The use of machine learning to predict adverse birth outcomes:

Empirical real world evidence from a human cohort study in Adama, Ethiopia

Stephanie Bol

Lund, January 2023

LUND
UNIVERSITY

Master's Thesis in
Biomedical Engineering

Faculty of Engineering, LTH
Department of Biomedical Engineering

Supervisors: Anna Oudin and Andreas Jakobsson

The use of machine learning to predict adverse birth outcomes: Empirical real world evidence from a human cohort study in Adama, Ethiopia

Author
Stephanie Bol

Figures
Created by the author if nothing else is indicated

Lunds Universitet
Institutionen för Biomedicinsk Teknik
Box 118
SE-221 00 Lund
Sverige

E-husets tryckeri
Lund 2023

Acknowledgements

Together with the faculty of medicine and the faculty of biomedical engineering, this master's thesis was written during the autumn semester of 2022 as the final work of my degree in Biomedical Engineering. I would like to thank the group at the faculty of medicine for letting me write this master's thesis with you and providing the exciting data you collected. A special thanks to my supervisor Anna Oudin who has been welcoming and helping since day one. I would also like to thank my supervisor at LTH, Andreas Jakobsson, for the guidance during the project. Lastly, thank you to Erin Flanagan for taking the time to explain all the features of the dataset to me.

Stephanie Bol

Abstract

Since Ethiopia has a high number of recorded adverse birth outcomes, the city of Adama was subjected to a study (Flanagan et al., 2022) that gathered data from 2085 pregnancies. This thesis utilizes that data to investigate the usage of machine learning in environmental epidemiology. Using the classification methods logistic regression, random forest, support vector classifier, and k-nearest neighbors, two different sampling methods were implemented to handle the imbalanced dataset. The original imbalanced dataset performed worst though similar to the undersampled dataset. Oversampling the dataset with SMOTE yielded the best result with the random forest classifier and had an AUC score of 0.72 and an f1-score of 0.85. With further work, more data, and higher evaluation scores, machine learning may be a way to implement preventable medicine in Adama, Ethiopia. However, more research is needed, especially with larger study populations, to improve the accuracy of these models and find the most important features to analyze for this region.

List of acronyms & abbreviations

AUC - Area under the ROC curve

FN - False negative

FP - False positive

FPR - False positive rate

KNN - k-nearest neighbors

LR - Logistic regression

ML - Machine learning

RF - Random forest

ROC - Receiver operating characteristic curve

SMOTE - Synthetic minority over-sampling technique

SVC - Support vector classifier

TN - True negative

TP - True positive

TPR - True positive rate

Contents

Chapter 1

Introduction

Substantial progress in improving the likelihood that a child will survive a pregnancy has been made over the past twenty years. However, there are still an estimated 2 million stillbirths occurring each year (Hug et al., 2020) and 2.4 million neonatal deaths within the first month of life were reported in 2020 (World Health Organization, 2022b). This is a prominent issue in low and lower-middle-income countries where 84% of all stillbirths occur, especially in Sub-Saharan Africa. There, the number of adverse birth outcomes has increased despite the worldwide decline (Hug et al., 2020). Most of these cases are mainly preventable, however, the cause of death is often not recorded which aggravates the process of finding the key factors in the regions (Aminu et al., 2014). Although surveys can help to get a better understanding of child mortality, the quality of data poses an issue (Hug et al., 2020).

Previous studies that examine the use of machine learning algorithms such as logistic regression for the prediction of stillbirths have been made on various data (Koivu and Sairanen 2020; Malacova et al. 2020). However, not many studies can be found using machine learning on data from Ethiopia which is particularly interesting due to it being one of the regions where the number of adverse birth outcomes is increasing (Hug et al., 2020). The dataset in this study has previously been used to investigate the possible correlation between adverse birth outcomes and ambient and indoor air pollution exposure in Adama, Ethiopia. After using statistical measures such as binary logistic regression, a tendency was seen between the factors but not statistical significant (Flanagan et al., 2022).

1.1 Purpose of the thesis

Due to the emergence of machine learning in preventive medicine, it was chosen as the method to investigate its use in an empirical dataset with real-world data. Adverse birth outcomes, including stillbirth and neonatal death, were selected as the target for the analysis. Six different living conditions were used as features. The aim was to test the viability of building a model using logistic regression, random forest, support vector classifier, and k-nearest neighbors to accurately predict the probability of adverse birth outcomes in Adama, Ethiopia. Due to the imbalanced nature of the dataset, it was undersampled and oversampled to investigate if it improved the evaluation metrics.

Logistic regression was the primary choice for classification since binary logistic regression was used in the study that previously processed this dataset without machine learning. The three other classifiers were implemented to create a broader comparison between the different techniques.

1.2 Report framework

The report is divided into the following parts:

- **Chapter 2: Background** To familiarise the reader with the subject, background information regarding adverse birth outcomes and machine learning with an imbalanced dataset will be presented.

- **Chapter 3: Method** The process of study setting, data processing, feature analysis, and machine learning are described.

- **Chapter 4: Results** Measured values from each sampling technique are presented with tables and graphs.

- **Chapter 5 & 6: Discussion and Conclusion** The results are discussed in addition to the optimal sampling technique for imbalanced epidemiological data. Future work and possible sources of error will be addressed.

Chapter 2

Background

2.1 Adverse birth outcomes

In this thesis, stillbirth (pregnancy losses after 22 weeks of gestation (Hug et al., 2020)) and neonatal death (deaths during the first 28 days among live births (World Health Organization, 2022a)) are referred to as adverse birth outcomes.

As mentioned in the introduction, many cases of adverse birth outcomes are predominantly preventable with better antenatal screening. Although there is an insufficient recording of the cause of death in many of the worst-affected regions, some trends have been noticed. A systematic review of 142 studies (Aminu et al., 2014) showed that the most common maternal factors reported as a cause of stillbirth were the following: Sexually transmitted diseases like syphilis and positive HIV in addition to diabetes and high blood pressure. In developing countries such as Ethiopia, the most common factors were a high maternal age (35 or above), parity (the number of births from a woman regardless of the health status of the child), gestational age at birth, and birth weight (Aminu et al., 2014). Other factors reported were the lack of education, socioeconomic characteristics, place of residence, lack of antenatal care, and previous stillbirth (Bhusal et al., 2019).

A noted tendency in studies on adverse birth outcomes in developed countries is that more attention and research are needed to better understand the underlying factors (McClure et al. 2006; Flanagan et al. 2022).

2.2 Machine learning

Using machine learning, a system that is given an input can be trained to produce answers on new data (Chollet, 2021). The tasks these machine learning algorithms are set to do are based on the idea that they will learn from the assignment and use it to improve their task execution. After training and testing the algorithm, predictions and decisions can be made. The more data the algorithm is given to train on, the more experience it gets that expectantly will make more accurate predictions (Ray, 2019). These predictions are tested using evaluation metrics that will be explained in section 2.4.2. In this study, four different classification models were used and tested. Logistic regression, support vector classifier, k-nearest neighbor, and random forest.

2.2.1 Logistic regression

Logistic regression (LR) is a model that has long been used for classification problems. Its simplicity and versatility have made it a good starter algorithm for getting to know the dataset (Chollet, 2021). Being a predicting method, logistic regression can be used to anticipate the probability of an event occurring with a set of input variables. It has the advantages of being computationally efficient and not easily affected by small noise. However, it is prone to overfitting in addition to only having a linear decision surface (Liu, 2011).

2.2.2 Support vector classifier

The support vector classifier (SVC) is a robust and popular method due to its simplicity, although it is mathematically complex and computationally expensive (Awad and Khanna, 2015). In this classifier, hyperplanes are defined as the decision boundary while kernels can be used to separate objects of different classes. The method uses generalization that reduces the probability of overfitting. Another advantage is that the SVC can handle contrasting data structures with the proper kernel. However, it can be hard to find the correct kernel. Furthermore, this method is lacking when used with a large dataset since the computation time increases and it does not provide probability estimates (Liu, 2011).

2.2.3 Random forest classifier

A classifier often used to handle imbalanced datasets is the random forest (RF) classifier. It employs decision trees which is a method of continuously splitting data defined as the nodes and the decisions as the leaves. A RF classifier uses multiple decision trees during the training stage to determine the class that the majority of the trees choose. It is executed with the help of a technique called bagging. During training, a random sample is repeatedly selected to fit the tree to the samples. This method does not have the same risk of overfitting as a normal decision tree considering the voting strategy which is a great advantage (Caie et al., 2021). Moreover, it is an efficient algorithm due to it being a tree traversal algorithm (Ray, 2019). However, the stochastic nature of tree traversal is also one of its disadvantages. The final trained model has a degree of unpredictability, making it hard to follow the path of decision (Caie et al., 2021).

2.2.4 k-nearest neighbors

k-nearest neighbors (KNN) is a simple and fast classifier since the majority of the work is performed during classification. The predictions are calculated for each instance based on the classes of k nearest neighbors that were found in the case library which holds the training dataset. The nearest neighbors are based on the shortest distance regarding the feature space with the classified instance, usually measured in Euclidean distance (Khoshgoftaar et al., 2007). One of the disadvantages of KNN is that the method becomes computationally intensive and noisy with irrelevant features that will impair the accuracy. This is mainly due to the distance computation of the nearest neighbors that are needed to classify unknown records. Considering that the method stores the training data and therefore deals with a large dataset, the computation becomes expensive (Ray, 2019).

2.2.5 Supervised learning

The most common case of machine learning today is supervised learning where targets (such as ”stillbirth” or ”not stillbirth”) are known. These targets are labeled and used when the artificial system is trying to map the input data to the targets (Chollet, 2021). A possible use of supervised learning is to have an artificial system that can predict the output when given new inputs it has never encountered before (Liu, 2011). An example could be teaching an algorithm to differentiate between pictures of apples and pears. In supervised learning, each picture is labeled as an apple or a pear, unlike unsupervised learning where the pictures are shown exclusively (Chollet, 2021).

Collecting labeled data needed for supervised learning is one of the disadvantages of the method due to the difficulty and expense of gathering enough quantity suitable for machine learning. Although these factors are important to consider, this method opens up the possibility to learn human behavior and the impact environmental factors can have. Compared to traditional statistical methods, supervised learning can perform analysis in a faster and more accurate way in some cases. However, it is still limited by the current hardware and algorithm designs (Liu, 2011).

2.3 Rare events data

Rare events data, or imbalanced data, is a common problem when working with epidemiology (King and Zeng, 2001). An example can be that the number of stillbirths occurring in a dataset compromises 5.5% whereas normal births are the vast majority at 94.5%. If not taken into account, this can be a problem. A strong bias is often shown toward the larger class, thus increasing the likelihood of error rates such as a large number of false negatives (FN). A result of this is that many of the data points from the minority class will be classified as the majority. It can therefore be an advantage to generate a more balanced dataset to combat these issues (Poolsawad et al., 2014).

2.3.1 Undersampling and Oversampling

A way to handle an imbalanced dataset is to undersample it. In other words, creating equally frequent classes by reducing the majority class to the same number of samples as the minority (see Figure 2.1). A simple method used to undersample is to apply random elimination in order to balance the distribution of the target. The technique is fast and works well with a large dataset (Mohammed et al., 2020). However, this may remove essential data points which could be crucial for the classification and possibly create a bias (Liu et al., 2008).

Another way to resample the data is oversampling, which is also illustrated in Figure 2.1. It involves increasing the frequency of the minority class to the same number as the majority. A common technique for oversampling is the synthetic minority over-sampling technique (SMOTE) which generates synthetic examples of the minority class. This is achieved by composing artificial cases of k-nearest class neighbors generated at random along the lines connecting the minority sample and its chosen amount of neighbors (Poolsawad et al., 2014). A disadvantage of this method is the possibility of it generating noisy data. Finding a fitting variation of SMOTE can combat the issue (Douzas and Bacao, 2017).

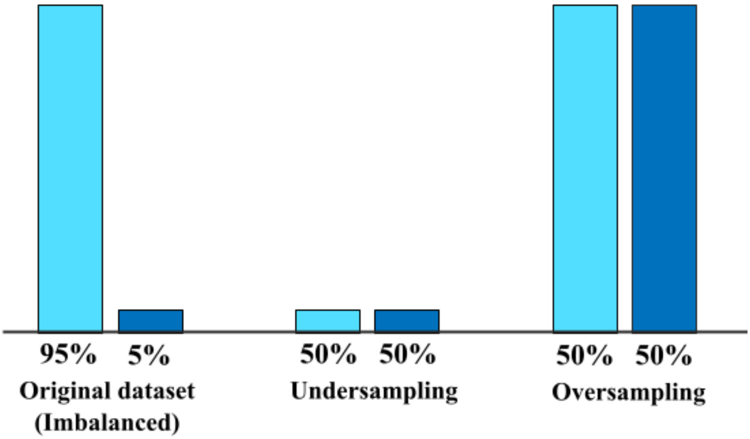


Figure 2.1: An illustration of how an imbalanced dataset can be resampled with two different methods, undersampling and oversampling.

2.4 Validation

2.4.1 k-fold cross validation

A recommended validation method for a small dataset is k-fold cross-validation. It works by splitting the data into a set number of partitions as in Figure 2.2. These are usually set to 5 and are identical. However, each split is divided into test- and training data where the partition is different each time. This splitting technique ensures that a smaller dataset can provide more data to train on. After the training of the splits, the final evaluation is completed on test data that the algorithm has not encountered before to evaluate the model (Chollet, 2021). A disadvantage of using k-fold cross-validation on imbalanced data is the chance of the validation set only containing samples from the majority class. This issue is solved using stratified sampling which ensures that the class proportions in the subset are equal to the proportions in the learning set. Therefore, samples from both the majority and minority will be present in each split (Berrar, 2019). It has been recommended to use stratified 10-fold cross-validation when processing real-world datasets (Kohavi, 1995).

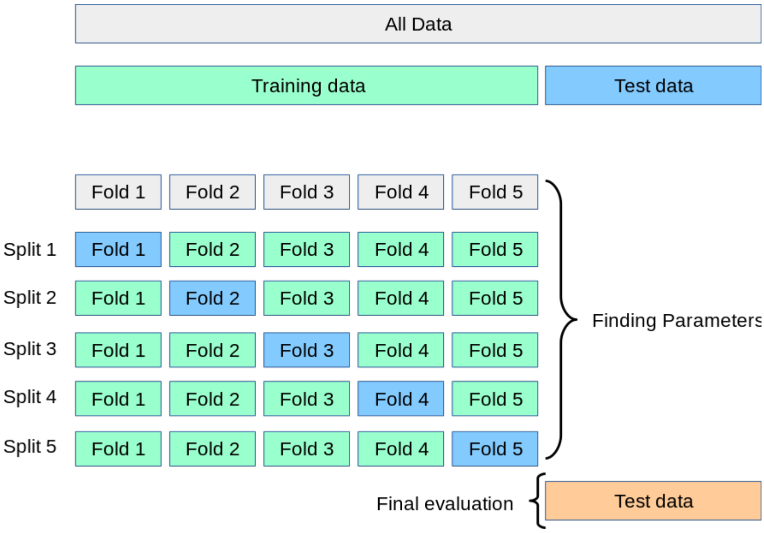


Figure 2.2: An illustration of k-fold cross validation with five partitions (Pedregosa et al., 2011).

2.4.2 Evaluation metrics

Different evaluation metrics can be used to measure the performance of an artificial system. Accuracy is a commonly used method, however, it is important to use it carefully. Some classifiers, such as logistic regression may underestimate the probability of uncommon events sharply, which can result in a bad prediction but a high accuracy (King and Zeng, 2001). Accuracy is calculated by dividing the number of correct predictions by the predicted number. Hence, if the artificial system is unable to accurately predict any of the minority classes, the score will still be high. The accuracy measurement is accordingly not favorable when working with an imbalanced dataset (He and Garcia, 2009).

Therefore, evaluation metrics such as the receiver operating characteristic curve (ROC) and area under the ROC curve (AUC) are better suited than accuracy for these cases (Mohammed et al., 2020). They are known as a good indicator of classifier performance and visual representation. ROC illustrates the performance of a model by plotting the two parameters true positive rate (TPR) and false positive rate (FPR). The calculation for TPR can be seen in Equation 2.1 where TP is the number of true positive and FN the false negative. TPR is also the same as the metric recall. The equation for FPR (see Equation 2.2) has false positive (FP) and true negative (TN) (He and Garcia, 2009). AUC on the other hand reflects the performance of the ROC curve and is useful for comparisons. A high AUC score indicates that the model performs well while a score close to 0.5 is no better than random (Huang and Ling, 2005).

$$TPR = \frac{TP}{TP + FN} = Recall \tag{2.1}$$

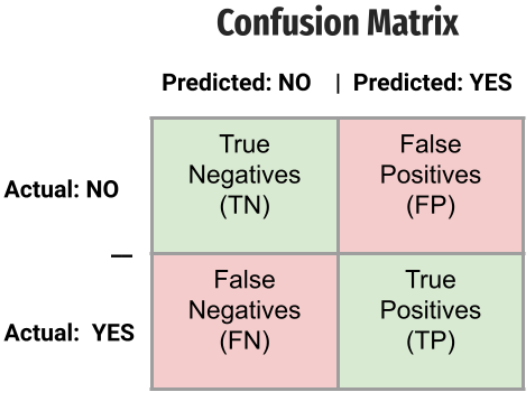
$$FPR = \frac{FP}{FP + TN} \tag{2.2}$$

ROC and AUC have their limitations when dealing with an imbalanced dataset considering it may overestimate the performance of the artificial system. Thus, it is advisable to supplement ROC with other evaluation metrics such as f1, recall, and precision. The evaluation metric recall is the same as TPR which can be seen in Equation 2.1 and it measures how well the algorithm classified the positive class correctly. Precision on the other hand measures how exact the algorithm is by looking at the positive labels and the ratio of them being labeled correctly, see Equation 2.3 (He and Garcia, 2009). Using these two metrics, f1 (see Equation 2.4) can be calculated which is a measure of how effective the classification is and can give more insight than other metrics such as accuracy (Sasaki et al., 2007). The combination of precision and recall can adequately evaluate the performance of the classifiers when dealing with an imbalanced dataset (He and Garcia, 2009).

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$f1 = \frac{2 \cdot (Precision \cdot Recall)}{(Precision + Recall} \tag{2.4}$$

The performance of the classification can be illustrated with a confusion matrix (see Figure 2.3). The amount of TP, FP, FN, and TN is clearly displayed in the matrix after validation. It is therefore an easy way to get an overview of the performance of the algorithm. It can be seen as a compliment to the other evaluation metrics (Zeng, 2020).



Confusion Matrix

	Predicted: NO	Predicted: YES
Actual: NO	True Negatives (TN)	False Positives (FP)
Actual: YES	False Negatives (FN)	True Positives (TP)

Figure 2.3: A model of a confusion matrix that illustrates its use and parameters.

Chapter 3

Method

3.1 Overview

The objective of the thesis was to investigate the use of machine learning on collected data from Adama, Ethiopia, with the challenge of a relatively small and imbalanced dataset with a high percentage of missing values. The workflow was divided into four parts which can be seen below:

- Study setting
- Data processing
- Feature analysis
- Machine learning

3.2 Study setting

The data used for this thesis were gathered from a study based in Adama, Ethiopia. The city had approximately 214 000 inhabitants when the data was collected (Flanagan et al., 2022). Adama is exposed to high emissions partly due to heavy traffic passing by the city center in addition to inadequate sustainable solid waste management (Hailemariam and Ajeme, 2014).

The cohort (ClinicalTrials.gov identifier number NCT03305991) was comprised of 2085 pregnancies that were recruited from November 2015 to February 2018. Re-pregnancies (n = 124) were excluded from this study. Through questionnaires, data regarding their socioeconomic, demographic, medical -and obstetric history were obtained. Three physical examinations were conducted for each pregnancy. In the cohort, 1616 women could be reached for a postnatal evaluation on-site or by phone to conclude the outcome of the pregnancy, adverse birth outcomes or not. This included miscarriage, neonatal death, and stillbirth (Flanagan et al., 2022). Only 1616 women with verified pregnancy outcomes were used in this study to ensure a reliable analysis. The two binary variables containing stillbirth (n = 69) and neonatal death (n = 16) were combined to create more data points for the target variable. However, the imbalance was still prevalent where adverse birth outcomes only consisted of 5.5% in the dataset.

3.3 Data processing

Using the statistical analysis software SPSS, the initial data analysis was executed. Due to missing labels on the majority of the variables, a consultation with the scientists who previously worked with the data was made. It resulted in a greater understanding and an elimination of 82 uncertain and non-relevant labels to ensure a credible result. An analysis of the missing pattern of the other values was then made. All variables with more than 30% missing values were detected and deleted. Imputation of these variables was deemed too uncertain. The remaining data were then polished, meaning that unclear data points were deleted and strings were made to numeric or scalar values.

3.4 Feature analysis

The program SPSS could not handle a large amount of data when executing multiple imputations which is an iterative form of stochastic imputation (Schafer, 1999). Therefore, binary logistic regression was calculated with the goal to remove all variables that scored a p-value higher than 0.2. This threshold was set in consideration of a large number of missing variables. In the case of similarities, for example, multiple features regarding education divided into two, three, or four categories. In such cases, the one with the lowest p-value was chosen. It resulted in the following six variables (see Table 3.1) that were used in the subsequent stages:

Table 3.1: The six selected feature variables and their respective values used in the machine learning algorithms.

Variables	**Values**
Hospitalization after birth	0 = No, 1 = Yes
Place of delivery	1 = Adama regional hospital, 2 = Adama health center, 3 = Gedda health center, 4 = Other facilities, 5 = Home delivery
First pregnancy	0 = No, 1 = Yes
Age	Scalar age between 13 - 40 years old
Level of education	1 = Illiterate and < 6 grades, 2 = 6 - 12 grades, 3 = Higher education
Living situation	0 = Has permanent residence, 1 = No permanent residence

3.4.1 Imputation

With the selection of the six feature variables, multiple imputation could be executed. It resulted in 1615 complete rows of data with seven columns (the target variable included).

3.5 Machine learning

Using the cloud service Google Colaboratory with the programming language Python, the machine learning (ML) stage was initiated. Adopting the pandas and sklearn library, the necessary tools for ML and analysis were used. After reading the file into the program, the dataset was split into features and targets. The target variable was adverse birth outcome (either stillborn death or neonatal death), whereas the features were composed of the remaining variables seen in Table 3.1. To form the data into a consistent format, data standardization was used and applied to the features. The four classifiers logistic regression (LR), random forest (RF), support vector classifier (SVC), and k-nearest neighbors (KNN) were used to evaluate the different performances in the models. The respective hyperparameters that control the learning process can be seen in Table 3.2.

Table 3.2: Hyperparameters for the four classifiers.

Classifiers	**Hyperparameters**
Logistic regression	max iteration = 1000

Random forest	n_estimators = 150
	random_state = 0

Support vector classifier	kernel = linear
	probability = True

k-nearest neighbors	default hyperparameters

3.5.1 Undersampling and Oversampling

To sample the dataset to reduce the majority class to the same size as the minority class (see Figure 2.1 for an illustration), 85 rows from the majority class were randomly chosen. The rest were removed. This resulted in a total of 170 rows.

The oversampling was built by importing imbalanced-learn's SMOTE function (Lemaître et al., 2017). The minority class was resampled to the same size as the majority through the creation of synthetic examples. The majority class consisted of 1530 data points, and the minority class originally had the size of 85 rows. Subsequently, the combined total became 3060 after oversampling. Using the newly sampled dataset, the models were built with the classifiers.

3.5.2 Test and train split

All the classification models were split into a train and test set through the tools from sklearn. The test size was set to 10% of the dataset and the function Stratify was used to ensure that the same proportions as observed in the original dataset were preserved in the split.

3.5.3 k-fold cross validation

The k-fold cross-validation model was created with ten splits, three repeats, and a random state set to 1. Splits were made stratified, meaning that the percentages of the different classes remained the same in each division. This was applied to all four classifiers, respectively.

3.5.4 Model evaluation

To observe the performance of the model in various constellations of sampling techniques and classifiers, different model evaluations were used. Firstly, the accuracy score was calculated using sklearn's accuracy function. F1, recall, precision, and AUC score were also calculated with sklearn. The mean of all the scores after the cross-validation was presented. Lastly, graphs displaying the confusion matrix as well as the ROC curve were plotted for each instance.

Chapter 4

Results

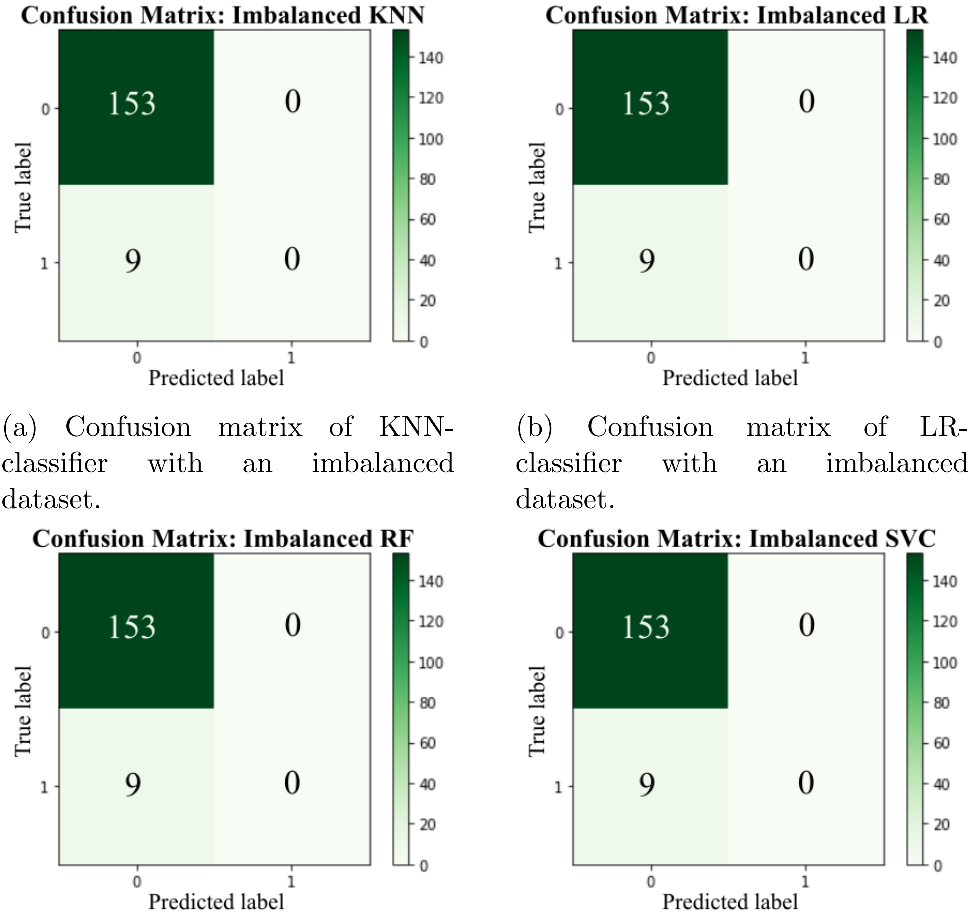
The results from the different machine learning algorithms were divided into three categories, the original and imbalanced dataset, the undersampled dataset, and lastly, the oversampled dataset. The same six features (see Table 3.1) were used for all the techniques. Each method was trained with the four classifiers, logistic regression, support vector classifier, random forest, and k-nearest neighbors. These were then evaluated with the evaluation metrics f1, recall, precision, AUC, and accuracy. Confusion matrices were plotted to further illustrate the result and balance between TP, TN, FP, and FN. For the interested reader, the ROC curves used to calculate the AUC scores can be seen in Table A.1 in the Appendix.

4.1 Imbalanced dataset

This dataset was the original without any sampling done. The target class had an imbalance of 94.5% for the majority class and 5.5% for the minority. The final evaluation scores after training, testing, and validation can be seen in Table 4.1 where f1, recall, precision, AUC score, and accuracy are displayed. In Figure 4.1 the four classifiers' respective confusion matrices are shown.

Table 4.1: Evaluation metrics for the four classifiers with the original imbalanced dataset.

Models	**F1**	**Recall**	**Precision**	**AUC**	**Accuracy**
LR	0	0	0	0.59	0.94
SVC	0	0	0	0.43	0.94
RF	0	0	0	0.47	0.94
KNN	0	0	0	0.48	0.94



(a) Confusion matrix of KNN-classifier with an imbalanced dataset.

(b) Confusion matrix of LR-classifier with an imbalanced dataset.

(c) Confusion matrix of RF with an imbalanced dataset.

(d) Confusion matrix of SVC with an imbalanced dataset.

Figure 4.1: Four confusion matrices with the respective classification model. See Figure 2.3 for further explanation of confusion matrices.

4.2 Undersampling

After undersampling the original imbalanced dataset, there were 170 entries remaining. Half the dataset was labeled as adverse birth outcomes, and the other 50% was marked as a healthy pregnancy. The result of training on the undersampled dataset for the four classifiers are visible in Table 4.2 as well as their confusion matrices in Figure 4.2.

Table 4.2: Evaluation metrics for the four classifiers with an undersampled dataset.

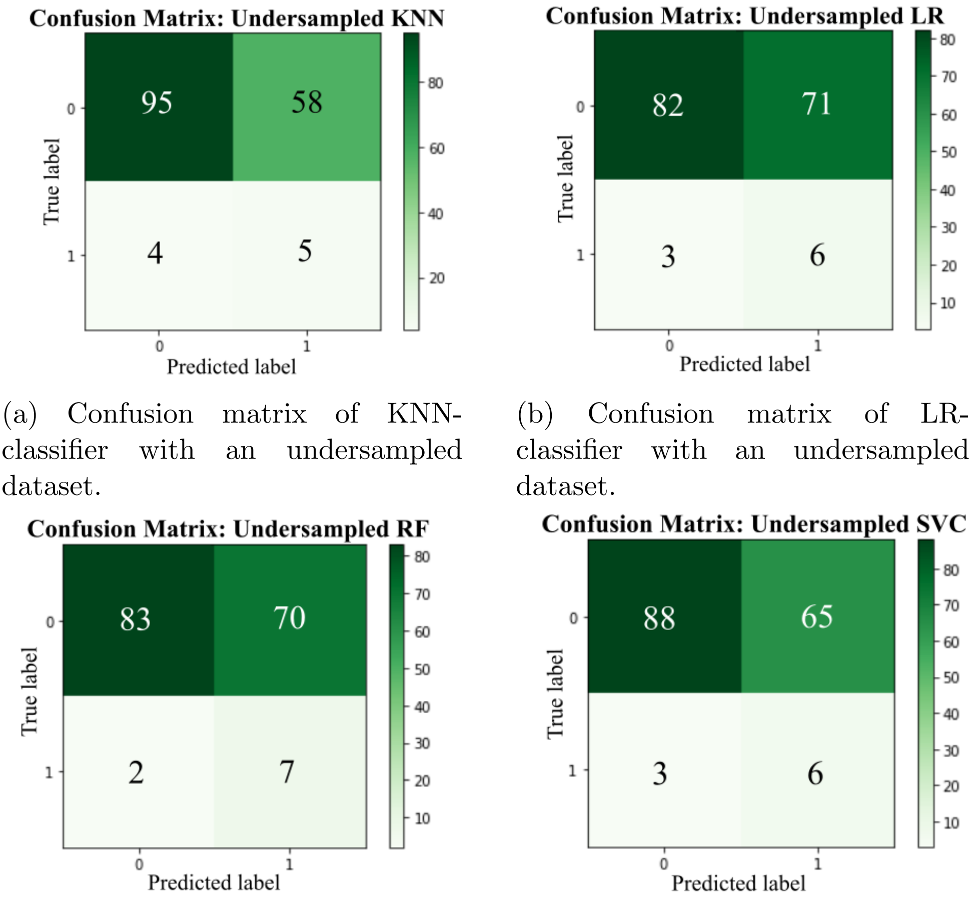
Models	**F1**	**Recall**	**Precision**	**AUC**	**Accuracy**
LR	0.54	0.52	0.56	0.60	0.54
SVC	0.51	0.50	0.54	0.62	0.58
RF	0.51	0.52	0.51	0.66	0.56
KNN	0.44	0.39	0.53	0.59	0.62

4.3 Oversampling

After oversampling the original dataset, it contained 3060 entries and had a balanced target class. The resulting evaluation scores can be seen in Table 4.3 in addition to the confusion matrices in Figure 4.3.

Table 4.3: Evaluation metrics for the four different classifiers with an oversampled dataset. The best-performing classifier, RF, is marked in bold for clarification.

Models	**F1**	**Recall**	**Precision**	**AUC**	**Accuracy**
LR	0.68	0.70	0.65	0.60	0.64
SVC	0.70	0.80	0.63	0.56	0.57
RF	**0.85**	**0.87**	**0.83**	**0.72**	**0.86**
KNN	0.71	0.62	0.84	0.51	0.77



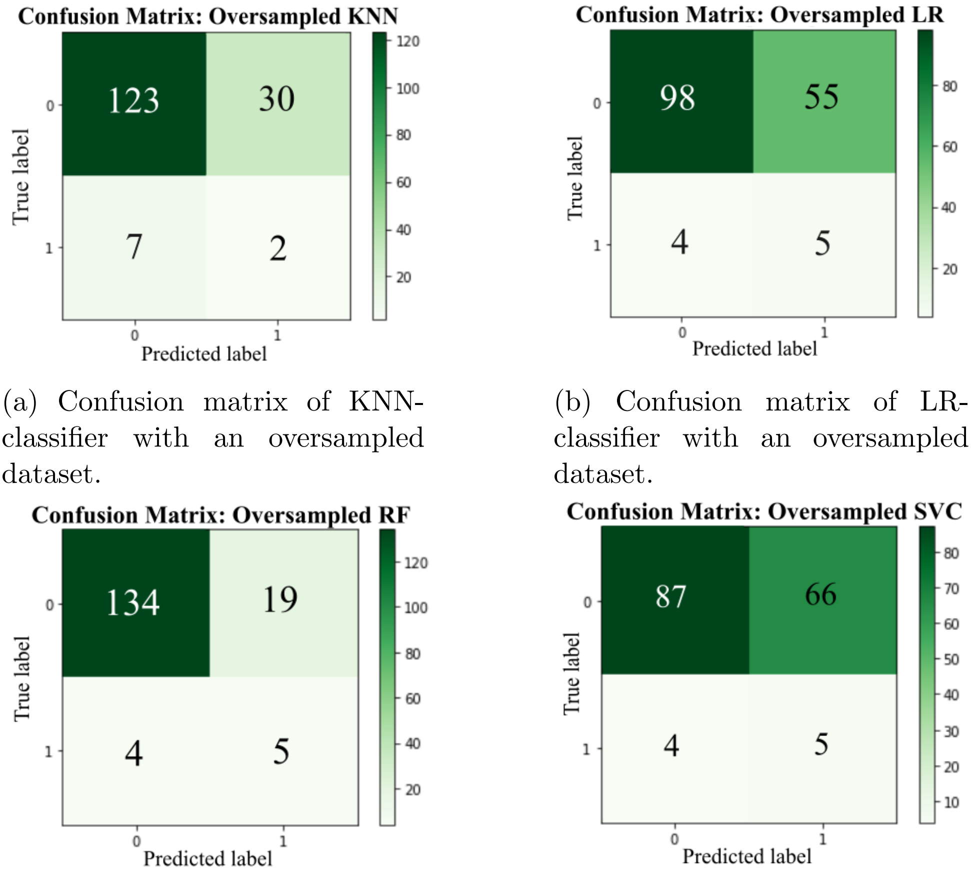
(a) Confusion matrix of KNN-classifier with an undersampled dataset.

(b) Confusion matrix of LR-classifier with an undersampled dataset.

(c) Confusion matrix of RF with an undersampled dataset.

(d) Confusion matrix of SVC with an undersampled dataset.

Figure 4.2: Four confusion matrices with the respective classification model on a undersampled dataset.



(a) Confusion matrix of KNN-classifier with an oversampled dataset.

(b) Confusion matrix of LR-classifier with an oversampled dataset.

(c) Confusion matrix of RF with an oversampled dataset.

(d) Confusion matrix of SVC with an oversampled dataset.

Figure 4.3: Four confusion matrices with the respective classification model on an oversampled dataset.

Chapter 5

Discussion

5.1 Imbalanced dataset

The imbalanced dataset had 94.5% of the target variable labeled as a normal pregnancy, and the other 5.5% was composed of adverse birth outcomes. In the result, there is a distinct difference between the metrics (see Table 4.1). Only by studying the accuracy, the result would look promising. However, as earlier mentioned in the background, accuracy gives a skewed outcome when given an imbalanced dataset (He and Garcia, 2009). It is apparent in Figure 4.1 that all the classification models have classified the test data as the majority class. This has resulted in false predictions of all the adverse birth outcomes as normal pregnancies. It is not surprising that the accuracy still is high since it takes the sum of the true predictions and divides it by the total number of predictions. It is unacceptable from a clinical point of view where the rare class often is the important one. If this evaluation criterion were to be used in this type of setting, it could lead to a high risk of misdiagnosing (Mazurowski et al., 2008).

Looking at the AUC score in Table 4.1, all the classifiers are close to 0.5 in score. This indicates that the models were nearly equal to random guessing (Lalkhen and McCluskey, 2008). Further, the score for the other evaluation metrics, f1, recall, and precision, in Table 4.1 were all approximately zero. Since all of them have TP as a nominator (see Equations 2.1 and 2.3), the values being zero is as expected since TP was zero (see Figure 4.1).

5.2 Undersampling

Reducing the majority class through random elimination to the same size as the minority, the size of the dataset decreased from 1615 data points to 170. This undersampling led to a distinct reduction of the size which may be the reason for the poor performance of all the classifiers. Even though the dataset was balanced, all the evaluation metrics are close to 0.5 (see Table 4.2). Instead of being at zero, the f1, recall, and precision metrics were closer to random. In that sense, undersampling is slightly better than the imbalanced dataset, however, the result is still very bad. Therefore, the algorithms used are not able to confidently predict if a person with the given features will have an adverse birth outcome.

Similar to the imbalanced dataset, there is no significant difference between the four classifiers when analyzing the result in Table 4.2 and Figure 4.2. Here, logistic regression has a slightly better f1-score, although it is not significant enough to draw any conclusions.

Since the core of machine learning is the input data, 170 entries are far too few to build a good model in this case. Comparing this to another study that used undersampling on an imbalanced dataset, their model yielded much better results (Mohammed et al., 2020). The main difference was the size of the datasets, they had 200 000 entries and 14 000 after undersampling. This strengthens the theory that size can be of great importance in this case. Another major drawback of using this resampling method is the potential that crucial data points may have been deleted in the removal process which prevents the algorithm from training on it (Mazurowski et al., 2008).

5.3 Oversampling

Using SMOTE, synthetic data points from the minority class were created to balance the dataset to 50/50. It increased the number of total data points from 1615 to 3060. The increase of size in combination with the balancing yielded the best result out of the three approaches. The model that used the random forest classifier generated the highest scores across all the metrics in Table 4.3. This is the greatest difference observed between the classifiers compared to Table 4.1 and 4.2.

Looking at the confusion matrices in Figure 4.3, there is a large number of FP in the classifiers LR and SVC. Not surprisingly, they did not score as well as RF. KNN on the other hand, had a hard time classifying TP, whereas the other three classifiers preformed equally good. During validation, they predicted 5 out of 9 cases correctly. Therefore, what distinguishes RF from the rest is the ability to predict the TN with 134 out of 153 correctly predicted. It should be noted that due to the function stratify being used during k-fold cross-validation, the same proportions as the original dataset was used to accurately reflect reality. Thus, the number of entries in the negative class (no adverse birth outcomes) was comprised of 94.5% whereas the positive class (adverse birth outcomes) had 5.5% data points in the validation.

A study by Mohammed et al. (2020) comparing various classifiers with an undersampled and oversampled dataset concluded that the oversampling method gave the highest scores and is suitable for small datasets. Their study also obtained the best result with the random forest classifiers. However, the authors concluded that the risk of overfitting increases when oversampling due to the identical reproductions of the instances in the minority class. Meaning that the model could give accurate predictions of the training data but perform poorly with new data. This can be checked by validating the algorithms on data that has not been trained on before.

5.4 Survey of the field

This study is an example of the various problems that occur when working with empirical real-world data from a human cohort, especially in developing countries where data collection can be challenging. In this thesis, the oversampling technique yielded the best result. This was most probably due to the small dataset (n = 1615) used for this analysis that made undersampling function very poorly. A large number of missing variables can also add uncertainty, however, multiple imputation is a well-established technique that deals with that issue when applied carefully (Sterne et al., 2009).

Other articles that have studied the possibilities of predicting the risk of stillbirth and preterm pregnancies with ML have yielded similar results. A study done by Koivu and Sairanen (2020) tested classifiers such as LR, artificial neural networks as well as gradient boosting decision trees to predict the risks of adverse birth outcomes with ML. Their best algorithm yielded an AUC score of 0.76 for early stillbirth and 0.63 for late stillbirth. In the present thesis, an AUC score of 0.72 (see Table

4.3) was achieved which is interesting due to this study being solely based on data from Ethiopia whereas Koivu and Sairanen (2020) validated their result on data gathered from New York City. It shows that as long as one overcomes the difficulty of collecting data in developing countries, there is a good possibility of creating an effective algorithm to predict adverse outcomes.

Contrary to the benefits of using ML on epidemiological data, another study from 2019 claims that machine learning does not have any performance benefit compared to logistic regression. After conducting a literature search between the years 2016 and 2017, they found no evidence that ML was superior to LR (Christodoulou et al., 2019).

An essential factor to study further is the selection of features. The feature Hospitalization after birth should possibly have been removed due to it being evident that stillbirth and neonatal death would follow with a hospitalization. It is also not useful when the objective is to find factors that cause adverse birth outcomes to create a predictive model. The place of delivery can also be seen as a possible source of bias. Most of the people that had adverse birth outcomes delivered the baby at Adama Regional hospital. Information about the different hospitals is lacking. Therefore, there may be a possibility that this hospital is the most specialized in managing pregnancy complications. Although the opposite can also be true. Without this information, this may not be a relevant feature.

Similar studies have confirmed that the other features, i.e. number of pregnancies, age, level of education as well as living situation are important risk factors regarding adverse birth outcomes (Aminu et al. 2014; Bhusal et al. 2019). A study done in the Assosa zone, located in western Ethiopia, highlighted predictors for neonatal death that turned out to be important for that region. These were age, prenatal visits, complications during pregnancy, and childbirth. They saw a rise in mortality due to the low access and use of obstetric services (Kidus et al., 2019). Given these studies, the binary logistic regression that was done in the pre-processing stage worked well when finding the best features for the target variable. This shows the importance of being aware of the most common causes for adverse birth outcomes when predicting the risks, along with the variations of risk factors depending on the specific culture, regulations, and procedures for that region.

5.5 Future work

The present thesis demonstrated that oversampling works well when dealing with imbalanced epidemiological data, especially with the random forest classifier. The models can be improved in several ways with additional time. Given that the dataset used included over 200 features, it would have been interesting to investigate whether alternative conclusions could have been made with other or more features. Changing the hyperparameters depending on the sampling method would also be interesting to try.

A literature search on 434 articles and 11 studies performed based on Mangold et al. (2021) revealed that the study with the best AUC score for predicting neonatal mortality using ML used linear discriminant analysis with 17 features. Therefore, it would have been interesting to test this supervised classification technique, as well as others, to examine a possible improvement of the score by changing the classifier.

5.6 Ethics

The data used to perform the machine learning analysis was gathered from participants of a study conducted in collaboration with Lund University and the Water and Public Health Department, Ethiopia Institute of Water Resources, Addis Ababa University, Ethiopia. All the participants of the study gave their written consent during the recruitment.

The study (Flanagan et al., 2022) that gathered the data was approved by The Ethical Review Board of the Ministry of Science and Technology in Addis Adaba, Ethiopia (310–046-2015) as well as the Lund University Ethical Committee (2015/364 and 2016/576). This study follows the GDPR rules used in the Data Protection Agreement and the data was therefore stored in a place where only the writer could access it. It was also agreed that the data will be deleted after the publication of the thesis to ensure the privacy of the study participants.

Due to the sensitive information that was collected from the study participants such as health data and their coordinates, respecting their privacy is crucial. All the participants were therefore pseudonyms and the data from the cohort was only available to researchers involved in the project. Lastly, the results presented cannot be traced to the individuals.

Chapter 6

Conclusions

To conclude, out of the three sampling techniques, oversampling with SMOTE yielded the best result for all four classifiers, logistic regression, support vector classifier, random forest, and k-nearest neighbors. With an AUC score of 0.72 and an f1-score of 0.85, the random forest classifier had the highest score. It was therefore the best-fitting classifier for this dataset. With further work, more data, and higher evaluation scores, machine learning can be of great use within the predictive medicine branch. Hopefully, the algorithm can find those who are at risk based on their environmental and preexisting health factors to then put in preventable measures to lower the risk of adverse birth outcomes in countries such as Ethiopia. However, in order to improve the evaluation scores of these models and identify the most essential characteristics to analyze for this region, additional research is required, particularly with larger study populations.

Bibliography

Aminu, M., Unkels, R., Mdegela, M., Utz, B., Adaji, S., and Van Den Broek, N. (2014). Causes of and factors associated with stillbirth in low-and middle-income countries: a systematic literature review. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121:141–153.

Awad, M. and Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines*, pages 39–66. Apress, Berkeley, CA.

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, 1:542–545.

Bhusal, M., Gautam, N., Lim, A., and Tongkumchum, P. (2019). Factors associated with stillbirth among pregnant women in Nepal. *Journal of Preventive Medicine and Public Health*, 52(3):154–160.

Caie, P. D., Dimitriou, N., and Arandjelović, O. (2021). Precision medicine in digital pathology via image analysis and machine learning. In *Artificial intelligence and deep learning in pathology*, pages 149–173. Elsevier.

Chollet, F. (2021). *Deep learning with Python.* Simon and Schuster.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22.

Douzas, G. and Bacao, F. (2017). Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert systems with Applications*, 82:40–52.

Flanagan, E., Oudin, A., Walles, J., Abera, A., Mattisson, K., Isaxon, C., and Malmqvist, E. (2022). Ambient and indoor air pollution exposure and adverse birth outcomes in Adama, Ethiopia. *Environment International*, 164:107251.

Hailemariam, M. and Ajeme, A. (2014). Solid waste management in Adama, Ethiopia: Aspects and challenges. *International Journal of Environmental and Ecological Engineering*, 8(9):670–676.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Huang, J. and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

Hug, L., Mishra, A., Lee, S., You, D., Moran, A., Strong, K. L., and Cao, B. (2020). A neglected tragedy the global burden of stillbirths: report of the UN inter-agency group for child mortality estimation, 2020. United Nations Children's Fund.

Khoshgoftaar, T. M., Golawala, M., and Van Hulse, J. (2007). An empirical study of learning from imbalanced data using Random Forest. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 310–317. IEEE.

Kidus, F., Woldemichael, K., and Hiko, D. (2019). Predictors of neonatal mortality in Assosa zone, Western Ethiopia: a matched case control study. *BMC pregnancy and childbirth*, 19(1):1–13.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2:1137–1145.

Koivu, A. and Sairanen, M. (2020). Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health information science and systems*, 8(1):1–12.

Lalkhen, A. G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing education in anaesthesia critical care & pain*, 8(6):221–223.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Liu, B. (2011). Supervised learning. In *Web data mining*, pages 63–132. Springer.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

Malacova, E., Tippaya, S., Bailey, H. D., Chai, K., Farrant, B. M., Gebremedhin, A. T., Leonard, H., Marinovich, M. L., Nassar, N., Phatak, A., et al. (2020). Stillbirth risk prediction using machine learning for a large cohort of births from western australia, 1980–2015. *Scientific reports*, 10(1):1–8.

Mangold, C., Zoretic, S., Thallapureddy, K., Moreira, A., Chorath, K., and Moreira, A. (2021). Machine learning models for predicting neonatal mortality: a systematic review. *Neonatology*, 118(4):394–405.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436.

McClure, E. M., Nalubamba-Phiri, M., and Goldenberg, R. L. (2006). Stillbirth in developing countries. *International Journal of Gynecology & Obstetrics*, 94(2):82–90.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Poolsawad, N., Kambhampati, C., and Cleland, J. (2014). Balancing class for performance of classification with a clinical dataset. In *Proceedings of the World Congress on Engineering*, volume 1, pages 1–6.

Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE.

Sasaki, Y. et al. (2007). The truth of the f-measure. *Teach tutor mater*, 1(5):1–5.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338.

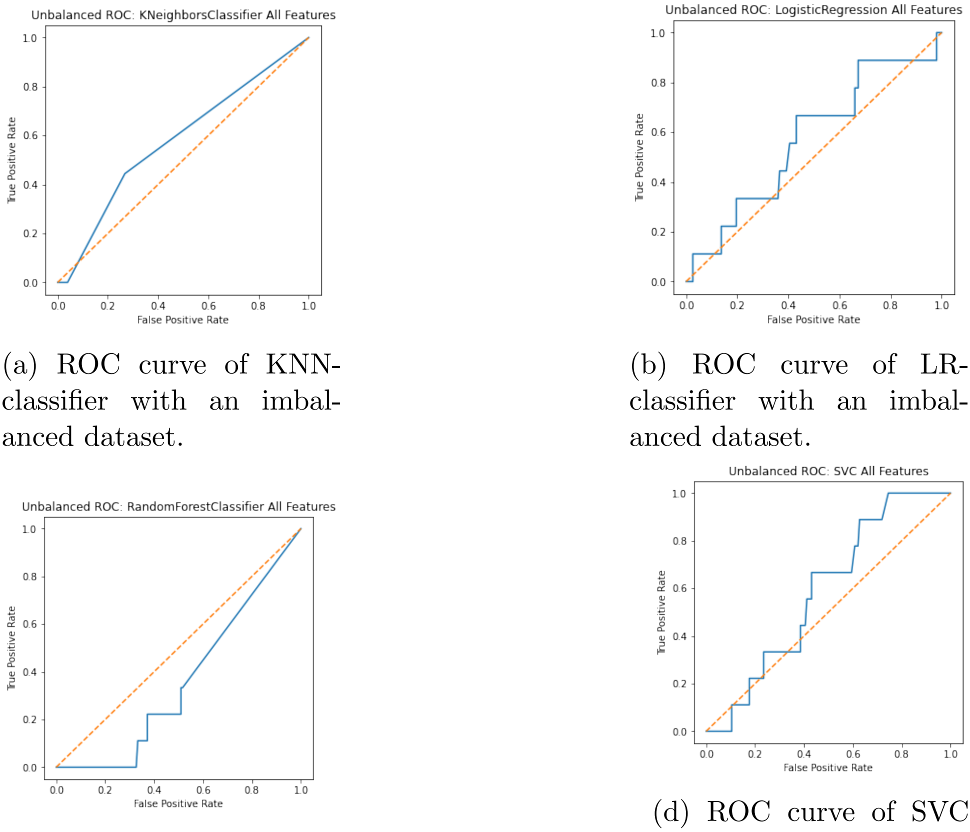
World Health Organization (2022a). Neonatal mortality rate (0 to 27 days) per 1000 live births) (sdg 3.2.2). https://www.who.int/data/gho/indicator-metadata-registry/imr-details/67#:~:text=Neonatal\%20deaths\%20(deaths\%20among\%20live,28th\%20completed\%20day\%20of\%20life. Visited on 2022-11-28.

World Health Organization (2022b). Newborn mortality. https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report2021#:~:text=There\%20are\%20approximately\%206700\%20newborn,in\%20child\%20survival\%20since\%201990. Visited on 2022-11-11.

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics-Theory and Methods*, 49(9):2080–2093.

Appendix

A.1 The resulting ROC curves



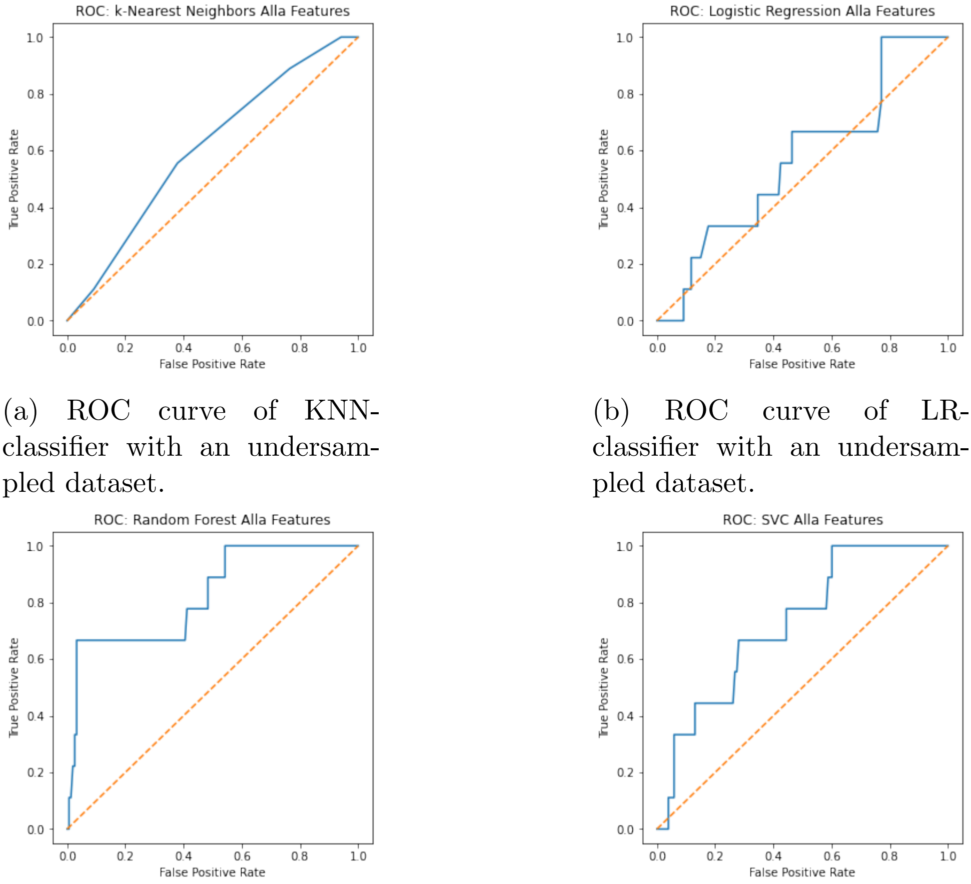
(a) ROC curve of KNN-classifier with an imbalanced dataset.

(b) ROC curve of LR-classifier with an imbalanced dataset.

(c) ROC curve of RF with an imbalanced dataset.

(d) ROC curve of SVC with an imbalanced dataset.

Figure 7.1: Four ROC curves on an imbalanced dataset with the respective classification model. The area under the curves are used to calculate the AUC score.



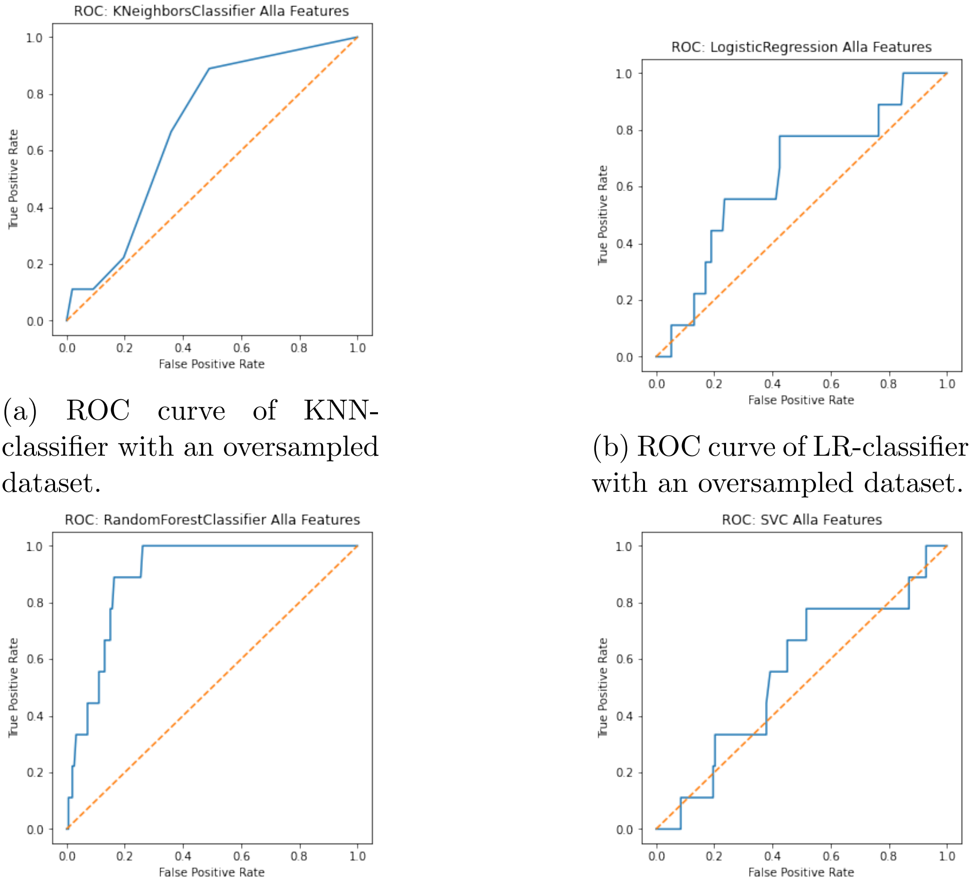
(a) ROC curve of KNN-classifier with an undersampled dataset.

(b) ROC curve of LR-classifier with an undersampled dataset.

(c) ROC curve of RF with an undersampled dataset.

(d) ROC curve of SVC with an undersampled dataset.

Figure 7.2: Four ROC curves on an undersampled dataset with the respective classification model.



(a) ROC curve of KNN-classifier with an oversampled dataset.

(b) ROC curve of LR-classifier with an oversampled dataset.

(c) ROC curve of RF with an oversampled dataset.

(d) ROC curve of SVC with an oversampled dataset.

Figure 7.3: Four ROC curves on an oversampled dataset with the respective classification model.