LUND UNIVERSITY

SCHOOL OF
ECONOMICS AND
MANAGEMENT

Predicting preterm birth with machine learning methods

by

Zélie Dresse

May 2022

Master’s Programme in Data Analytics and Business Economics

DABN01 - Master Essay I

Supervisor: Ana Rodriguez-Gonzalez

Abstract

Preterm birth is a leading cause for birth complications and neonatal mortality in the world. It remains difficult to predict whether a preterm birth will occur, which hinders the possible use of prevention treatments. This thesis investigates the use of machine learning models in the prediction of spontaneous preterm birth. In addition, possible heterogeneous performance of these models among different racial groups is explored. Using birth certificate data, retrieved from the Natality Birth Data Sets in the National Vital Statistics System, machine learning models were trained and evaluated. Four machine learning methods are employed: logistic regression, random forests, eXtreme gradient boosting and neural networks. The models' performance is similar across methods, the logistic regression model achieved the lowest test AUC of 0.6710 and the lowest TPR of 30.14% at the 10% FPR level. The eXtreme gradient boosting model performed best with a test AUC of 0.6994 and TPR of 34.15%. All models performed similarly for both black and non-black women. These results confirm previous evidence that this type of easily accessible patient data does not seem to be sufficient to construct high-performing machine learning models.

Keywords: preterm birth, prediction, machine learning, race

Acknowledgements

First, I would like to thank my supervisor Ana Rodriguez-Gonzalez for her guidance and feedback throughout the writing of this thesis.

In addition, I want to thank my parents, family and friends for their support and encouragement throughout my studies.

Table of Contents

List of Tables

List of Figures

1 Introduction

Preterm birth (PTB) remains one of the main causes of birth complications and neonatal mortality in both low-income and high-income countries. The World Health Organization (2018) (WHO) reports that every year around 15 million babies are born preterm. In addition, it is the leading cause for death under five years old. For the United States, the Centers for Disease Control and Prevention (2021) (CDC) reports a preterm birth rate of 10.1% in 2020. Their numbers also reflect the heterogeneity in preterm birth as the rate among African-American women is 14.4%, while the rate stands at 9.1% for white women and 9.8% for Hispanic women. This overall rate of preterm birth has been quite stable for about the last ten years (March of Dimes Perinatal Data Center, 2021).

Apart from the direct relation between preterm birth and neonatal mortality, PTB has many other undesirable effects in the later life of the surviving infant. Currie and Almond (2011) provide an extensive literature review supporting the theory that early-life events before the age of five can have long term impacts in later life. There is a large body of literature specifically concerning the effect of PTB on the infant in their later life. Associations between PTB and a higher risk of certain disabilities, low educational attainment, lower earnings or receiving social security benefits have been found (Moster, Lie & Markestad, 2008). Further, evidence exists that extremely preterm children report poorer health-related quality of life at the adolescent age (Wolke et al., 2013), perform worse in cognitive and mathematical test at age 11 (Simms et al., 2013) and have an increased incidence of ADHD (Bhutta et al., 2002). Other studies performing sibling-comparisons find that there is an increased incidence of infant mortality and autism among PTB babies (D'Onofrio et al., 2013), worse school performance for early PTB children (Ahlsson et al., 2015) and a clear relation between early PTB and language delay (Zambrana et al., 2021)

There exist some prevention methods that are currently used for clearly at-risk mothers, such as mothers having had a previous preterm birth or with a short cervical length. In particular, cervical cerclage and vaginal progesterone have been shown to be effective in reducing the frequency of early PTB and neonatal morbidity (da Fonseca, Damiao & Moreira, 2020). Flood and Malone (2012) also mention the success of clinics dedicated to preterm birth in increasing gestation length and reducing PTB complications.

Given the negative consequences of PTB and as prevention methods are available, it would be desirable to be able to identify mothers who are likely to have a preterm birth. However, it is fairly hard to predict whether a woman is at risk for a preterm delivery. While there are some factors that seem to increase the risk of preterm birth, the majority of women with a preterm delivery have no clear risk factor (Vogel et al., 2018). This is especially the case for women giving birth for the first time as a previous preterm birth is the main indicator for preterm delivery. Glover and Manuck (2018) provide an overview of the existing screening methods for spontaneous preterm birth. There are various methods with varying results. A common method is an ultrasonographic assessment of the cervical length. However, findings seem to indicate that only a small number of low-risk women who end up having a PTB can be identified with such a test. Other existing or upcoming methods are screening through fetal fibronectin, biomarkers, serum proteomics and genetic factors. However, there is no clear method at the moment that is easily used and has a high predictive ability.

The main challenge according to Glover and Manuck is the heterogeneous nature of PTB. As rates show, African-American or non-Hispanic black women are affected by preterm birth more often. Previous studies suggest that the mechanisms behind PTB are different for non-Hispanic black women and this makes it difficult to derive one screening method that fits all. Manuck (2017) explores this racial difference further. She finds that the interpregnancy interval, which is an important indicator for PTB, can only explain 4% of the racial disparity. Socioeconomic factors also explain only a small part of the racial variation. In addition, it is found that biomarkers are clearly race-dependent, the genetic variation between races could play a role and that the vaginal microbiome is more diverse for non-Hispanic black women, which is associated with PTB. This evidence shows that race matters in the case of preterm birth and should be kept in mind when trying to predict PTB.

This previous evidence indicates that predicting preterm birth is quite difficult and an accurate and feasible medical method has not been found yet. However, predicting preterm birth is a crucial step in order to prevent it and reduce the amount of preterm birth related complications.

The aim of this thesis is to construct a machine learning model capable to predict preterm birth based on information found in the birth certificate. This data is found in the Natality Birth Data sets from the National Vital Statistics System, made available by the National Center for Health Statistics (NCHS) (National Center for Health Statistics, 2022; NBER, 2022). This thesis considered the data from years 2016 until 2020. More specifically, only information available in the first and second trimester is used, as this would enable prediction at that time in the pregnancy. In addition, it will be investigated whether models perform as well for black as for non-black women and if not, how this can be resolved. Four methods are explored: logistic regression, random forests, eXtreme gradient boosting (XGBoost) and neural networks. The two main metrics that are used are: Area Under the Receiver Operating Curve (AUC) and the true positive rate (TPR) at the 10% false positive rate (FPR) level.

I find that all four models perform quite similarly. The AUC scores on a test set range from 0.6710, for the logistic regression model, to 0.6994, for the XGBoost model. This is in line with results in existing literature. When considering the TPR rates on test data, the values range from 30.14% (logistic regression) to 34.15% (XGBoost). These results indicate that the models only have limited predictive ability and are possibly not powerful enough to be implemented in a clinical setting. Further research and improvements are necessary to make this possible. Regarding the possible heterogeneous performance between black and non-black women, it is found that models perform similarly between these two groups. In addition, training models separately does not substantially improve performance.

This thesis adds to existing literature by further exploring ML methods in order to find a preterm birth prediction model. A large amount of recent data has been used, containing variables that should be easily available during the pregnancy. Furthermore, an additional focus was placed on possible differences in performance between two main subgroups, black and non-black women. It was shown that the trained models do not suffer from this kind of heterogeneous performance.

This thesis is structured as follows. Section 2 provides an overview of the existing literature concerning the application of machine learning in health and the use of machine learning methods for preterm birth prediction. In Section 3, the data, variables and sample selection are explained and some summary statistics are displayed. The following section, Section 4, provides an explanation of the methods used. Then, I move on to the results of the models in Section 5. Next, Section 6 contains the discussion. Finally, I finish with some concluding remarks.

2 Literature Review

In this section, I first discuss the existing literature concerning the use of machine learning in health applications. Next, some papers with a similar aim as this thesis, namely predicting preterm birth with ML methods, are discussed.

2.1 Machine learning applications in health

Machine learning has often been mentioned as a promising evolution in health care. Deo (2015) provides an overview of machine learning in medicine. He finds that while much has been written about the potential of machine learning in medicine, few applications have successfully been implemented in health care. There are two main ways in which machine learning can contribute to medicine. The first is by assisting in learning patterns from unlabeled data, so-called unsupervised learning. The potential contribution of ML is the highest in this task as it is a difficult task for humans. The second is by carrying out supervised learning. In this case an algorithm mostly learns what a doctor already does, for example predicting a health outcome, but the algorithm can reduce workload and maybe improve prediction.

There are a few areas in which it is shown that machine learning and artificial intelligence is valuable and can improve prediction accuracy. First of all, it seems that it is possible to train algorithms for the interpretation of medical images. Erickson et al. (2017) studies the use of ML in radiology and finds that there is a wide range of applications. Commonly used methods are deep learning, support vector machines (SVM), naive Bayes, decision trees and K-nearest neighbors. They find that these ML models reduce time spent on interpretation. In addition, it can be used to help interpretation of challenging tasks such as pulmonary embolism segmentation, polyp detection and brain tumor segmentation. Some studies have developed models that can detect breast cancer in mammography images and outperform human radiologists consistently (Ragab et al., 2019; McKinney et al., 2020). Furthermore, Al'Aref et al. (2019) discuss applications of ML in cardiology and find promising results in applications such as the interpretation of electrocardiograms, analysis of two-dimensional echocardiography, coronary artery calcium scoring, heart failure diagnosis and classification and many others.

A different source of medical data, in which machine learning can be useful, are electronic health records (EHR). Recently, more and more medical information is documented in these records, which has increased the research of possible ML applications by exploiting this data. Shickel et al. (2018) explore the use of deep learning in these applications. They identify five main ML applications in the literature. Machine learning models can be used to extract information from EHR, perform EHR representation learning, predict health outcomes, apply computational phenotyping and de-identify clinical data. While I am not using actual electronic health record data, the information found in the birth certificate is very similar to what would be found in a patient's health record.

Finally, there exists a body of literature on the application of machine learning in prenatal health. One example is the prediction of vaginal birth after a previous caesarean delivery, Lipschuetz et al. (2020) have used gradient boosting and random forest models to predict this. The gradient boosting model performed the best with an Area Under the Curve (AUC) of 0.793. Another application is the prediction of down syndrome in the first or second trimester, results indicate that the performance is comparable or slightly better than existing methods (Koivu et al., 2018; He et al., 2021).

Aside from the potential of ML in medicine, it is important to keep certain ethical challenges in mind. As McCoy et al. (2020) point out racial bias and the underrepresentation of certain subgroups, such as women, is prevalent in traditional medicine, machine learning can be a way to solve these issues but can also worsen them if used uncarefully. It is therefore important to keep these problems in mind and actively address them. An important example is the Framingham heart study, in which models where trained on a predominantly white population, leaving the model to perform poorly on other populations. It became clear in the previous section, that there are clear racial differences in the prevalence of preterm birth. It will therefore be an additional focus in this thesis.

Other challenges mentioned by Char, Shah and Magnus (2018) are disparities in the goals of users and the intent behind the design of a model, changing physician-patient relations and a need to rethink confidentiality and core ethics in order to be able to process sensitive data.

2.2 Preterm birth prediction with machine learning methods

In recent years, some researchers have started to explore how machine learning could be used for preterm birth prediction. Wlodarczyk et al. (2021) provide an overview of existing work in this area, which data they use and which methods. Four main sources of data used in these types of studies are identified: electrohysterography measurements, electronic health records, transvaginal ultrasounds and uterine electromyography. In what follows, I will only focus on the literature concerning electronic health records as this is the type of data used in this thesis. In general, one of the main technical challenges discussed by Wlodarczyk et al. is the class imbalance that exists in these datasets, as preterm birth prevalence usually ranges around 10%. Common methods to deal with this are oversampling and undersampling. Some commonly used methods for prediction are logistic regression, support vector machines (SVM), decision trees and neural networks.

There are some papers with a similar objective as this thesis. A good example is the work of Koivu and Sairanen (2020), who predict early stillbirth, late stillbirth and preterm birth using a linked birth-death pregnancy CDC dataset. They employ logistic regression, feed-forward neural networks, gradient boosting decision tree and ensemble learning. When testing the models on external test data, the AUC for preterm birth ranges from 0.62 to 0.67. Weber et al. (2018) perform a similar analysis on a smaller Californian dataset with an additional focus on the disparity between non-Hispanic black women and white women. Models were trained separately for both groups and combined. They utilized (penalized) logistic regression, random forest, k-nearest neighbors and generalized additive modelling. The AUC was 0.62 and 0.63 for non-Hispanic black and for white respectively and 0.67 for the combined models.

Lee and Ahn (2019) report similar AUC values ranging from 0.62 to 0.64. They uncover that the most important predicting variables are BMI, cervical length, age, hypertension and diabetes mellitus. Some of these are also reported by Belaghi, Beyene and McDonald (2021), who find that the strongest predictors in their models are diabetes, previous abortions and abnormal pregnancy-related plasma protein A concentrations. Their models based on information in the first trimester lead to AUCs ranging from 0.55 to 0.59, which increase to 0.58-0.64 when including second trimester information. Finally, the inclusion of pregnancy complications leads to an increase up to an AUC of 0.8.

Another noteworthy paper is written by Raja, Mukherjee and Sarkar (2021) who predict preterm birth in rural India, combining a feature selection based on entropy, oversampling and the training of logistic regression, SVM and decision tree models. They achieve a sensitivity of 0.71 for the decision tree, 0.83 for the logistic regression and 0.89 for SVM.

Finally, Rocha et al. (2021) try to predict the week of delivery using different ML methods, such as eXtreme gradient boosting, elastic net, ridge and lasso regression, linear regression and decision trees. Again, all models perform quite similarly, with eXtreme gradient boosting being the best. They are able to estimate the delivery within two weeks. They find that the most important variables are previous C-sections, number of prenatal care visits, age, the availability of ultrasound examination in the care network and the share of primary care teams in the municipality registering the oral care consultation.

In all of these papers the different methods usually lead to similar performance. When looking across the different studies, there is not one machine learning method that stands out compared to the others. The purpose of this thesis is to further add to this literature by exploring a few ML methods and their performance on more recent data. In addition, it will be investigated whether these models perform as well for black as for non-black women.

3 Data

3.1 Dataset description

The models in this thesis are trained and evaluated on the Natality Birth Data from the National Vital Statistics System, made available by the National Center for Health Statistics (NCHS) (National Center for Health Statistics, 2022; NBER, 2022). These public-use datasets consist of information on all live births occurring in the United States and are based on information retrieved from the birth certificates. This information includes demographic data of the parents, information about the mother’s health, about the birth itself and the baby’s health.

3.2 Sample selection

I consider five years of data and include the births from 2016 until 2020, which leads to a total of 18,999,808 births. In order to reach a complete dataset, observations with missing values in relevant categories are removed. The removal of observations should not lead to too much of an information loss due to the large size of the data set. In addition, a quick investigation of these observations made sure that these missing values were mostly random and thus their removal would not lead to a bias. This brings the total amount of observations to 16,973,359 births. For two categorical variables with a high number of missing values, marital status and paternity acknowledged, it was decided to include a category indicating that the information is missing as the observations were not always missing randomly.

For the prediction of spontaneous preterm birth, only spontaneous births were included in the final dataset. This is usually done in the literature as these are the kind of births for which prediction would be the most valuable. A non-spontaneous preterm birth is usually a medical decision made by a doctor, for example due to certain complications. It is planned and thus does not need to be predicted by the model. In practice, this was done by only including either vaginal births without induction of labor or cesarean births without induction of labour and with a trial of labour. This inclusion criteria reduces the number of observations to 8,473,853.

Finally, the dataset is split in three sets, the training set contains 60% of the data (5,084,311 observations) and the validation and test set both contain 20% of the data (1,694,771 observations).

3.3 Variable description

The goal of our model is to perform a binary classification, indicating whether the birth was a preterm birth or not. The outcome variable, preterm birth, is equal to True if the gestational period is equal to 36 weeks or less. This variable is constructed based on the obstetric estimate of gestation, a measure which has been used by the NCHS since 2014.

In order to perform this prediction, a wide range of variables are included. In general, only variables that would already be available in the first or second trimester are selected, as this would enable to predict the preterm birth at that time.

The first set of variables contains general information about the mother. This includes the mother's age but also demographic variables such as information about US nativity, resident status, race and Hispanic origin of the mother. Next, more information that can signal the socio-economic status of the mother is included: the mother's educational level, whether she is a recipient of the Special Supplemental Nutrition Program for Women, Infants and Children (WIC) and the payment source of the delivery. Information about the father was mostly excluded as these lead to a high number of observations with missing values. However, marital status of the mother, a variable indicating whether paternity was acknowledged and a variable indicating if the information about the father's age was available, are included in order to partly cover this factor.

The next set of variables pertains to health conditions. First, a set of risk factors and present infections are included. These variables are concerning pre-pregnancy and gestational diabetes, pre-pregnancy and gestational hypertension, hypertension eclampsia, previous preterm birth, the use of infertility treatment, fertility enhancing drugs or assisted reproductive technology, previous cesareans, Gonorrhea, Syphilis, Chlamydia, Hepatitis B and Hepatitis C. Further information about previous pregnancies is included in the form of the number of prior births now living, prior births now dead, prior other pregnancies, an indicator for the first birth, an indicator for the first pregnancy and the birth interval. Other more general health information included is the mother's height and her BMI. Regarding health behavior, information about the mother's smoking behavior before and during the first and second trimester of the pregnancy is also included.

Finally, there are binary variables indicating whether prenatal care was begun in the first or second trimester or not all. The final included variable is the sex of the born infant, in which the assumption is made that this can be reasonably well predicted in the second trimester.

3.4 Summary statistics

The final dataset consists of 7,795,112 full term births and 678,741 preterm births, which is an overall preterm birth prevalence of 8.01 percent. When considering race, the prevalence among black women is equal to 11.25 percent compared to 7.24 percent for non-black women. This clearly reflects the pattern found in the literature.

Appendix A contains some descriptive statistics of the variables included in the dataset. In what follows, I will briefly discuss them. As mentioned before, black women are clearly overrepresented in the preterm births. Other patterns are a higher percentage of unmarried women, lower educated women, women receiving WIC, more births paid for with Medicaid and a lower percentage of paternity acknowledged for preterm births. This seems to indicate that women with a lower or more vulnerable socio-economic status are more affected by preterm birth.

Regarding the mother's health, results are not very surprising. There are clearly more women in the preterm group that smoke, have a high BMI, have risk factors and infections. Especially the risk factors gestational diabetes, gestational hypertension and previous preterm birth are more present among women having a preterm birth. It also seems that there is a higher percentage of women not having received prenatal care among the preterm births. Finally, the percentage of female births are lower for preterm births. These basic statistics seem to indicate that it might be possible to predict spontaneous preterm birth with these variables, as they vary between full term and preterm births.

4 Methodology

This thesis explores different machine learning methods in order to solve the binary classification problem of classifying births as a preterm birth or not. Four different methods are explored: logistic regression models, random forest models, eXtreme gradient boosting models and neural networks.

To ensure that final models still perform well on new data and avoid overfitting, it is extremely important to perform data splitting. As this is a big dataset, cross-validation does not seem necessary and I have simply split the data in three different sets. Candidate models are first trained on the training dataset, after which their performance is evaluated on the validation set. The tuning of parameters is also done by using the validation set. Based on these results, the final models are selected and are then evaluated on the test set. Using this method ensures that results are generalizable to new unseen data.

The main metrics used are Area Under the Receiver Operating Characteristics Curve (ROC-AUC) and the True Positive Rate (TPR) at the 10 percent False Positive Rate (FPR). These measures are more suitable than, for example, accuracy as they take into account class imbalance and allow for different classification thresholds. In addition, I focus on these metrics as the goal would be to achieve a good TPR while having only a reasonable ratio of false positives. It is more important to correctly identify preterm births as such than to avoid falsely predicting a few preterm births that end up being a full term birth.

The data analysis part of this thesis has been performed in Python 3.8 (Van Rossum & Drake, 2009). The most commonly used packages are NumPy (Harris et al., 2020), pandas (Mckinney, 2010; The pandas development team, 2021), Matplotlib (Hunter, 2007), Scikit-learn (Pedregosa et al., 2011), xgboost (xgboost developers, 2021) and Keras (Chollet & others, 2015).

Wlodarczyk et al. (2021) finds that a wide range of machine learning methods has been used in order to predict preterm birth. The choice of methods depends on different factors. For example, the use of support vector machine models seems promising. However, it is not performed in this thesis due to the large size of the dataset. The “no free lunch” theorem in statistics and machine learning explains why so many different methods are being tried out and used. This theorem says that there is not one method that clearly outperforms all others: the performance depends on the specific data set and on the problem at hand (James et al., 2021). Our findings in the literature review were quite similar. Therefore, I decided to try out several methods and compare their performance.

An obvious first choice to perform classification is a logistic regression, which can be seen as a baseline model. Inspired by previous literature (Weber et al., 2018; Lee & Ahn, 2019; Koivu & Sairanen, 2020), I decided to apply two ensemble methods, random forests and boosting. It has been shown that ensemble methods improve performance compared to training single models (Sagi & Rokach, 2018). In addition, these models include non-linearities in the model, in contrast to the linear logistic regression model. For boosting, the popular XGBoost implementation is chosen. This boosting system is a very effective system that incorporates regularization and uses limited computational resources (Chen & Guestrin, 2016). Finally, I experiment with an artificial neural network model due to its proven performance in a wide range of areas (Abiodun et al., 2018).

4.1 Logistic Regression

Logistic regression is a linear model for classification. The aim is to model the expected value of the outcome variable (y) given the predictors (x), $\mathbb{E}(y_i|\boldsymbol{x_i})$. In the case of logistic regression, the form for the expected value that the outcome variable is equal to one given the features is set as in Equation 1 (Hastie, Tibshirani & Friedman, 2017b). This assures that predicted probabilities are bounded between zero and one.

$$\mathbb{E}(y_i|\boldsymbol{x_i}) = p(y_i = 1|\boldsymbol{x_i}) = \frac{\exp\,(\boldsymbol{x_i^T\beta} + \beta_0)}{1 + \exp\left(\boldsymbol{x_i^T\beta} + \beta_0\right)}$$

Equation 1. Logistic regression probabilities.

In practice, logistic regression models are fitted by maximizing the log-likelihood or equivalently, minimizing the negative log-likelihood. The log-likelihood is given by:

$$\log[L(\boldsymbol{\beta})] = \sum_{i=1}^{N}\{y_i(\boldsymbol{x_i^T}\boldsymbol{\beta} + \beta_0) - \log[1 + \exp(\boldsymbol{x_i^T}\boldsymbol{\beta} + \beta_0)]\}$$

Equation 2. Log-likelihood logistic regression.

In Scikit-learn, ridge (L2) regularization is automatically applied (scikit-learn developers, 2022), meaning that a squared penalty term is added to the minimizing function. This is done in order to avoid overfitting. Finally, from Equation 2 and by adding the regularization term, the cost function in Equation 3 can be derived, this is the function that Scikit-learn is minimizing when fitting a logistic regression.

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\boldsymbol{\beta^T}\boldsymbol{\beta} + \sum_{i=1}^{n}\log\left(exp\left(-y_i(\boldsymbol{x_i^T}\boldsymbol{\beta} + \beta_0)\right) + 1\right)$$

Equation 3. Loss function logistic regression.

Standardization is applied to help with the convergence of the models and for the application of certain feature selection methods. In order to deal with the high class imbalance in the dataset, I chose to apply random undersampling. In random undersampling, random observations in the majority class are removed until a certain majority/minority class ratio is reached. In this case the ratio was 1:1. Due to the large size of the dataset, the random loss in observations does not seem to lead to too much of loss in information. Other undersampling and oversampling techniques are available, but random undersampling proved to be the most feasible as it helped reducing the size of the dataset and thus helped with computation.

Finally, feature selection was performed using ANOVA F-test feature selection, in which the variables with the highest ANOVA F-value were selected. The number of variables selected are chosen by the researcher. Two other feature selection methods, variance thresholding and selection based on computed chi-squared statistics, were also experimented with. However, the ANOVA F-value seemed to select features the best way.

4.2 Random Forest

Random forest is an ensemble method for classification. An ensemble method is a method that trains several models and then combines them to achieve better results. More specifically, random forest models combine many decision-tree models. A decision tree is a model that is developed by asking a set of questions about the features, each of these questions being represented by a node. After going through the tree and answering the different questions, you arrive at a final point, a leaf, which returns the prediction for the outcome. Such a tree is fitted by splitting nodes in a way that minimizes a certain loss function. The easiest way to understand a decision tree is visually. Figure 1 shows an example of a binary classification decision tree.

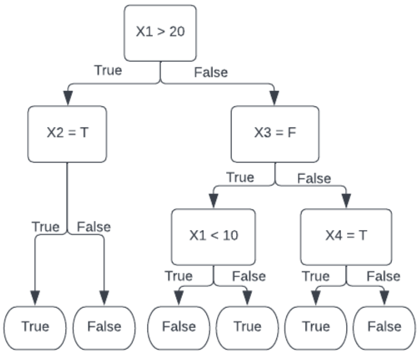

Figure 1. Illustration of a decision tree.

One possible issue is that a decision tree can have high variance. A solution to this is bootstrap aggregating or bagging. A bagging method trains multiple trees on different bootstrap samples of the same data and then averages these models. Figure 2 illustrates how a bagging classification model could look like. This ensemble technique smooths out the prediction and thus reduces the variance. Random forest is such a bagging model, in which decision trees are trained with the important restriction that in each step where a new split is decided on, only a few randomly selected features are candidates for that split. This leads to a possible further improvement in variance by aggregating less correlated trees. (Hastie, Tibshirani & Friedman, 2017c).

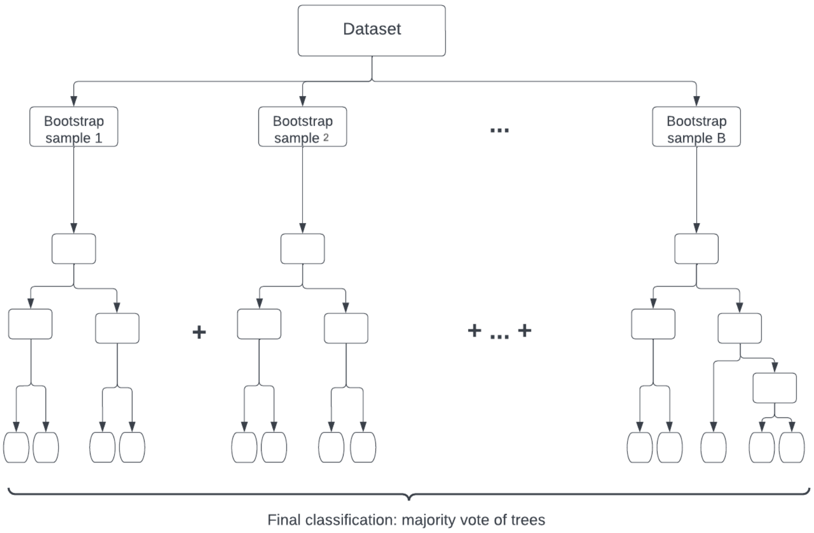

Figure 2. Illustration of a bagging model.

To summarize, the classification random forest algorithm goes as follows (Hastie, Tibshirani & Friedman, 2017c):

1. A number (B) of bootstrap samples are drawn from the training data
2. A tree is trained on each bootstrap sample by repeating the following steps for each split, until a certain minimum node size is reached
    a. Randomly select m variables from all the predictor variables
    b. Pick the best split among these variables
    c. Split the node in two according to this best split
3. The output is the ensemble of trees.

In order to get to a prediction for a new observation, the predictions of all the trees in the random forest are taken and a majority vote will decide on the final prediction.

There are two main parameters to be tuned in the analysis: the depth of the tree and the number of features randomly drawn in each step. Regarding the depth, in theory a fully grown tree is best, although this leads to overfitting. As for the number of chosen features, the default chosen value in a classification model is the square root of the number of variables. Other parameters to consider are the criterion to evaluate the quality of a split, either Gini impurity or the information gain, the minimum number of observations to reach a terminal leaf, and the number of trees to be estimated.

4.3 eXtreme Gradient Boosting

Apart from bagging methods, there is another category of ensemble methods, so-called boosting methods. eXtreme Gradient Boosting (XGBoost) is such a boosting method that also implements regularization. The idea of boosted trees is that many simple decision trees, called weak learners, are sequentially trained to modified versions of the same dataset to arrive at a final model that reduces variance and bias (Hastie, Tibshirani & Friedman, 2017a). The dataset used in the training is reweighted in each iteration, and observations that were incorrectly predicted at the previous step will receive higher weights in the next iteration. The difference with bagging methods becomes clearer when visualizing it like in Figure 3.

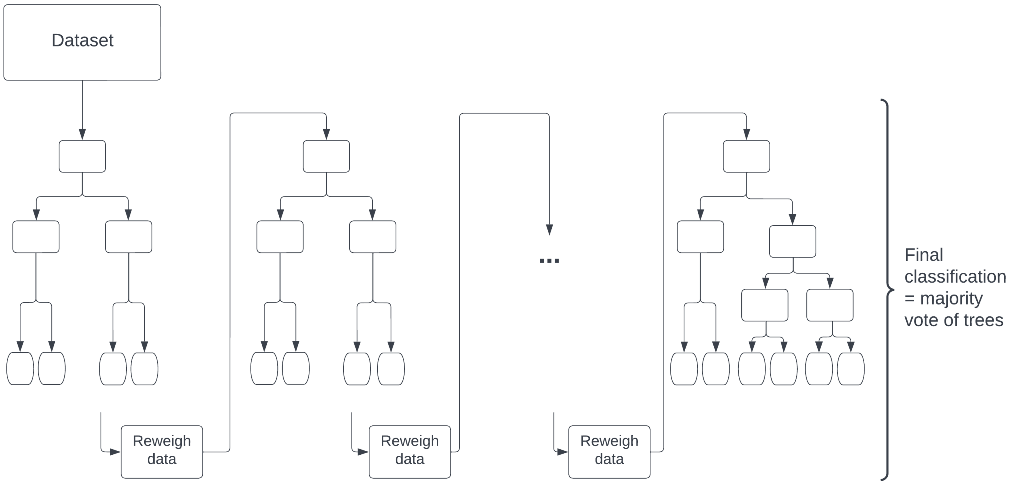

Figure 3. Illustration of a boosting model.

Gradient boosting is a generalization of boosting, which allows it to optimize any loss function, as long as it is differentiable. The algorithm used to optimize this function uses gradient descent, in which steps are made towards the minimum by moving against the direction of the gradient of the function. Finally, XGBoost, is an optimized, efficient implementation of gradient boosting. Some of its advantages are that it implements regularization for the trees and allows parallel processing, which reduces computation time significantly.

There are several hyperparameters that should be tuned in XGBoost. The method used in this thesis is based on the method proposed by Analytics Vidhya (2016). Early stopping is used to avoid overfitting. First, the depth of the tree and the minimum weight in a child are tuned. As explained in the XGBoost documentation (xgboost developers, 2021), this second variable is the minimum sum of instance weight in a child, a child being similar to a split. This is a form of regularization. Next, gamma is tuned. This variable indicates the minimum reduction in loss required in order to make a split, another form of regularization. After that, subsample ratio is optimized, in XGBoost this parameter means the ratio of observations randomly sampled before growing the trees in order to avoid overfitting. At the same time, the subsample ratio of columns is also tuned. This is the same parameter as in random forests which decided how many variables are randomly selected to be a candidate for the next split. Furthermore, L2 regularization is applied and the related lambda parameter is tuned. Finally, the learning rate, the step size used in the optimizing algorithm, and the number of estimated trees must be considered.

4.4 Neural Networks

The final method used in this thesis is a neural network model. More specifically, feedforward neural networks will be trained. Chapter 6 in the book Deep Learning by Goodfellow, Bengio and Courville (2016) provides an extensive explanation of feedforward neural networks.

A neural network can be seen as a chain of functions. Each layer in a network is a function that takes the output of the previous layer as its input. Figure 4 shows us a possible feedforward neural network. Our features or inputs are considered as the input layer and are entered into the first hidden layer, which contains eight so called hidden inputs or neurons. The second hidden layer has four hidden inputs, which take the output of the first hidden layer as their input.

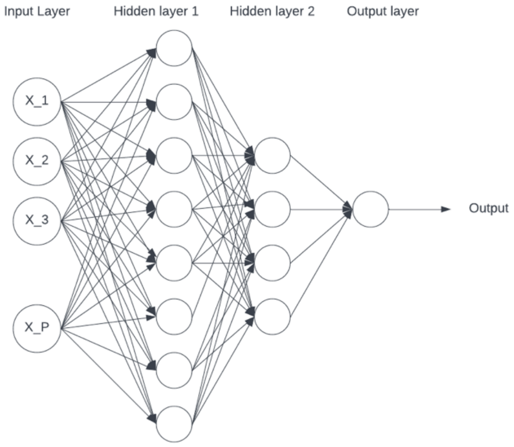

Figure 4. Illustration of a feedforward neural network.

In each layer, weights and a bias are applied to the inputs. For example, in the first layer, the following values (Equation 4) are calculated in each of the eight hidden units and are the outputs of this layer.

$$h_k^{(1)} = w_{k0}^{(1)} + \sum_{j=1}^{P} w_{kj}^{(1)} X_j , \quad k = 1, \dots , 8$$

Equation 4. Output of the first hidden layer.

Next, the same calculation is done in the second hidden layer, where the outputs of the first hidden layer are taken as the input.

$$h_k^{(2)} = w_{k0}^{(2)} + \sum_{j=1}^{8} w_{kj}^{(2)} h_j^{(1)}, \quad k = 1, \dots 4$$

Equation 5. Output of the second hidden layer.

Finally, the same formula is applied in the output layer.

$$y = \beta_0 + \sum_{j=1}^{4} \beta_j h_j^{(2)}$$

Equation 6. Output of the output layer

The term feedforward means that the information flows from the inputs to the output through different layers without passing feedback from a layer output to previous layers. This is the case for this specific network. It takes the p input variables, passes them through two hidden layers, the first with eight units and the second with four units, to then pass it through the final output layer which gives us the output, in our case the probability that the observation belongs to class 1 (or the True class).

The above explanation was a simplified version of what is usually done. In order to introduce nonlinearities in the model, activation functions are almost always applied to each layer. In this thesis, the rectified linear unit (ReLU) is chosen as the activation function in the hidden layers, which is usually the default choice in modern neural networks. This function is defined as $f(z) = \max(0, z)$. So, the true output of the first hidden layers will be:

$$f\left(h_k^{(1)}\right) = \max\left(0, w_{k0}^{(1)} + \sum_{j=1}^{P} w_{kj}^{(1)} X_j\right)$$

Equation 7. ReLU activation function applied to the first hidden layer.

As we are training a binary classification model, the sigmoid function is chosen as the activation function for the output layer. This function, defined by Equation 8, ensures that the output is bounded between zero and one. This is desirable as the output in our model is the probability that the birth is a preterm birth.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Equation 8. Sigmoid function.

The actual fitting of the neural network model happens through backpropagation. The weights are initially set to a random number close to zero. Then, an optimizing method is used to update these weights towards optimal values that minimize a certain loss function. In this thesis, the Adam optimizer is used, which is also a gradient descent method.

Again, there are some parameters that need to be tuned. Of course, one must decide on how many layers and how many units in each layer the model should have. Next, the learning rate of the optimizer has to be carefully tuned. Another parameter is the number of epochs or the number of times the algorithm cycles through the whole dataset while updating the weights. This is chosen by applying early stopping in order to avoid overfitting. Within one epoch, the model is trained multiple times on smaller subsets of the data, a so-called batch, in order to update weights more often and reduce computation, as computing a gradient on the whole data is quite heavy. Therefore, one must also choose the batch size, which is the number of samples included in one batch, typical choices are 128, 256 or 512. Finally, I consider two other parameters. First, the initial weights that are given to the algorithm can play a role and different weights should be experimented with. At last, it could also be beneficial to apply L2-regularization, this is performed by choosing the final parameter lambda.

5 Results

5.1 Logistic Regression

5.1.1 General results

The first model (Model 1) is a baseline logistic regression model. This model was trained on a randomly undersampled and standardized dataset. Next, feature selection using the ANOVA F-value was tested. A model was trained for each set of selected variables by the algorithm, going from selecting only one variable to all variables. The AUC scores of these models were then evaluated using the validation set. Figure 5 shows the results. Clearly, the feature selection selects good variables in the beginning, as the AUC increases quite a lot. After around 30, the increases in validation AUC are only small and it might be interesting to only include a part of them for computational purposes. However, there is a remarkable jump around 45 variables, where the AUC shoots up a bit again. Given these insights, a model (Model 2) comprising of the 30 first selected variables and the variable leading to the additional increase, which is birth interval, was trained.

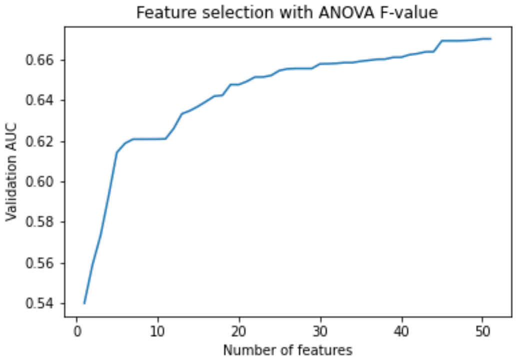

Figure 5. Validation AUC scores for logistic regression models with the number of selected variables by ANOVA F-test going from one to fifty-one.

Figure 6 shows the Receiver Operating Characteristics (ROC) curve on the test data for both models. Applying variable selection decreases the test AUC only slightly from 0.6710 to 0.6634. In summary, Model 1 achieves an AUC of 0.6710 and a TPR of 30.14% at the 10% FPR level. Model 2 results in an AUC of 0.6634 and a TPR of 29.25% at the 10% FPR level.

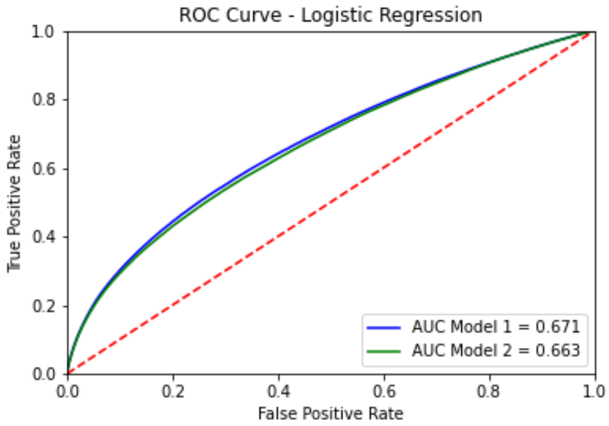

Figure 6. ROC curve on the test data for two logistic regression models.

Figure 7 plots the ten largest (in absolute terms) coefficient values in Model 2. A previous preterm birth is clearly an important factor in the prediction, which is in line with the findings in the literature. Furthermore, the birth interval, gestational hypertension, having had no prenatal care and gestational diabetes are the variables with the highest coefficients in this model.

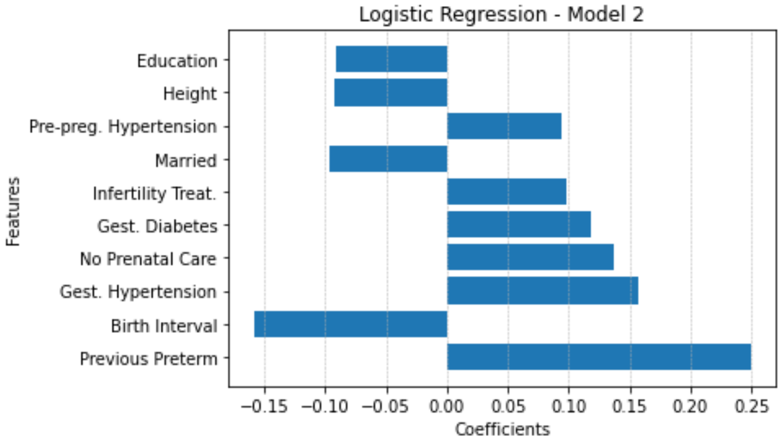

Figure 7. Coefficient values for the variables with the ten largest coefficients in Model 2.

5.1.2 Heterogeneous performance

In this section, we consider the performance of Model 1 across two groups, black women and non-black women. Figure 8 shows the ROC curves and AUC score for those specific observations. It is quite clear that the model performs equally well for both groups. The same is reflected in the TPR at 10% FPR of 29.17% and 29.66%, for black and non-black women respectively.

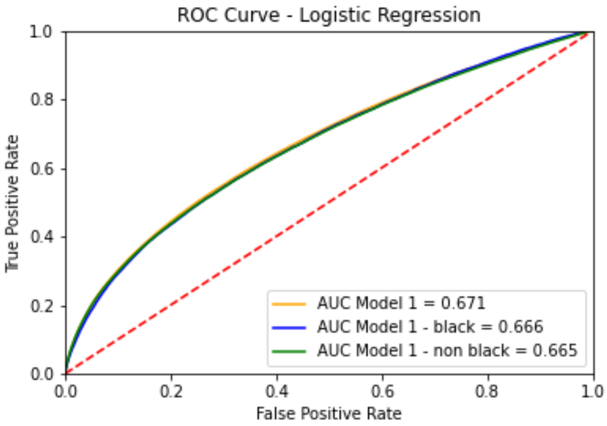

Figure 8. ROC curve for Model 1, in yellow on the whole test set, blue only on black women and green on only non-black women.

Next, I trained two separate models for these groups in order to explore whether this would improve the performance. This did slightly improve the AUC for black women to 0.672, which is similar to the overall performance of Model 1. The TPR at 10% FPR becomes 30.23% for black women and 29.73% for non-black women. However, this is a minimal improvement.

5.2 Random Forest

5.2.1 General

In the random forest model, it was again decided to use a randomly undersampled and standardized dataset. After careful tuning of the hyperparameters, maximum tree depth and the number of randomly selected features in each split decision were set to 21 and 9 respectively. The number of trees was set to 100, as training more trees did not improve the results further and the entropy criterion was used. This led to the final model (Model 3) with a test AUC score of 0.6939 and a TPR of 33.28% at the 10% FPR level. Figure 9 shows the corresponding ROC curve.

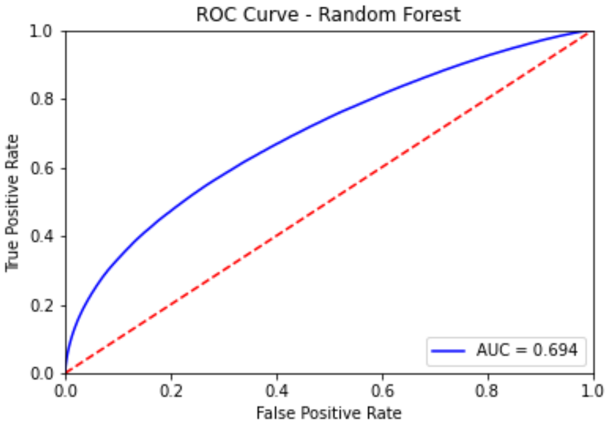

Figure 9. ROC curve on the test data set for the random forest model.

For a random forest model, it is possible to calculate a variable's importance in terms of the decrease in impurity attributed to that variable. Figure 10 shows the variable importance of the ten most important variables. Some variables are the same as the ones identified in the logistic regression but there are also some newcomers, such as age, prior births now alive, race and the payment method. Again, the birth interval and previous preterm birth play the major roles in the prediction.

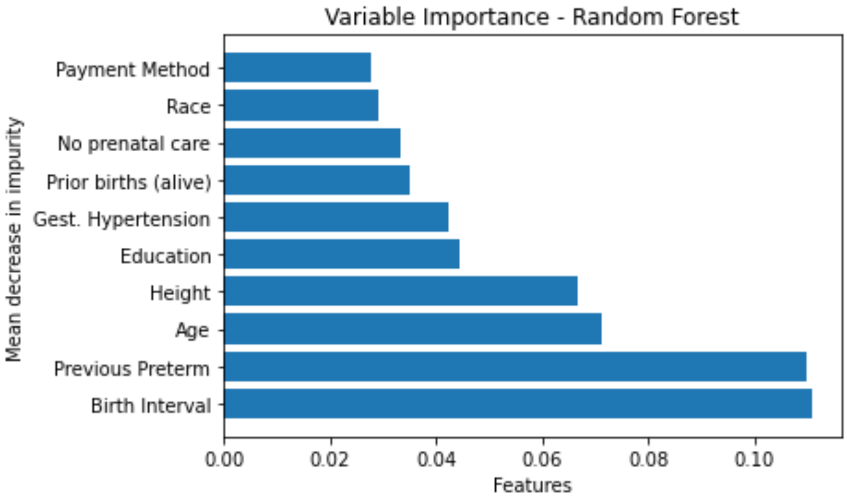

Figure 10. Variable importance in terms of mean decrease in impurity for the ten most important variables in the random forest model.

5.2.2 Heterogeneous performance

Similar to the logistic regression model, Model 3 performs equally well for black and non-black women. The AUC for black women is 0.6873 and 0.6874 for non-black women. The TPR at 10% FPR is 31.98% and 32.73% respectively.

When training separate models for these two groups and tuning the parameters separately, the performance is almost exactly the same. The AUC becomes 0.6867 and 0.6872 and the rates 32.77% and 32.50%, for black women and non-black women respectively.

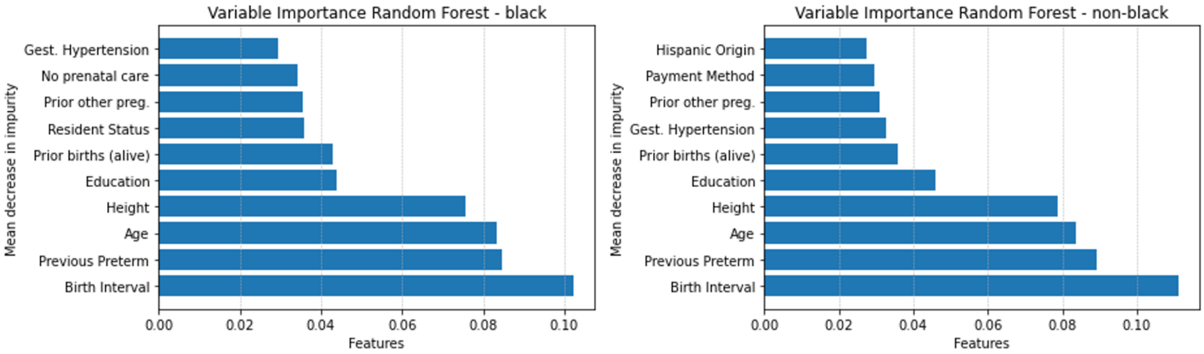

Figure 11. Variable importance from separately trained random forest models on each subgroup.

It might be interesting to consider whether the same variables are important in these separate models. Figure 11 shows the variable importance for the models trained separately on the two subgroups. While the six most important variables are the same, there are some differences after that. For black women the resident status and having had no prenatal care has some predictive importance, while for non-black women the payment method and the Hispanic origin seem to be more relevant.

5.3 eXtreme Gradient Boosting

5.3.1 General

Next, the XGBoost model was trained on the undersampled and standardized dataset. The tuning of the hyperparameters led to the following choices: a learning rate of 0.2, 233 estimated trees (chosen by applying early stopping), a maximum depth of 6 and a minimum weight in a child of 4. Furthermore, the minimal reduction gamma was set to 0, the subsample ratio of the columns to 0.5, the subsample ratio to 0.9 and the lambda for the L2 regularization to 80. This leads to Model 4 with the ROC curve shown in Figure 12. The AUC is equal to 0.6994 and the TPR at 10% FPR is equal to 34.15%, the best results so far.

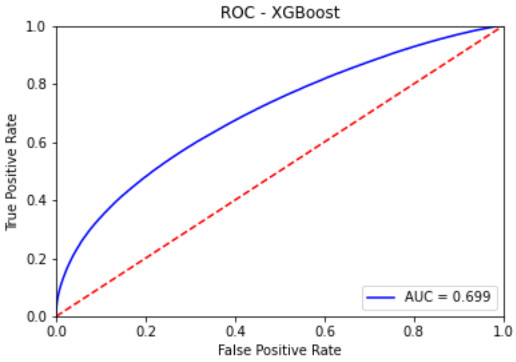

Figure 12. ROC curve on the test data for the XGBoost model.

Figure 13 shows us the 10 most important variables again. Surprisingly, height now seems to be the most important variable and a previous preterm birth is not included at all.

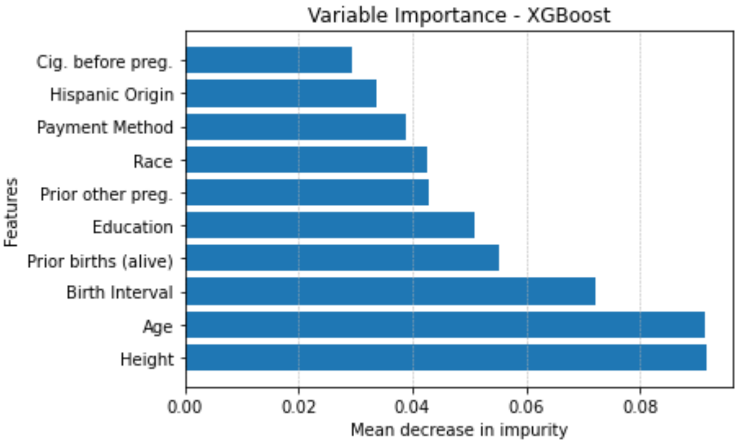

Figure 13. Variable importance in terms of mean decrease in impurity for the ten most important variables in the XGBoost model.

5.3.2 Heterogeneous performance

Once again, performance across our two main subgroups were not very different. The AUC for black women was 0.6933 and the TPR 33.16%. For non-black women, these scores were 0.6932 and 33.53%. When training and tuning separate models for the subgroups, the performance was still not improved, with an AUC of 0.6909 and 0.6907 and rates of 33.34% and 33.33%.

5.4 Neural Networks

5.4.1 General

Finally, a neural network was trained. The hyperparameters were carefully tuned using the validation set and the following choices were made. The model consists of 2 hidden layers, the first with 128 hidden units and the second with 64 hidden units. The ReLU activation function was applied to these layers, while the sigmoid function was applied to the output layer. In addition, L2 regularization with lambda equal to 0.001 was applied and a random normal weight initialization was used. Next, the learning rate for the Adam optimizer was equal to 0.0001, 100 epochs were trained and a batch size of 128 was used in this process. Finally, early stopping was applied with a patience of 20.

This model resulted in a test AUC of 0.6964 and a TPR of 33.88% at the 10% FPR level. There is no obvious method to evaluate the variable importance in a neural network.

5.4.2 Heterogeneous performance

Figure 14 shows the AUC curve for this overall model and for the two subgroups. It is again clear that performance is very similar across the groups and the same goes for the rate which is 33.10% and 33.27%, for black and non-black respectively. Training the models separately does not improve performance at all and the AUC drops to 0.6890 and 0.6883, while both TPR rates are around 32.96%

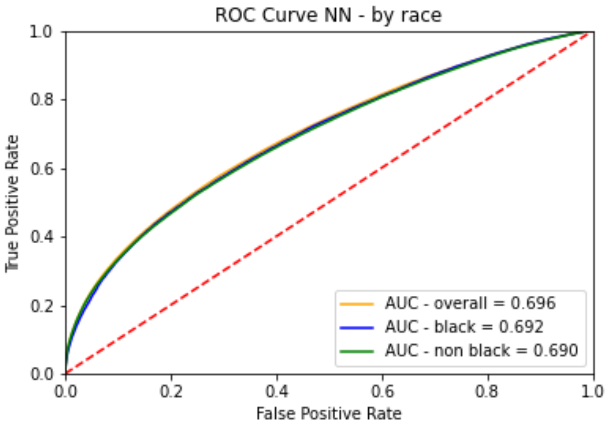

Figure 14. ROC curve for the neural network model, in yellow on the whole test set, blue only on black women and green on only non-black women.

5.5 Overall results

Table 1 summarizes the found AUC scores and TPRs. Overall, results are quite similar between methods and little improvement compared to the basic logistic regression is found when using more advanced and computationally heavy methods such as random forest, boosting and neural networks. However, we can conclude that the XGBoost model reached the highest test AUC score and highest true positive rate of 34.15% at the 10% false positive rate level.

	AUC	TPR at 10% FPR
Model 1 - LR	0.6710	30.14%
Model 3 - RF	0.6939	33.28%
Model 4 - XGB	0.6994	34.15%
Model 5 - NN	0.6964	33.88%

Table 1. Overall results of the trained models.

6 Discussion

The goal of this thesis was to build a machine learning model that could predict preterm birth based on information available in the first or second trimester of the pregnancy. Four methods were explored: logistic regression, random forests, eXtreme gradient boosting and neural networks. These models resulted in test AUC scores ranging from 0.6710 to 0.6994. When considering the true positive rates ranging from 30.14% to 34.15% at the 10% FPR level, it becomes clear that these models are not very successful at identifying preterm birth while accepting a reasonable amount of mistakes. Another finding is that the four methods perform quite similarly, even though logistic regression is a quite simple linear model compared to for example neural networks, the performance of the complex models are only slightly better.

In comparison with previous literature, where AUC scores were ranging from 0.62 to 0.67, the trained models are slightly better. For the TPR, Koivu and Sairanen (2020) find results around 27-31%. Again, our results improved this performance slightly. Note that some of the papers in the literature focus on predicting preterm birth for first-time mothers.

One important note is that AUC was used as a metric when tuning models. For example, in the random forest models when performing searches among possible parameter values, the AUC was used to evaluate the models and pick the best one. The implication of this is that the TPR could maybe be improved a bit more if the focus was set on this metric but it is improbable that big improvements would be achieved.

While the models have extensively been tested on a quite large test set and this should ensure the generalizability, one could wonder if the models would also work on other data sets. The birth certificates are of course only available after the birth, so the information would need to be retrieved from hospital data sets or specific surveys made for this purpose. It is not certain that models would achieve the same performance if these data sets were a bit different. A similar question is whether the models can be extended to other countries or time frames. Hopefully, the model should not be too sensitive to a certain time period as they were trained over a five-year period, including the COVID-19 year 2020. In addition, there is no obvious reason to suspect that the mechanisms behind preterm birth would wildly vary in a short time span.

One technical constraint in this thesis was the computational resource available. As the dataset was quite large and in order to keep computation time reasonable, it was decided to use random undersampling. This tackled the class imbalance problem posed in the data, while also reducing the size of the training data considerably. A negative side-effect of random undersampling is that part of the observations is removed and it is not certain that these observations do not contain crucial information. It might therefore be possible that better models could be achieved by employing other methods, such as random oversampling, SMOTE or a combination of over- and undersampling.

Another limitation of the models could lie in the variables that are important for prediction. When looking at the variable importance, some variables stand out. A previous preterm birth, the birth interval and the number of prior births and pregnancies are quite important predictors in the models. These are all variables that are related to previous pregnancies and would thus not be available for first-pregnancy mothers. This indicates that it would be even harder to predict a preterm birth for these mothers. When fitting the final models determined in the analysis on this subset of mothers, the results displayed in Table 2 are found. It is clear that the performance has declined and especially the TPR at 10% FPR level are lower.

	AUC	TPR at 10% FPR
Model - LR	0.6348	23.90%
Model - RF	0.6338	23.27%
Model - XGB	0.6417	24.81%
Model - NN	0.6396	24.67%

Table 2. Results of the models on the subset of first-pregnancy mothers.

While these models could be used in a medical setting to possibly detect preterm births that would otherwise not be expected, one could wonder whether the models' performance is good enough to be implemented. However, it might be possible to use them as a basis for similar but more advanced models. It is likely that a hospital record would contain more health information than a birth certificate that could possibly help the prediction. The information of certain simple medical tests such as blood tests or other standard examinations done early in the pregnancy could then be added to these models. This information might be easy to uncover or might even already be available in other data sets and could further improve models' performance.

Finally, a note about heterogeneous performance among different racial groups. Another goal of this thesis was to investigate whether this was an issue in these models and if there was a possible solution to it. However, it became quite clear that the models did not suffer from this, as performance was similar for both black and non-black women. One possible explanation is that the dataset was representative for the American population and models were trained on this whole dataset. Issues would more likely arise if models were only trained on predominantly white or rich populations and then expanded to the wide population. Another factor is that race was available in the dataset and could be included as a variable in the models, so models could take this into account. This could be an issue in other countries where race is not reported as frequently as it is in the United States.

7 Conclusion

The main purpose of this thesis was to employ machine learning methods in order to predict preterm birth. Such a model would improve identification of at-risk mothers in the first or second trimester and enable them to receive treatments that could reduce the risk of a preterm birth. Models were trained on the Natality Birth Data provided by the NCHS, including the years 2016 until 2020. This data set contains information retrieved from the birth certificates of all live births in the United States. Four methods were applied to the data: logistic regression, random forests, eXtreme gradient boosting and neural networks. An additional aim of the thesis was to investigate whether trained models suffered from heterogeneous performance between two subgroups, black and non-black women.

When evaluating the models on a test data set, the following results were found. The logistic regression model performed worst with a test AUC of 0.6710 and a TPR of 30.14%. The three other models improved performance slightly. The random forest model achieved an AUC of 0.6939 and a TPR of 33.28% and the neural network model had an AUC of 0.6964 and TPR of 33.88%. While its performance was very similar, the XGBoost model performed best with an AUC of 0.6994 and a TPR of 34.15%. Regarding the heterogeneous performance, the main conclusion is that all these models performed very similarly for both black and non-black women.

This thesis adds to previous papers by further exploring how ML methods can be used in order to predict PTB. More recent data was used and of course the choices in tuning and constructing the models differed. The results in this thesis are in line with previous attempts and indicate that it is quite hard to use this type of patient data in order to predict preterm birth. Additionally, I also focused on ensuring that the models work for both black and non-black women.

As mentioned before, the performance of the models is probably not sufficient for it to be implemented in a clinical setting. Another question is whether the performance of the models would differ when applied to other data sets, in other countries, or in another time period. Another important weakness in the models is that the main variables contributing to the prediction of PTB are variables relating to a previous pregnancy. This means that it would be even harder to build a successful model for first-time mothers. The additional analysis presented in the discussion confirms that this is the case.

Further research and extensions of the models are necessary in order to find a model that performs well enough. One possible suggestion would be to add more and other variables. It is likely that hospital records contain other relevant variables that could improve prediction. In addition, it should be investigated if there are more possible predictors that could easily be detected and included, such as for example biomarkers, fetal fibronectin or genetic factors. Hopefully, this could lead to a model with a high predictive ability that would enable easy and accurate prediction of preterm birth. This would in turn make it possible to administer prevention treatments and possibly reduce preterm birth rates in the world.

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. & Arshad, H. (2018).State-of-the-Art in Artificial Neural Network Applications: A Survey, *Heliyon,* vol. 4, no. 11

Ahlsson, F., Kaijser, M., Adami, J., Lundgren, M. & Palme, M. (2015).School Performance after Preterm Birth, *Epidemiology,* vol. 26, no. 1**,** pp. 106-11

Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., Maliakal, G., van Rosendael, A. R., Beecy, A. N., Berman, D. S., Leipsic, J., Nieman, K., Andreini, D., Pontone, G., Schoepf, U. J., Shaw, L. J., Chang, H. J., Narula, J., Bax, J. J., Guan, Y. & Min, J. K. (2019).Clinical Applications of Machine Learning in Cardiovascular Disease and Its Relevance to Cardiac Imaging, *Eur Heart J,* vol. 40, no. 24**,** pp. 1975-1986

Analytics Vidhya. (2016). *Complete Guide to Parameter Tuning in Xgboost with Codes in Python* [Online]. Available online: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/#h2_10 [Accessed 4 May 2022]

Belaghi, R. A., Beyene, J. & McDonald, S. D. (2021).Prediction of Preterm Birth in Nulliparous Women Using Logistic Regression and Machine Learning, *Plos One,* vol. 16, no. 6

Bhutta, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M. & Anand, K. J. (2002).Cognitive and Behavioral Outcomes of School-Aged Children Who Were Born Preterm: A Meta-Analysis, *JAMA,* vol. 288, no. 6**,** pp. 728-37

Centers for Disease Control and Prevention. (2021). *Reproductive Health - Preterm Birth* [Online]. Available online: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm [Accessed 29 April 2022]

Char, D. S., Shah, N. H. & Magnus, D. (2018).Implementing Machine Learning in Health Care - Addressing Ethical Challenges, *New England Journal of Medicine,* vol. 378, no. 11**,** pp. 981-983

Chen, T. Q. & Guestrin, C. (2016).Xgboost: A Scalable Tree Boosting System, *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining***,** pp. 785-794

Chollet, F. & others. (2015). *Keras* [Online]. Available online: https://keras.io

Currie, J. & Almond, D. (2011). Chapter 15 - Human Capital Development before Age Five. in: Card, D. & Ashenfelter, O. (eds.) *Handbook of Labor Economics.* Elsevier pp. 1315-1486

D'Onofrio, B. M., Class, Q. A., Rickert, M. E., Larsson, H., Langstrom, N. & Lichtenstein, P. (2013).Preterm Birth and Mortality and Morbidity: A Population-Based Quasi-Experimental Study, *JAMA Psychiatry,* vol. 70, no. 11**,** pp. 1231-40

da Fonseca, E. B., Damiao, R. & Moreira, D. A. (2020).Preterm Birth Prevention, *Best Pract Res Clin Obstet Gynaecol,* vol. 69**,** pp. 40-49

Deo, R. C. (2015).Machine Learning in Medicine, *Circulation,* vol. 132, no. 20**,** pp. 1920-30

Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. (2017).Machine Learning for Medical Imaging, *Radiographics,* vol. 37, no. 2**,** pp. 505-515

Flood, K. & Malone, F. D. (2012).Prevention of Preterm Birth, *Semin Fetal Neonatal Med,* vol. 17, no. 1**,** pp. 58-63

Glover, A. V. & Manuck, T. A. (2018).Screening for Spontaneous Preterm Birth and Resultant Therapies to Reduce Neonatal Morbidity and Mortality: A Review, *Seminars in Fetal & Neonatal Medicine,* vol. 23, no. 2**,** pp. 126-132

Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep Feedforward Networks. *Deep Learning.* MIT Press pp. 164-223

Harris, C. R., Millman, J. K., van der Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernandez del Rio, J., Wiebe, M., Peterson, P., Gerar-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020).Array Programming with Numpy, *Nature,* vol. 585, no. 7825**,** pp. 357-362

Hastie, T., Tibshirani, R. & Friedman, J. (2017a). Boosting Methods. *The Elements of Statistical Learning.* 2nd ed.: Springer pp. 337 - 341

Hastie, T., Tibshirani, R. & Friedman, J. (2017b). Logistic Regression. *The Elements of Statistical Learning.* 2nd ed.: Springer pp. 119-120

Hastie, T., Tibshirani, R. & Friedman, J. (2017c). Random Forests. *The Elements of Statistical Learning.* 2nd ed.: Springer pp. 587 - 589

He, F. L., Lin, B., Mou, K., Jin, L. Z. & Liu, J. T. (2021).A Machine Learning Model for the Prediction of Down Syndrome in Second Trimester Antenatal Screening, *Clinica Chimica Acta,* vol. 521**,** pp. 206-211

Hunter, J. D. (2007).Matplotlib: A 2d Graphics Environment, *Computing in Science & Engineering,* vol. 9, no. 3**,** pp. 90-95

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R, 2nd:

Koivu, A., Korpimaki, T., Kivela, P., Pahikkala, T. & Sairanen, M. (2018).Evaluation of Machine Learning Algorithms for Improved Risk Assessment for Down's Syndrome, *Comput Biol Med,* vol. 98**,** pp. 1-7

Koivu, A. & Sairanen, M. (2020).Predicting Risk of Stillbirth and Preterm Pregnancies with Machine Learning, *Health Information Science and Systems,* vol. 8, no. 1

Lee, K. S. & Ahn, K. H. (2019).Artificial Neural Network Analysis of Spontaneous Preterm Labor and Birth and Its Major Determinants, *J Korean Med Sci,* vol. 34, no. 16

Lipschuetz, M., Guedalia, J., Rottenstreich, A., Persky, M. N., Cohen, S. M., Kabiri, D., Levin, G., Yagel, S., Unger, R. & Sompolinsky, Y. (2020).Prediction of Vaginal Birth after Cesarean Deliveries Using Machine Learning, *American Journal of Obstetrics and Gynecology,* vol. 222, no. 6

Manuck, T. A. (2017).Racial and Ethnic Differences in Preterm Birth: A Complex, Multifactorial Problem, *Seminars in Perinatology,* vol. 41, no. 8**,** pp. 511-518

March of Dimes Perinatal Data Center. (2021). *2021 March of Dimes Report Card* [Online]. Available online: https://www.marchofdimes.org/mission/reportcard.aspx

McCoy, L. G., Banja, J. D., Ghassemi, M. & Celi, L. A. (2020).Ensuring Machine Learning for Healthcare Works for All, *BMJ Health Care Inform,* vol. 27, no. 3

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J. & Shetty, S. (2020).International Evaluation of an Ai System for Breast Cancer Screening, *Nature,* vol. 577, no. 7788**,** pp. 89-94

Mckinney, W. (2010) Published. Data Structures for Statistical Computing in Python. in*:* van der Walt, S. & J., M., eds. Proceedings of the 9th Python in Science Conference, 2010. pp. 56-61

Moster, D., Lie, R. T. & Markestad, T. (2008).Long-Term Medical and Social Consequences of Preterm Birth, *N Engl J Med,* vol. 359, no. 3**,** pp. 262-73

National Center for Health Statistics. (2022). *Nvss - National Vital Statistics System* [Online]. Available online: https://www.cdc.gov/nchs/nvss/ [Accessed 9 May 2022]

NBER. (2022). *Vital Statistics Natality Birth Data* [Online]. Available online: https://www.nber.org/research/data/vital-statistics-natality-birth-data [Accessed Feb 26 2022]

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011).Scikit-Learn: Machine Learning in Python, *Journal of Machine Learning Research,* vol. 12**,** pp. 2825-2830

Ragab, D. A., Sharkas, M., Marshall, S. & Ren, J. C. (2019).Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines, *Peerj,* vol. 7

Raja, R., Mukherjee, I. & Sarkar, B. K. (2021).A Machine Learning-Based Prediction Model for Preterm Birth in Rural India, *J Healthc Eng,* vol. 2021

Rocha, T. A. H., de Thomaz, E. B. A. F., de Almeida, D. G., da Silva, N. C., Queiroz, R. C. d. S., Andrade, L., Facchini, L. A., Sartori, M. L. L., Costa, D. B., Campos, M. A. G., da Silva, A. A. M., Staton, C. & Vissoci, J. R. N. (2021).Data-Driven Risk Stratification for Preterm Birth in Brazil: A Population-Based Study to Develop of a Machine Learning Risk Assessment Approach, *The Lancet Regional Health - Americas,* vol. 3**,** pp. 100053

Sagi, O. & Rokach, L. (2018).Ensemble Learning: A Survey, *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery,* vol. 8, no. 4

scikit-learn developers. (2022). *Linear Models - 1.11 Logistic Regression* [Online]. Available online: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression [Accessed 4 May 2022]

Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. (2018).Deep Ehr: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (Ehr) Analysis, *IEEE J Biomed Health Inform,* vol. 22, no. 5**,** pp. 1589-1604

Simms, V., Gilmore, C., Cragg, L., Marlow, N., Wolke, D. & Johnson, S. (2013).Mathematics Difficulties in Extremely Preterm Children: Evidence of a Specific Deficit in Basic Mathematics Processing, *Pediatr Res,* vol. 73, no. 2**,** pp. 236-44

The pandas development team. (2021). Pandas-Dev/Pandas: Pandas. 1.2.4 ed.: Zenodo.

Van Rossum, G. & Drake, F. L. (2009). Python 3 Reference Manual: CreateSpace

Vogel, J. P., Chawanpaiboon, S., Moller, A. B., Watananirun, K., Bonet, M. & Lumbiganon, P. (2018).The Global Epidemiology of Preterm Birth, *Best Pract Res Clin Obstet Gynaecol,* vol. 52**,** pp. 3-12

Weber, A., Darmstadt, G. L., Gruber, S., Foeller, M. E., Carmichael, S. L., Stevenson, D. K. & Shaw, G. M. (2018).Application of Machine-Learning to Predict Early Spontaneous Preterm Birth among Nulliparous Non-Hispanic Black and White Women, *Annals of Epidemiology,* vol. 28, no. 11**,** pp. 783-789

Wlodarczyk, T., Plotka, S., Szczepanski, T., Rokita, P., Sochacki-Wojcicka, N., Wojcicki, J., Lipa, M. & Trzcinski, T. (2021).Machine Learning Methods for Preterm Birth Prediction: A Review, *Electronics,* vol. 10, no. 5

Wolke, D., Chernova, J., Eryigit-Madzwamuse, S., Samara, M., Zwierzynska, K. & Petrou, S. (2013).Self and Parent Perspectives on Health-Related Quality of Life of Adolescents Born Very Preterm, *J Pediatr,* vol. 163, no. 4**,** pp. 1020-6 e2

World Health Organization (2018). *Preterm Birth* [Online]. Available online: https://www.who.int/news-room/fact-sheets/detail/preterm-birth [Accessed 29 April 2022]

xgboost developers. (2021). *Xgboost Parameters* [Online]. Available online: https://xgboost.readthedocs.io/en/latest/parameter.html [Accessed 4 May 2022]

Zambrana, I. M., Vollrath, M. E., Jacobsson, B., Sengpiel, V. & Ystrom, E. (2021).Preterm Birth and Risk for Language Delays before School Entry: A Sibling-Control Study, *Dev Psychopathol,* vol. 33, no. 1**,** pp. 47-52

Appendix A: summary statistics

	Data Type	Full Term	Preterm
Age	Numerical	28.37 (5.74)	28.30 (6.03)
Young mother (age < 20)	Boolean	5.77%	6.99%
Old mother (age >= 35)	Boolean	14.82%	16.63%
US born	Boolean	75.91%	80.09%
Residence status	Categorical		
Resident		72.28%	68.00%
Intrastate nonresident		25.34%	29.02%
Interstate nonresident		2.08%	2.83%
foreign resident		0.30%	0.16%
Race	Categorical		
White		73.67%	67.82%
Black		14.89%	21.69%
AIAN		0.96%	1.10%
Asian		7.41%	6.12%
NHOPI		0.33%	0.35%
More than one		2.74%	2.91%
Hispanic origin	Categorical		
Non-Hispanic		74.38%	75.79%
Mexican		15.12%	13.99%
Puerto Rican		1.85%	2.10%
Cuban		0.52%	0.49%
Central and South American		4.42%	3.89%
Other Hispanic		3.71%	3.75%
Marital Status			
Married	Boolean	53.73%	45.95%
Unmarried	Boolean	35.23%	45.58%
Information unknown	Boolean	11.04%	8.47%
Mother's education	Categorical		
8th grade or less		4.00%	3.38%
9th - 12th grade		9.76%	12.96%
High school graduate		25.49%	29.82%
Some college credit		19.30%	20.75%
Associate degree		7.88%	7.84%
Bachelor's degree		21.40%	15.98%
Master's degree		9.71%	7.18%
Doctorate degree		2.84%	2.09%
Low education	Boolean	38.87%	46.16%
High education	Boolean	12.55%	9.26%
Paternity Acknowledged			
Yes	Boolean	79.19%	75.64%
No	Boolean	9.68%	15.75%
Information unknown	Boolean	11.13%	8.62%
Father's age available	Boolean	89.00%	82.45%
Payment method	Categorical		
Medicaid		41.45%	49.88%
private insurance		49.12%	42.29%
self-pay		5.51%	4.11%
other		3.92%	3.71%

	Data Type	Full Term	Preterm
WIC	Boolean	35.68%	40.00%
HEALTH INFORMATION			
Start of prenatal care			
No prenatal care	Boolean	1.59%	5.10%
Started in first trimester	Boolean	77.08%	75.03%
Started in second trimester	Boolean	16.70%	16.10%
BMI			
BMI underweight	Boolean	3.83%	4.82%
BMI overweight	Boolean	39.32%	40.10%
BMI obese	Boolean	8.35%	12.06%
Height		64.13	63.86
		(2.82)	(2.84)
Cigarettes			
Cigarettes before (daily)	Numerical	0.90	1.59
		(4.26)	(5.63)
Cigarettes 1st trimester (daily)	Numerical	0.55	1.08
		(3.03)	(4.27)
Cigarettes 2nd trimester (daily)	Numerical	0.41	0.85
		(2.49)	(3.66)
Cigarettes (yes/no)		5.35%	10.05%
RISK FACTORS	All: boolean		
Pre-pregnancy diabetes		0.34%	1.53%
Gestational diabetes		4.29%	7.54%
Pre-pregnancy hypertension		0.73%	2.45%
Gestational hypertension		2.95%	7.20%
Hypertension Eclampsia		0.10%	0.41%
Previous preterm Birth		2.34%	10.28%
Infertility treatment used		1.04%	2.42%
Fertility Enhancing Drugs		0.52%	1.08%
Asst. Reproductive technology		0.60%	1.53%
Previous Cesarean		4.18%	5.84%
Number of cesareans		0.05	0.07
		(0.25)	(0.33)
INFECTIONS	All: boolean		
Gonorrhea		0.29%	0.53%
Syphilis		0.10%	0.19%
Chlamydia		1.88%	2.61%
Hepatitis B		0.22%	0.22%
Hepatitis C		0.33%	0.89%
Infant is female	Boolean	49.05%	45.09%
PREVIOUS PREGNANCIES			
Prior alive	Numerical	1.07	1.22
		(1.31)	(1.46)
Prior dead	Numerical	0.01	0.02
		(0.17)	(0.22)
Prior other pregnancies	Numerical	0.37	0.48
		(0.81)	(0.99)
First birth	Boolean	41.59%	39.75%
First pregnancy	Boolean	34.61%	31.64%

	Data Type	Full Term	Preterm
Birth interval	Categorical		
0-3 months (plural delivery)		0.18%	4.37%
4-11 months		0.78%	2.65%
12 to 17 months		5.01%	6.87%
18 to 23 months		8.18%	6.92%
24 to 35 months		14.96%	10.81%
36 to 47 months		8.95%	7.15%
48 to 59 months		5.78%	5.15%
60 to 71 months		3.95%	3.76%
72 months and over		10.62%	12.57%
Not applicable (first birth)		41.59%	39.75%