reduce Phase
Feature Selection Reduce Phase
DT-FS1
DT-FS2
DT-FSm
Aggregating feature scores
DT-FSF
Choosing high-score features as top features
Top Features 1
SVM-FS1
SVM-FS2
SVM-FSm
Aggregating feature scores
SVM-FSF
Choosing high-score features as top features
Top Features 2
RF-FS1
RF-FS2
RF-FSm
Aggregating feature scores
RF-FSF
Choosing high-score features as top features
Top Features 3
Testing Data1
Testing Data2
Testing Datam
Testing Data
Classification Reduce Phase
Testing Data
DT1
DT2
DTm
OutputD1
OutputD2
OutputDm
Input for E1, E4, E5, E7, E8, E11, E12, E14
SVM1
SVM2
SVMm
OutputS1
OutputS2
OutputSm
Input for E2, E4, E6, E7, E9, E11, E13, E14
RF1
RF2
RFm
OutputR1
OutputR2
OutputRm
Input for E3, E5, E6, E7, E10, E12, E13, E14
DT-W1
DT-W2
DT-Wm
SVM-W1
SVM-W2
SVM-Wm
RF-W1
RF-W2
RF-Wm
Input for Ensembles with ER2 (E8-E14)
E1: Ensemble DTs with ER1
E2: Ensemble SVMs with ER1
E3: Ensemble RFs with ER1
E4: Ensemble DTs and SVMs with ER1
E5: Ensemble DTs and RFs with ER1
E6: Ensemble SVMs and RFs with ER1
E7: Ensemble DTs, SVMs and RFs with ER1
E8: Ensemble DTs with ER2
E9: Ensemble SVMs with ER2
E10: Ensemble RFs with ER2
E11: Ensemble DTs and SVMs with ER2
E12: Ensemble DTs and RFs with ER2
E13: Ensemble SVMs and RFs with ER2
E14: Ensemble DTs, SVMs and RFs with ER2
Acc1
AUC1
Acc2
AUC2
Acc14
AUC14
ER1: Aggregating the predicted class labels from the classifiers with majority voting
ER2: Aggregating the predicted class labels from the classifiers with weighted voting

Fig. 3 The schema of the reduce phase for data classification and finding the most important features for it

- Module for classification reduce phase (CRP).

FSRP aggregates the feature scores obtained by CFSMP1. Inputs of FSRP are DT-FS$_i$, SVM-FS$_i$ and RF-FS$_i$ for $1 \le i \le$ m. For calculating the aggregated feature scores (DT-FSF, SVM-FSF and RF-FSF), the summary statistics of the feature scores is computed and the mean value is considered as the final feature score obtained from each classification and feature selection model. For example, DT-FSF for *j*th feature is the mean value of DT-FS$_i$ ($1 \le i \le$ m) for the *j*th feature indicating the final score of the *j*th feature assigned by Decision Trees.

The features with non-zero scores are selected as top features and three lists of top features are constructed. Top Features 1, Top Features 2 and Top Features 3 are corresponding to top features ranked by decision trees, SVM, and RF, respectively. For example, Top Features 1 are the features having non-zero scores in DT-FSF.

CRP calculates the predicted value for each testing data of Map phase with applying each classification and feature selection model trained in Map phase to the testing data. Testing data used in CRP is the union of all local testing data sets of Map Phase. The predicted values are aggregated based on two different ensemble rules. The first ensemble rule (ER1) is the majority voting of the predicted values and the second one (ER2) is the weighted voting of them. The weights contributing to ER2 are the scores assigned to the feature selection and classification models in Map phase (DT-W$_i$, SVM-W$_i$ and RF-W$_i$ for $1 \leq i \leq$ m).

14 different ensemble methods (E_k for $1 \leq k \leq 14$) are constructed and evaluated to find the best ensemble method. The main difference between them is their base classifiers and the ensemble rule used for aggregating the base classifiers. For example, E4 is an ensemble classifier which base classifiers are decision trees and SVMs and uses ensemble rule ER1 for aggregating the base classifiers.

Each ensemble method is applied to the testing data and the accuracy and AUC is calculated.

Findings

According to the mentioned above details, the algorithm consists of Map and reduce phases. Feature scores are calculated in Map phase based on three different models including decision trees, support vector machines and random forests. Table 3 lists the summary statistics of the feature scores assigned by parallel decision trees in Map phase. Since the large number of features, the features with positive value of the 3rd quantile are listed in Table 3.

As listed in Table 3, 5-top features for predicting preterm birth from term labor (spontaneous premature delivery from provider-initiated labor) based on decision tree scores are mother age, education level, city, prenatal gender and number of pregnancies (province, city, education level, mother age and number of pregnancies).

The summary statistics of the feature scores assigned by parallel support vector machines are listed in Table 4.

As listed in Table 4, 5-highest ranked features based on support vector machines include having pregnancy risk factors, having gestational diabetes, having cardiovascular disease, mother age and maternal underlying diseases for preterm birth prediction. Top-5 features discriminating spontaneous from provider-initiated preterm birth are mother age, having pregnancy risk factors, having gestational diabetes, father nationality and having cardiovascular disease.

Parallel random forests have assigned scores to the features in Map phase which are summarized in Table 5.

Now, it is time to compare the predictive power of the features obtained the highest scores based on parallel decision trees (Top Features 1) and SVM (Top Features 2). Table 6 compares the performance of the ensemble classifiers in the reduce phase when applying to different datasets. The last grid in Table 6 considers all features listed in Top Features 1, Top Features 2 and/or Top Features 3.

Performance measures included in Table 6 are the weighted average of the measures obtained from two classification problems solved in this study: classifying data into preterm and term birth and classifying preterm birth data into spontaneous and provider-initiated premature deliveries.

As shown by Table 6, the ensemble classifier E14 outperforms other ensemble classifiers when applied to dataset with all features and/or union of Top Features 1 and Top Features 2. E14 is an ensemble of decision trees, SVMs and RFs aggregating the base classifiers with ER2.

Discussion

Our experimental results show that many of the considered features have assigned zero scores. High sparsity degree of the most features is probably the main reason for having a score of zero. For example, Arteriosus Truncus (a kind of cardiovascular prenatal abnormality) is very sparse and cannot be a predictive variable for preterm birth prediction with high discriminative ability.

A color map based on the weights assigned to the features is shown in Table 7.

Where green, orange, yellow, blue and gray colors are associated with weights greater than 1, in the intervals (0.5, 1], (0.1, 0.5], (0.05, 0.1] and (0, 0.05], respectively. The most important features are colored by green at least once in their associated row. The second (third) most important ones are shown in orange (yellow) once or more and has no green (green or orange) cell in their own row. If all cells in a row in Table 7 are gray or colorless (at least one gray-colored cell must exist), the corresponding feature has least important. If all cells of a row are colorless, the corresponding feature has no importance (these rows are not included in Table 7).

As listed in Table 7, all colored features can be used for early prediction of preterm birth (either provider-initiated or spontaneous premature delivery). Green-colored features have highest importance. They are having pregnancy risk factors, having gestational diabetes, having cardiovascular diseases and mother age (for predicting all considered classes), chronic blood pressure (for preterm birth prediction), father nationality and city (for discriminating provider-initiated from spontaneous preterm birth). All these features but father nationality are selected by at least two of three methods applied in this study.

Table 3 Summary statistics of the feature scores assigned by parallel decision trees (Top Features 1)

F. name	Score mean	Score median	1st quantile	3rd quantile
Feature scores for discriminating term birth from other classes (preterm and spontaneous preterm birth)				
Having pregnancy risk factors	0/03	0/02	0/01	0/03
Having gestational diabetes	0	0	0	0/01
Maternal underlying disease	0	0	0	0/01
Chronic blood pressure	0	0	0	0/01
Preeclampsia risk factors	0/01	0	0	0/02
Early rupture of the amniotic sac	0/01	0/01	0	0/02
Number of pregnancies	0/04	0/04	0/04	0/05
Number of previous deliveries	0/03	0/03	0/02	0/04
Miscarriage number	0/02	0/02	0/01	0/03
Prenatal gender	0/05	0/05	0/04	0/06
Education level	0/12	0/12	0/09	0/13
Consanguinity with spouse	0/03	0/03	0/02	0/04
Address	0/03	0/02	0/02	0/04
Mother age	0/20	0/20	0/18	0/22
Province	0/01	0/01	0	0/02
City	0/08	0/08	0/06	0/10
Feature scores for discriminating spontaneous preterm birth from provider-initiated preterm birth				
Having pregnancy risk factors	0/03	0/03	0/03	0/03
Having gestational diabetes	0/01	0/01	0/01	0/01
Chronic blood pressure	0/01	0/01	0/01	0/01
Prenatal abnormality	0/02	0/02	0/02	0/02
Preeclampsia risk factors	0/01	0/01	0/01	0/01
IUGR	0/01	0/01	0/01	0/01
Placental abruption	0/01	0/01	0/01	0/01
Meconium-stained amniotic fluid	0/01	0/01	0/01	0/01
Irregular fetal heartbeat	0/01	0/01	0/01	0/01
Early rupture of the amniotic sac	0/03	0/03	0/03	0/03
Placenta accreta	0/01	0/01	0/01	0/01
Number of pregnancies	0/06	0/06	0/06	0/06
Number of previous deliveries	0/02	0/02	0/02	0/02
Miscarriage number	0/03	0/03	0/03	0/03
Province	0/11	0/11	0/11	0/11
Prenatal gender	0/04	0/04	0/04	0/04
Education level	0/10	0/10	0/10	0/10
Consanguinity with spouse	0/03	0/03	0/03	0/03
Address	0/02	0/02	0/02	0/02
City	0/10	0/10	0/10	0/10
Mother age	0/08	0/08	0/08	0/08

The first step to prevent preterm birth is the management of pregnancy risk factors, diagnosis of gestational diabetes and maternal underlying diseases and remote monitoring of blood pressure to reduce the risks and complications of preeclampsia. Management and monitoring these features can be helpful for reducing the probability of preterm labor.

As shown in Table 7, province is a predictor of preterm birth lying in the category of second most important features. A province such as Sistan and Baluchestan shows high rate of preterm birth near 10%. It seems that the lack of appropriate pregnancy care centers and lack of regular monitoring the pregnant women can lead to preterm birth in this province. It is suggested a future research to identify the factors increasing preterm birth rate in Sistan and Baluchestan and other similar provinces.

Another feature belonging to second highest important features for discriminating provider-initiated premature birth from spontaneous preterm delivery is the mother's education

Table 4 Summary statistics of feature scores assigned by parallel support vector machines (Top Features 2)

F. name	Score mean	Score median	1st quantile	3rd quantile
Feature scores for discriminating term birth from other classes (preterm and spontaneous preterm birth)				
Having pregnancy risk factors	1/86	1/97	1/81	2
Having gestational diabetes	1/69	1/72	1/50	1/98
Having cardiovascular diseases	1/41	1/32	1/087	1/60
Maternal underlying disease	1/22	1/11	1	1/41
Chronic blood pressure	0/98	1	0/93	1/08
Prenatal abnormalities	0/83	0/90	0/69	1
HIV+	0/76	0/83	0/60	1
Preeclampsia risk factors	0/61	0/64	0/47	0/75
IUGR	0/51	0/52	0/35	0/63
Infant mortality in previous pregnancies	0/41	0/45	0/27	0/57
Still birth in previous pregnancies	0/31	0/27	0/16	0/44
Type-1 or type-2 diabetes	0/23	0/21	0/08	0/31
Hepatitis B	0/18	0/15	0/04	0/28
Cleft lip with or without cleft palate	0/04	0	0	0/05
Placental abruption	0/02	0	0	0/02
Meconium-stained amniotic fluid	0/01	0	0	0/01
Miscarriage number	0	0	0	0
Prenatal gender	0/05	0/02	0	0/05
Education level	0/11	0/06	0/01	0/17
Consanguinity with spouse	0/17	0/11	0/02	0/30
Address	0/25	0/20	0/05	0/37
Mother age	1/34	1/32	1/03	1/62
Chorioamnionitis	0/13	0/11	0/02	0/18
Alcohol or drug addiction	0	0/07	0	0/16
Smoking	0/07	0/03	0	0/09
Province	0/35	0/30	0/10	0/57
City	0/48	0/49	0/18	0/67
Prenatal other abnormalities	0/62	0/64	0/38	0/92
Father nationality	0/74	0/87	0/49	1
VDRL+	0/68	0/72	0/53	0/88
Feature scores for discriminating spontaneous preterm birth from provider-initiated preterm birth				
Having pregnancy risk factors	1/35	1/3	1/15	1/47
Having gestational diabetes	1/18	1/14	1	1/31
Having cardiovascular diseases	1/05	0/98	0/93	1/10
Chronic blood pressure	0/92	0/91	0/77	1
Prenatal abnormalities	0/70	0/70	0/59	0/82
HIV+	0/61	0/65	0/52	0/72
VDRL+	0/49	0/51	0/45	0/62
Preeclampsia risk factors	0/30	0/34	0/19	0/39
IUGR	0/22	0/21	0/17	0/28
Still birth in previous pregnancies	0/14	0/13	0/07	0/17
Type-1 or type-2 diabetes	0/10	0/06	0/01	0/09
Hepatitis B	0/05	0/01	0	0/03
Chorioamnionitis	0/03	0	0	0
Alcohol or drug addiction	0/02	0	0	0
Smoking	0/02	0	0	0
Cleft lip with or without cleft palate	0/01	0	0	0
Placental abruption	0/01	0	0	0
Irregular fetal heartbeat	0	0	0	0/01
Early rupture of the amniotic sac	0/01	0	0	0/01

Table 4 (continued)

F. name	Score mean	Score median	1st quantile	3rd quantile
Placenta accreta	0/01	0	0	0/02
IVF	0/06	0/03	0	0/09
Number of pregnancies	0/07	0/03	0/01	0/10
Number of previous deliveries	0/11	0/07	0/04	0/13
Miscarriage number	0/13	0/08	0/05	0/16
Prenatal gender	0/45	0/40	0/23	0/57
Mother nationality	0/50	0/42	0/27	0/58
Education level	0/59	0/55	0/36	0/75
Consanguinity with spouse	0/69	0/65	0/56	0/80
Address	0/87	0/86	0/71	0/94
Province	0/97	0/92	0/84	0/97
City	1/10	1/02	1	1/08
Father nationality	1/23	1/13	1/07	1/19
Having maternal underlying disease	1/36	1/28	1/22	1/49
Mother age	2/08	2/10	2/07	2/17

level. As shown in, top category includes women having doctor of medicine degree. They know how to take care of themselves. Therefore, the reason of preterm birth high rate in this category is not related to lack of good monitoring and management of the risk factors. A probable cause of preterm birth in this group may be the high-stress level in the workplace or high mother age. Therefore, it is suggested to identify which features cause high rate of preterm birth among pregnant women having doctor of medicine degree.

Another category of education level having high preterm birth rate are women only can read and write. It is probable that they do not know the methods for monitoring and management of preterm risk factors. It is required to supply this group with simple user-friendly software, animation clips and video clips to train them the most important instructions.

Other important predictors of preterm birth are city (a member of the highest important group) and address (belonging to the second most important features). Pregnant women in remote cities and villages has no access to good monitoring services. They need to be equipped with IOT gadgets and smartphone applications for telemedicine and remote monitoring their situations to predict preterm birth.

Predictive features for preterm birth can be divided into two different categories: features that can be controlled and monitored and features that are not manageable. For example, the city and province in which the couple is living cannot be changed in most cases. Therefore, they lie in the second category. A future research direction can be investigating which factors are the main cause of increasing the preterm birth rate in some cities and how these factors can be controlled to reduce the preterm birth rate in these cities.

Features of the first category are having pregnancy risk factors, having gestational diabetes, maternal underlying diseases, chronic blood pressure, preeclampsia risk factors, having cardiovascular diseases, HIV+, IUGR, type-1 or type-2 diabetes, hepatitis B, Chorioamnionitis, alcohol or drug addiction, smoking, VDRL+, early rupture of the amniotic sac, clef lip with or without clef palate, placental abruption, Meconium-stained amniotic fluid and prenatal abnormalities. Some of these features can be avoided such as smoking. Some others should be managed by regular monitoring the patient such as cardiovascular diseases. The preventive strategies and monitoring the patient can reduce the risk of preterm birth according to the previous studies. Some previous studies have considered factors associated with each preterm birth risk factor such as identifying factors associated with placental abruption [27] or predictive factors of gestational diabetes [28]. Therefore, monitoring the patients can be performed by management and monitoring of the factors associated with each risk factor. For example, it has been proposed to prevent obesity with lifestyle interventions in pregnancy and reduce the risk of gestational diabetes [29].

As listed in Table 8, among the top features selected in this study, the impact of many features have been proven in the previous studies. But, to the best of our knowledge, there is no scoring and ranking method in the previous studies to measure and compare their predictive power for preterm birth detection. On the other hand, subclass or superclass of some of the top features for preterm delivery prediction (not exactly the same features) are considered in the previous studies. For example, previous studies have confirmed that the socioeconomic characteristics of pregnant women are associated with preterm birth [35]. The mother's education level is a factor describing the socioeconomic properties.

Previous studies have identified other important predictors of preterm birth which are not considered in this study. For example, Vestgaard et al. have shown that preterm birth

Table 5 Summary statistics of the feature scores assigned with parallel Random Forests (Top Features 3)

F. name	Score mean	Score median	1st quantile	3rd quantile
Feature scores for discriminating term birth from other classes (preterm and spontaneous preterm birth)				
Having pregnancy risk factors	0/20	0/20	0/19	0/21
Having gestational diabetes	0/19	0/18	0/17	0/20
Having cardiovascular diseases	0/12	0/12	0/11	0/13
Maternal underlying disease	0/08	0/08	0/06	0/10
Chronic blood pressure	0/06	0/06	0/05	0/07
Prenatal abnormalities	0/04	0/04	0/04	0/04
HIV+	0/04	0/04	0/03	0/04
Preeclampsia risk factors	0/03	0/03	0/03	0/03
IUGR	0/03	0/03	0/03	0/03
Infant mortality in previous pregnancies	0/03	0/02	0/02	0/03
Still birth in previous pregnancies	0/02	0/02	0/02	0/02
Type-1 or type-2 diabetes	0/02	0/02	0/02	0/02
Hepatitis B	0/02	0/02	0/01	0/02
Chorioamnionitis	0/01	0/01	0/01	0/02
Alcohol or drug addiction	0/01	0/01	0/01	0/02
Smoking	0/01	0/01	0/01	0/01
VDRL+	0/03	0/04	0/03	0/04
Feature scores for discriminating spontaneous preterm birth from provider-initiated preterm birth				
Having pregnancy risk factors	0/15	0/15	0/14	0/17
Having gestational diabetes	0/12	0/11	0/11	0/13
Having cardiovascular diseases	0/10	0/10	0/10	0/10
Chronic blood pressure	0/09	0/09	0/08	0/09
Prenatal abnormality	0/08	0/08	0/08	0/08
HIV+	0/07	0/07	0/06	0/07
Preeclampsia risk factors	0/04	0/04	0/04	0/04
IUGR	0/03	0/03	0/03	0/04
Still birth in previous pregnancies	0/03	0/03	0/03	0/03
Type-1 or type-2 diabetes	0/03	0/03	0/03	0/03
Hepatitis B	0/03	0/03	0/03	0/03
Chorioamnionitis	0/03	0/03	0/03	0/03
Alcohol or drug addiction	0/02	0/02	0/02	0/03
Smoking	0/02	0/02	0/02	0/02
Cleft lip with or without cleft palate	0/02	0/02	0/02	0/02
Placental abruption	0/01	0/01	0/01	0/02
Meconium-stained amniotic fluid	0/01	0/01	0/01	0/01
Irregular heart beats	0/01	0/01	0/01	0/01
Early rupture of the amniotic sac	0/01	0/01	0/01	0/01
Placental accreta	0/01	0/01	0/01	0/01
IVF	0/01	0/01	0/01	0/01
VDRL+	0/05	0/05	0/05	0/06

for women with type-1 diabetes and insufficient vitamin D is twice compared to women with vitamin D insufficiency [37]. Wang et al. have shown that women with low mental health have higher rate of preterm birth [38]. According to Goldenberg et al. study, the strongest predictive feature for preterm birth is the cervical length [39]. It is suggested measuring and adding these predictors such as the amount of vitamin D in the pregnant women and the cervical length to the studied features for increasing the accuracy of preterm birth prediction in the future studies.

This study has a main limitation including:

- The previous studies have shown high predictive power of some features for preterm birth occurrence such as having previous spontaneous preterm birth [40, 41]. But they have not been registered in IMAN registry. There-

Table 6 Comparing the performance of the ensemble classifiers used in the reduce phase

Included Features	Ensemble Classifier	Accuracy	AUC	Included Features	Ensemble Classifier	Accuracy	AUC
All features	E1	69	58	Top Features1	E1	60	50
	E2	69	57		E2	59	50
	E3	70	59		E3	62	51
	E4	68	58		E4	63	52
	E5	72	60		E5	62	53
	E6	74	61		E6	63	52
	E7	77	63		E7	64	53
	E8	73	60		E8	63	51
	E9	72	59		E9	63	53
	E10	73	61		E10	63	51
	E11	74	62		E11	64	53
	E12	75	62		E12	62	51
	E13	76	64		E13	64	54
	E14	81	68		E14	72	59
Top Features2	E1	61	50	Top Features1 + Top Features2 + Top Features3	E1	69	59
	E2	59	48		E2	68	57
	E3	63	51		E3	71	60
	E4	61	50		E4	70	59
	E5	62	51		E5	71	60
	E6	64	53		E6	75	62
	E7	66	55		E7	76	63
	E8	63	51		E8	72	60
	E9	64	52		E9	72	59
	E10	63	51		E10	74	62
	E11	64	52		E11	74	62
	E12	66	54		E12	76	64
	E13	68	56		E13	78	66
	E14	73	61		E14	81	68

fore, they are not included in this study. It is suggested to add them to the considered features in the future researches.

In this paper, a Map-reduce based algorithm is proposed for preterm birth prediction. It can be used as a decision support system (DSS) for gynecologists. The framework of the DSS system (named as PTB-DSS) and its workflow for preterm birth prediction is depicted in Fig. 4.

Text in blue shows the main steps of PTB-DSS workflow containing requesting for data from maternal database, retrieving data, training preterm birth prediction model, taking data describing the important features for a pregnant woman, applying the model to input data, classifying the pregnant woman data to spontaneous premature, provider-initiated preterm or term birth classes, and displaying the predicted class label to the user.

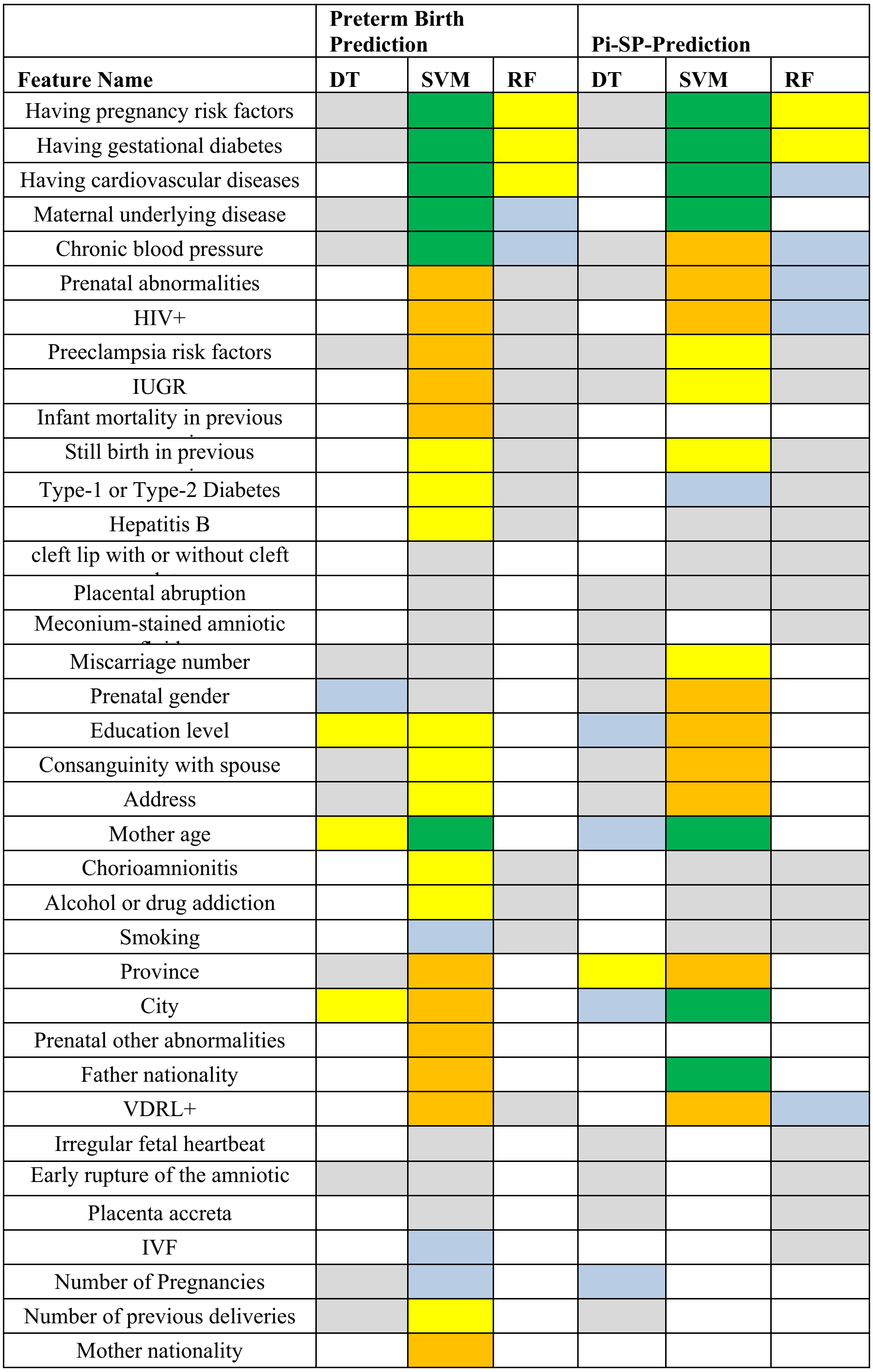
If a pregnant woman is diagnosed at high risk of preterm birth according to PTB-DSS, two different strategies are proposed:

1. She should be monitored via remote telehealth applications and IOT sensors for controlling her preterm birth risk factors and reducing the risk of preterm birth.
2. She can be educated by a smart-phone application for how to care from the premature neonatal and giving her more information about prematurity issues (to be ready if preterm birth occurs).

Conclusion

Machine learning models for big data analytics are proposed in this study for predicting preterm birth and ranking the predictive features. The proposed model can predict premature delivery with an accuracy of 81% and AUC of 68%. According to the findings of this study, high-ranked predictive features are having pregnancy risk factors, having gestational diabetes, mother age, having cardiovascular diseases, maternal underlying diseases, education level, city, prenatal gender and number of pregnancies. For reducing the risk of preterm birth, it is recommended to remotely monitor and manage the high-ranked

Table 7 Color map based on weights assigned to the features by parallel DT, parallel SVM and parallel RF



	Preterm Birth Prediction			Pi-SP-Prediction		
Feature Name	**DT**	**SVM**	**RF**	**DT**	**SVM**	**RF**
Having pregnancy risk factors	▨	🟩	🟨	▨	🟩	🟨
Having gestational diabetes	▨	🟩	🟨	▨	🟩	🟨
Having cardiovascular diseases		🟩	🟨		🟩	🟦
Maternal underlying disease	▨	🟩	🟦		🟩	
Chronic blood pressure	▨	🟩	🟦	▨	🟧	🟦
Prenatal abnormalities		🟧	▨	▨	🟧	🟦
HIV+		🟧	▨		🟧	🟦
Preeclampsia risk factors	▨	🟧	▨	▨	🟨	▨
IUGR		🟧	▨	▨	🟨	▨
Infant mortality in previous		🟧	▨			
Still birth in previous		🟨	▨		🟨	▨
Type-1 or Type-2 Diabetes		🟨	▨		🟦	▨
Hepatitis B		🟨	▨		▨	▨
cleft lip with or without cleft		▨			▨	▨
Placental abruption		▨		▨	▨	▨
Meconium-stained amniotic		▨		▨		▨
Miscarriage number	▨	▨		▨	🟨	
Prenatal gender	🟦	▨		▨	🟧	
Education level	🟨	🟨		🟦	🟧	
Consanguinity with spouse	▨	🟨		▨	🟧	
Address	▨	🟨		▨	🟧	
Mother age	🟨	🟩		🟦	🟩	
Chorioamnionitis		🟨	▨		▨	▨
Alcohol or drug addiction		🟨	▨		▨	▨
Smoking		🟦	▨		▨	▨
Province	▨	🟧		🟨	🟧	
City	🟨	🟧		🟦	🟩	
Prenatal other abnormalities		🟧				
Father nationality		🟧			🟩	
VDRL+		🟧	▨		🟧	🟦
Irregular fetal heartbeat		▨		▨		▨
Early rupture of the amniotic	▨	▨		▨		▨
Placenta accreta		▨		▨		▨
IVF		🟦				▨
Number of Pregnancies	▨	🟦		🟦		
Number of previous deliveries	▨	🟨		▨		
Mother nationality		🟧				

features in regular intervals with smartphone applications and IOT gadgets. Therefore, identifying the factors associated with the top features and finding the ways to monitor and manage them is suggested as the future research opportunities. Moreover, it is required to provide pregnant women with high risk of preterm birth with education about how to care from premature newborns as another future research direction. Some previous studies have considered impact of other features on preterm birth such as

Table 8 Some of top features selected in this study and related findings in the previous studies

Feature name	Related findings in the previous studies
Having gestational diabetes	According to the previous studies, gestational diabetes has impact on preterm birth. Gestational diabetes can lead to hypertension and/or obesity [30] and these two factors can increase the risk of preterm birth. Moreover, gestational diabetes can increase the risk of preterm delivery [31]
Maternal underlying disease	Maternal illness is associated with preterm birth [32]
Chronic blood pressure	Hypertension has impact on preterm birth [31, 32]
Preeclampsia risk factors	Preeclampsia has indirect impact on preterm birth [33]
Miscarriage number	Previous miscarriage can increase the risk of preterm labor [34]
Prenatal gender	Male fetal gender is a risk factor of preterm birth [10]
Education level	Mother social characteristics have impact on preterm birth [35]
Address	Environmental variables such as exposure to certain environmental pollutants have impact on preterm birth [31, 32]
Mother age	Mothers younger than 18 or older than 35 years old have higher risk of preterm birth [31]
Province	Environmental variables have impact on preterm birth [32]
City	Environmental variables have impact on preterm birth [32]
Having cardiovascular diseases	Vascular disease is a risk factor of spontaneous preterm birth [5]
Type-1 or Type-2 Diabetes	Preterm birth for women with type-1 diabetes occurs five times more frequent than healthy women [9]. Diabetes increases the risk of preterm labor [31]
Alcohol or drug addiction	It is known as risk factor of preterm birth in public web sites [31, 36]
Smoking	It is known as risk factor of preterm birth in public web sites [31, 36]
VDRL+	Some vaginal infections such as bacterial vaginosis and Trichomoniasis can lead to premature delivery [31]
Prenatal other abnormalities	Prenatal developmental abnormalities are risk factors of preterm delivery [31]
Father nationality	Black race is a risk factor of spontaneous preterm birth [5]. Ethnicity is associated with preterm labor [31]

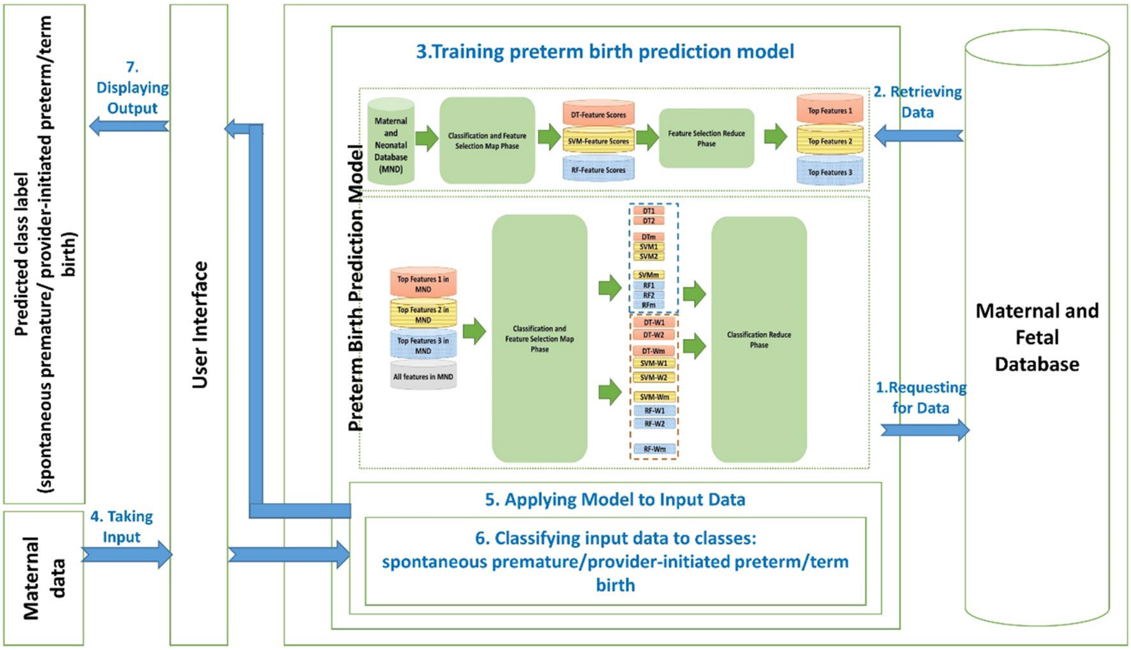


Fig. 4 the framework of the decision support system for preterm birth prediction (PTB-DSS) and its workflow

previous spontaneous preterm delivery, prenatal screening results, different types of infectious diseases, BMI, vitamin D insufficiency and cervical length. It is suggested that adding them individually to the considered features and analyze the augmented data again. Finally, other ensemble rules and ensemble algorithms can be used to improve the performance of the preterm birth prediction in the future studies.

Acknowledgements We thank Dr. Mohammad Heidarzadeh and Health and Medical Education Ministry of Iran—Neonatal Health Office for allowing us to have access to the dataset of IMAN registry and allowing us to do research on this dataset and publish the obtained results.

Author contributions Conceptualization: TK, NK and MMS. Data curation: TK, NK and MMS. Formal analysis: TK, NK and MMS. Funding acquisition: there is no funding. Investigation: TK and NK. Methodology: TK, NK and MMS. Project administration: Khatibi T. Software: TK and NK. Supervision: TK and MMS. Validation: TK, NK and MMS. Visualization: TK, NK. Writing—original draft: TK, NK and MMS. Writing—review & editing: TK and MMS.

Compliance with ethical standards

Conflict of interest The authors declare that there are no conflicts of interest.

Appendix: More details about the feature selection and classification methods (decision trees, support vector machines and random forests)

In this section, we review the methods for classification and feature selection which will be used in this paper. These methods include decision trees, support vector machines (SVM) and random forests (RF).

Decision trees

In 1970s and 1980s, decision tree algorithm (ID3) has been proposed by Quinlan [20]. Further, other extensions of decision tree algorithm such as CART, C4.5 (later C5.0) and J48 have been developed.

For building decision trees, the training set is recursively partitioned in top-down and divide and conquer method. The tree-structure of decision trees made it easy to interpret them. By navigating decision trees, some association rules can be extracted describing the decision tree. The main elements of decision trees are its nodes. The nodes have two different types including internal node (parents) and leaves (terminals). The best discriminating feature is selected to partition data in each internal node. For measuring the discriminative ability of the features, impurity measures such as Gini Index, Information Gain (Entropy) and misclassification error are used as formulated in Eqs. (1–3) [23]:

$$\text{Gini Index}(D) = 1 - \sum_{j=1}^{C} p_{j|D}^{2}, \tag{1}$$

$$\text{Information Gain}(D) = -\left(\sum_{j=1}^{C} p_{j|D} \log_2^{p_{j|D}}\right), \tag{2}$$

$$\text{MisclassificationError}(D) = 1 - \max_{1 \le j \le C}\left(p_{j|D}\right), \tag{3}$$

where D is a node of the decision tree, C is the number of classes and $p_{j|D}$ is the probability that an arbitrary data record in D belongs to class C_j. The lower values of impurity measures denote better splits.

Each leaf node in the tree is labeled by the majority class label of its data. After inducing decision trees based on the training set, they can be applied to the testing sets to predict their class label. For this purpose, each testing record is fed to the induced decision tree as input data. The split criterion on the root node of the tree is assessed for the input data and the valid branch is followed by the input data. This process continues till the input data reaches to a leaf node and its class label is assigned to the input data.

Decision trees may be prunned to avoid from overfitting [23].

Decision trees have good performance for classifying linearly separable data. Each decision tree has assigned importance scores to the features based on their discriminative ability. These scores can be used for ranking the features.

Support vector machines (SVM)

SVM is a classifier trying to find the optimal hyperplane to separate data records of different classes. Default version of SVM is appropriate for classifying linearly separable data. But, it can classify nonlinear separable data using nonlinear kernels.

The equation of the optimal hyperplane for separating different classes is formulated as Eq. (4):

$$\sum_{i=1}^{m} w_i \cdot \mathrm{FP}_i + \sum_{i=1}^{m} b_i = 0, \tag{4}$$

where m is the number of the features, w_i is the weight of the ith feature FP_i in the hyperplane equation and b_i is the ith element of the bias vector.

Each data record is classified by SVM according to Eq. (5):

$$y = \begin{cases} +1 & \text{if} \sum_{i=1}^{m} w_i \cdot \mathrm{FP}_i + \sum_{i=1}^{m} b_i \ge 0 \\ -1 & \text{if} \sum_{i=1}^{m} w_i \cdot \mathrm{FP}_i + \sum_{i=1}^{m} b_i < 0 \end{cases}. \tag{5}$$

SVMs are known as strong classifiers having good performance for many datasets. The weights of the features in the hyperplane equation in SVM can be used for feature ranking.

Random forests (RF)

RF is an ensemble classifier of T independent decision trees. Training set of each decision tree in RF is a bootstrap sample of the original training set. For selecting the split criterion in each node of a decision tree, a random subset of features is considered and the best split is found based on the impurity measures. Decision trees are pruned to avoid over fitting and depth of decision trees is not more than a user-defined maximal depth.

Therefore, RF performance depends on two user-defined parameters including the number of the trees and the tree maximal depth.

Each tree is applied to the out-of-bag (OOB) data to estimate its performance. OOB is a subset of the original training set which does not contribute for training the decision tree.

RF is a classifier that can be used for feature selection, too. It assigns scores to the features as Eq. (6):

$$\text{Importance}\left(F_i\right) = \frac{1}{T}\sum_{t=1}^{T} \text{VI}\left(F_i \cdot T_t\right), \tag{6}$$

where Importance (F_i) is the importance score assigned to F_i with RF. T is the number of decision trees in RF and VI is the variable importance of F_i in the tree T_t in RF.

VI is calculated as Eq. (7):

$$\text{VI}\left(F_i \cdot T_t\right) = \sum_{s \in T_t} \Delta I\left(F_i \cdot T_t\right), \tag{7}$$

where ΔI is the amount of the impurity reduction after partitioning data records based on the splitting feature F_i at node s in the tree T_t. For calculating impurity, the mentioned impurity measures such as Gini Index, Information Gain and misclassification error can be used.

RF have a good performance and its accuracy is comparable with SVMs for many datasets.

References

1. World Health Organization (2018) Preterm Birth. World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/preterm-birth. Accessed Jan 2019
2. Renzo GC, Tosto V, Giardina I (2018) The biological basis and prevention of preterm birth. Best Pract Res Clin Obstet Gynaecol 52:13–22. https://doi.org/10.1016/j.bpobgyn.2018.01.022
3. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L, Lawn JE (2012) National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet 379(9832):2162–2172. https://doi.org/10.1016/S0140-6736(12)60820-4
4. Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, Rudan I, Campbell H, Cibulskis R, Li M, Mathers C, Black RE (2012) Child Health Epidemiology Reference Group of WHO and UNICEF. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet 379:2151–2161. https://doi.org/10.1016/S0140-6736(12)60560-1
5. Goldenberg RL, Culhane JF, Iams JD, Romero R (2008) Epidemiology and causes of preterm birth. Lancet 371:75–84. https://doi.org/10.1016/S0140-6736(08)60074-4
6. Ville Y, Rozenberg P (2018) Predictors of preterm birth. Best Pract Res Clin Obstet Gynaecol 52:23–32. https://doi.org/10.1016/j.bpobgyn.2018.05.002
7. Iams JD (2003) Prediction and early detection of preterm labor. Obstet Gynecol 101(2):402–412. https://doi.org/10.1016/S0029-7844(02)02505-X
8. Son M, Miller ES (2017) Predicting preterm birth: cervical length and fetal fibronectin. Semin Perinatol 41(8):445–451. https://doi.org/10.1053/j.semperi.2017.08.002
9. Colstrup M, Mathiesen ER, Damm P (2013) Pregnancy in women with type 1 diabetes: have the goals of St. Vincent's declaration been met concerning fetal and neonatal complications? J Matern Fetal Neonatal Med 26(17):1682–1686. https://doi.org/10.3109/14767058.2013.794214
10. Peelen MJ, Kazemier BM, Ravelli AC, Ghroot CJ, Post JA, Mol BW, Hajenius PJ, Kok M (2016) Impact of fetal gender on the risk of preterm birth, a national cohort study. Acta Obstetricia et Gynecologica Scandinavica (AOGS) 95(9):1034–1041. https://doi.org/10.1111/aogs.12929
11. Weber A, Darmstadt GL, Gruber S, Foeller ME, Carmichael SL, Stevenson DK, Shaw GM (2018) Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. Ann Epidemiol 28(11):783–789. https://doi.org/10.1016/j.annepidem.2018.08.008
12. Mailath-Pokorny M, Polterauer S, Kohl M, Kueronyai V, Worda K, Heinze G, Langer M (2015) Individualized assessment of preterm birth risk using two modified prediction models. Eur J Obstet Gynecol Reprod Biol 186:42–48. https://doi.org/10.1016/j.ejogrb
13. Elaveyini U, Devi SP, Rao KS (2011) Neural networks prediction of preterm delivery with first trimester bleeding. Arch Gynecol Obstet 283(5):971–979. https://doi.org/10.1007/s00404-010-1469-2
14. Huang T, Lan L, Fang X, An P, Min J, Wang F (2015) Promises and challenges of big data computing in health sciences. Big Data Res 2(1):2–11. https://doi.org/10.1016/j.bdr.2015.02.002
15. Genuer R, Poggi JM, Tuleau-Malot C, Villa-Vialaneix N (2017) Random forests for big data. Big Data Res 9:28–46. https://doi.org/10.1016/j.bdr.2017.07.003
16. Chu C, Kim S, Lin Y, Yu Y, Bradski G, Ng A (2010) Olukotun K Map-reduce for machine learning on multicore. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (eds) Advances in neural information processing systems (NIPS 2010). NIPS, Vancouver, pp 281–288
17. Sun Z, Fox G (2012) Study on parallel SVM based on MapReduce. In: Proceedings of the international conference on parallel and distributed processing techniques and applications
18. Xu K, Wen C, Yuan Q, He X, Tie J (2014) A MapReduce based parallel SVM for email classification. J Netw 9(6):1640–1647. https://doi.org/10.4304/jnw.9.6.1640-1647

19. You ZH, Yu JZ, Zhu L, Li S, Wen ZK (2014) A MapReduce based parallel SVM for large-scale predicting protein–protein interactions. Neurocomputing 145:37–43. https://doi.org/10.1016/j.neucom.2014.05.072
20. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106. https://doi.org/10.1007/BF00116251
21. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
22. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks, Monterey
23. Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. Morgan Kauffmann, Burlington
24. Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20:1–25. https://doi.org/10.1007/BF00994018
25. Collobert R, Bengio S, Bengio Y (2001) A Parallel mixture of SVMs for very large scale problems. Neural Comput 14:1105–1114. https://doi.org/10.1162/089976602753633402
26. Khalili N, Moradi-Lakeh M, Heidarzadeh M (2019) Low birth weight in Iran based on Iranian Maternal and Neonatal Network (IMAN). Med J Islam Repub Iran (MJIRI) 33:30. https://doi.org/10.34171/mjiri.33.30
27. Spinillo A, Capuzzo E, Colonna L, Solerte L, Nicola S, Guaschino S (1994) Factors associated with abruptio placentae in preterm deliveries. Acta Obstetricia et Gynecologica Scandinavica (AOGS) 73(4):307–312
28. Kouhkan A, Khamseh ME, Moini A, Pirjani R, Valojerdi AE, Arabipoor A, Hosseini R, Baradaran HR (2018) Predictive factors of gestational diabetes in pregnancies following assisted reproductive technology: a nested case–control study. Arch Gynecol Obstet 298(1):199–206. https://doi.org/10.1007/s00404-018-4772-y
29. Langer O (2018) Prevention of obesity and diabetes in pregnancy: is it an impossible dream? Am J Obstet Gynecol (AJOG) 218(6):581–589. https://doi.org/10.1016/j.ajog.2018.03.014
30. Bryson CL, Ioannou GN, Rulyak SJ, Critchlow C (2003) Association between gestational diabetes and pregnancy-induced hypertension. Am J Epidemiol 158(12):1148–1153. https://doi.org/10.1093/aje/kwg273
31. NIH (2017) What are the risk factors for preterm labor and birth? https://www.nichd.nih.gov/health/topics/preterm/conditioninfo/who_risk. Accessed 28 Jan 2019
32. Steer P (2005) The epidemiology of preterm labour. BJOG 112(s1):1–3. https://doi.org/10.1111/j.1471-0528.2005.00575.x
33. Morisaki N, Ogawa K, Urayama KY, Sago H, Sato S, Saito S (2017) Preeclampsia mediates the association between shorter height and increased risk of preterm delivery. Int J Epidemiol 46(5):1690–1698. https://doi.org/10.1093/ije/dyx10
34. Oliver-Williams C, Fleming M, Wood AM, Smith GC (2015) Previous miscarriage and the subsequent risk of preterm birth in Scotland, 1980–2008: a historical cohort study. BJOG 122(11):1525–1534. https://doi.org/10.1111/1471-0528.13276
35. Chiavarini M, Bartolucci F, Gili A, Pieroni L, Minelli L (2012) Effects of individual and social factors on preterm birth and low birth weight: empirical evidence from regional data in Italy. Int J Public Health 57(2):261–268. https://doi.org/10.1007/s00038-011-0311-3
36. CDC (2018) Center for disease control and prevention website. https://www.cdc.gov/features/prematurebirth/index.html. Accessed 28 Jan 2019
37. Vestgaard M, Secher AL, Ringholm L, Jensen JE, Damm P, Mathiesen ER (2017) Vitamin D insufficiency, preterm delivery and preeclampsia in women with type 1 diabetes—an observational study. Acta Obstetricia et Gynecologica Scandinavica (AOGS) 96(10):1197–1204. https://doi.org/10.1111/aogs.1318
38. Wang P, Liou SR, Cheng CY (2013) Prediction of maternal quality of life on preterm birth and low birthweight: a longitudinal study. BMC Pregnancy Childbirth 13(1):124. https://doi.org/10.1186/1471-2393-13-124
39. Goldenberg RL, Mercer BM, Meis PJ, Copper RL, Das A, McNellis D (1996) The preterm prediction study: fetal fibronectin testing and spontaneous preterm birth. Obstet Gynecol 87(5):643–648. https://doi.org/10.1016/0029-7844(96)00035-X
40. Alijahan R, Hazrati S, Mirzarahimi M, Pourfarzi F, Ahmadi Hadi P (2014) Prevalence and risk factors associated with preterm birth in Ardabil, Iran. Iran J Reprod Med 12(1):47–56
41. Vakilian K, Ranjbaran M, Khorsandi M, Sharafkhani N, Khodadost M (2015) Prevalence of preterm labor in Iran: a systematic review and meta-analysis. Int J Reprod Biomed (Yazd) 13(12):743–748

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Toktam Khatibi[1,2] · Naghme Kheyrikoochaksarayee[3] · Mohammad Mehdi Sepehri[1,2]

Naghme Kheyrikoochaksarayee
kheyri.naghmeh@gmail.com

Mohammad Mehdi Sepehri
mehdi.sepehri@modares.ac.ir

[1] Healthcare Systems Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran 14117-13114, Iran

[2] Healthcare Systems Engineering, Hospital Management Research Center (HMRC), Iran University of Medical Sciences (IUMS), Tehran, Iran

[3] Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran 14117-13114, Iran