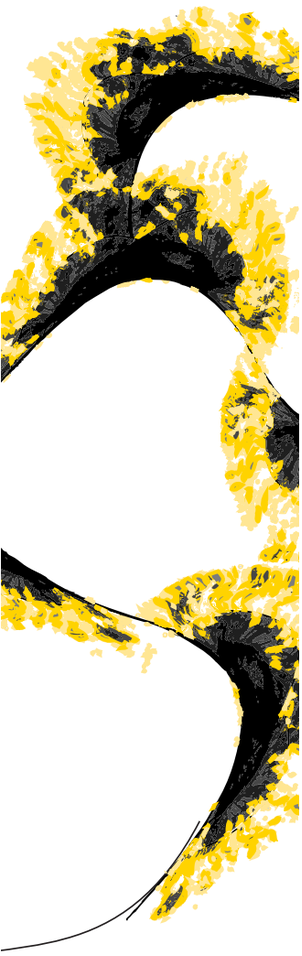


RAM ROBOTICS AND MECHATRONICS

PREDICTING TSB VALUES USING TCB MEASUREMENTS AND PATIENT CHARACTERISTICS WITH MACHINE LEARNING

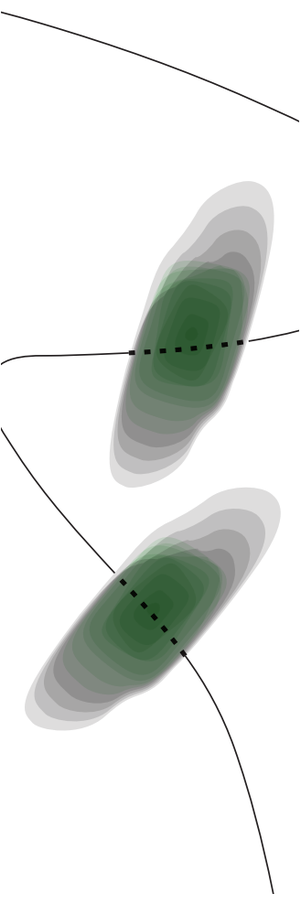
D.J. (Dries) Cavelaars

BSC ASSIGNMENT

Committee:
dr. ir. F. van der Heijden
E.I.S. Hofmeijer, MSc
dr. A. van Houselt

January, 2021

002RaM2021
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands



UNIVERSITY OF TWENTE. | TECHMED CENTRE UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Abstract

Jaundice can be a serious medical condition which affects mostly newborn infants. High levels of bilirubin in the bloodstream are the main cause of this. These bilirubin levels, total serum bilirubin (TSB), need to be determined frequently, in order to detect early onset of jaundice and start treatment. The traditional method of determining the TSB is by analysing blood samples in a laboratory. The frequent venipunctures required to obtain these blood samples cause a high strain on the newborn. A non-invasive technique to determine the bilirubin levels is through the skin, using a transcutaneous bilirubinometry (TcB) device. This technique is preferred because it is non-invasive, cheaper and faster than the traditional method. However, TcB measurements are less accurate, especially for preterm infants. This group is also more susceptible to jaundice. Therefore it is of interest to increase the accuracy of a TcB measurement.

This study uses the data of $n = 101$ newborn infants that are preterm (median gestational age: 30.5 weeks, range: 28.0 to 35.7). This data includes TcB measurements on five body locations and patient characteristics, along with the actual TSB measurements. Machine learning is applied to map the TcB measurements more accurately to a TSB value. Two models have been realized: a linear regression model and a decision tree. The root mean square error (RMSE) of the linear regression model is $21.9\,\mu\text{mol}\,\text{L}^{-1}$, and that of the decision tree is $30.4\,\mu\text{mol}\,\text{L}^{-1}$. The specified error of the TcB measurement device for preterm infants without phototherapy is $27.4\,\mu\text{mol}\,\text{L}^{-1}$, and $39.0\,\mu\text{mol}\,\text{L}^{-1}$ after phototherapy. The data used to train the models contain a mix of measurements with and without phototherapy. The error of the measurement data ranges from 30.2 to $85.2\,\mu\text{mol}\,\text{L}^{-1}$, depending on the body measurement location.

The linear model reduces the number of unnecessary blood samples from 201 to 69, and has an RMSE that is lower than the specified error of the TcB device, and can therefore be accepted as a valid model to predict TSB values.

Conflict of interests The author declares to have no financial interest in any TcB measuring devices.

Contents

Nomenclature

APGAR	method to indicate health of a newborn right after birth
cross-validation	technique to determine optimal hyperparameters
hyperbilirubinemia	condition in which the bilirubin levels in the blood are high
jaundice	medical condition that causes yellowing of the skin and whites of the eye, caused by hyperbilirubinemia
MSE	Mean Squared Error: parameter used by MATLAB to train a model
neonate	a newborn child
predictor variable	independent variable used to predict the response variable
response variable	dependent variable that is predicted by the model
RMSE	Root Mean Square Error: indication of model accuracy
TcB	Transcutaneous Bilirubinometry: non-invasive technique to measure bilirubin levels through the skin
test set	part of dataset used to determine the accuracy a final model
training set	part of dataset used to train the model during building
TSB	Total Serum Bilirubin: bilirubin concentration in blood
validation set	part of dataset used to validate settings of a model after training
venipuncture	method of drawing blood from a vein for use as a blood sample

1 Introduction

1.1 Problem statement

Jaundice is a disease that affects mostly newborns. It is caused by hyperbilirubinemia [1], elevated levels of bilirubin in the bloodstream. High levels of this pigment are common in neonates, and are mostly harmless. About 80% of neonates [2] undergo jaundice within their first week of life. Bilirubin levels rise due to an increased red cell breakdown, and the young liver's inability to effectively excrete bilirubin. When the concentration of bilirubin reaches a certain threshold, it can cause severe brain damage (bilirubin encephalopathy, the poisoning of the brain due to bilirubin), cerebral palsy, and if left untreated, even death. Neonatal jaundice is accountable for 1.3% of deaths in the early-neonatal period (postnatal age: 0 to 6 days) in 2016, globally [3]. Jaundice treatments include blood exchange transfusion and phototherapy [4].

The levels of this pigment can be determined by collecting blood samples, yielding the Total Serum Bilirubin (TSB). This method, however, causes a high strain on the newborn, since their blood volume is limited. For the reason that the bilirubin levels can rise rapidly, the TSB needs to be measured frequently. This causes an even bigger strain.

A non-invasive method that can determine the bilirubin level is via a Transcutaneous Bilirubinometry (TcB) measurement [4]. TcB literally means the measurement of bilirubin through the skin. A TcB device uses two optical paths that direct light of specific wavelengths into the skin, making this method painless and convenient. The devices are point-of-care devices [5], meaning they can be carried to the patient and their values read off instantly. This leads to faster results and a decrease in costs. McClean et al. [5] determined that one TcB screening is 5 to 20 times cheaper than TSB, and 4 times faster.

Unfortunately, TcB measurements are not very accurate. Therefore, the TSB values must still regularly be checked using the traditional method.

1.2 Current research

A solution would be to create a model that can accurately predict the actual TSB values, using the TcB measurements, and possibly taking into account certain patient characteristics. This gives the advantage that hyperbilirubinemia can be detected earlier on, while simultaneously reducing the strain on the neonate's blood system by avoiding blood tests.

Multiple research papers [6, 7, 8, 9] have been published that discuss the statistical correlation between acquired TcB and the actual TSB values. Schmidt et al. [6] show that the correlation between TcB and TSB ranges from 0.79 to 0.92, for different gestational ages. The data used in their research is shown in figure 1. It gives an indication how the TcB relates to the TSB values.

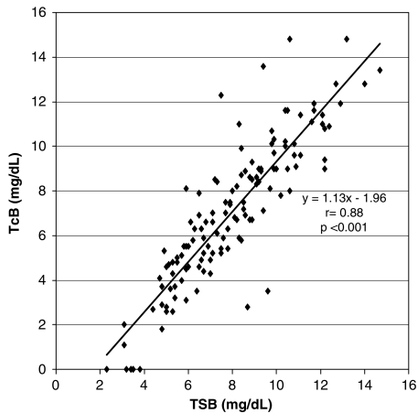


Figure 1: Relation between TcB and TSB values, from Schmidt et al. [6]

The predictive power of TcB measurements [7] depends on multiple factors, such as skin anatomy, gestational age, and measurement body location. The TcB measuring seems to underestimate [8] the TSB values for light and medium skin colours while overestimating in darker skin colours. The studies by Schmidt et al. [6] and Karen et al. [9] suggest that the accuracy of the TcB measurements decreases with higher levels of serum bilirubin. The TcB measurement device used in this study is not recommended for use on preterm babies, due to its reduced accuracy.

A recent research paper by Raba et al. [4] investigates the possibility to use TcB measurements to predict whether phototherapy will eventually be necessary. Their conclusion was that TcB performs as a poor predictor for phototherapy. This was done without the use of machine learning. In fact, no papers have been found that apply machine learning to improve the accuracy of TcB measurements.

1.3 Goal of assignment

The goal of this bachelor assignment is to develop a model that can derive the correct TSB values based on the less accurate TcB values, along with certain patient characteristics. This improves the usability of TcB measurements. Additionally, a statistical analysis will be performed to indicate the significance of this model. The model will be generated using machine learning. Multiple coding languages will be considered, such as MATLAB and Python.

The model will be created based on actual measurements performed on newborn patients. A dataset from 101 different preterm infants is provided at the beginning of the assignment, which means that data collection is not part of the scope of this bachelor assignment. The dataset consists of TSB values, TcB measurements and certain patient characteristics.

A regression model is desired since the reference value (TSB) is a continuous variable. Regression models fall under supervised machine learning. There exist, however, different regression techniques, which can incorporate different kinds of input data. Some predictor variables are continuous, such as TcB, while others are classified, for instance some patient characteristics. These different techniques will need to be researched, and an optimal method will be determined.

This study focuses on improving the predictive power of existing TcB devices, by processing their measurements differently. It does not include developing new detection methods.

A theoretical background on bilirubin and the machine learning techniques is given in chapter 2 and applied in the method, chapter 3. Next, the results are presented in chapter 4. These results are discussed in chapter 5, along with a conclusion in chapter 6.

Combining biology and machine learning to improve healthcare is a broad task, which makes this a true Advanced Technology bachelor assignment.

1.4 Research questions

The research questions that have arisen from the goal of this assignment are formulated as follows:

- Which machine learning model can be used to predict TSB values, based on TcB values and patient characteristics?
- How accurate is the prediction of the model?
- Which patient characteristics are meaningful?

2 Theory

The theory behind the different models that will be implemented during this assignment, is presented in this chapter. There exist different types of machine learning algorithms [10], namely supervised, unsupervised and reinforcement learning. Supervised machine learning models are trained using known, correct combinations of input and output data. The predictions for new, unseen output data are based on the training of the model.

A machine learning model as a whole is a function which maps the independent variables (predictors) to a dependent variable (response).

Firstly, more background on the pigment bilirubin is provided. Then a linear regression model is discussed, followed by a decision tree model.

2.1 Bilirubin

Hyperbilirubinemia occurs when the bilirubin levels in the blood rise so much, that the bilirubin diffuses into the surrounding tissues, such as the skin. This build-up starts in the trunk, and moves to the limbs. The accumulation of bilirubin can then be observed as a yellow color shift in the skin and whites of the eyes. A TcB device measures the concentration of bilirubin in the skin, from which it estimates the corresponding bilirubin blood concentration. The TcB meter used in this study (Draeger Jaudice Meter JM-105) does this by emitting light at two wavelengths into the skin, and measuring the reflected light intensities [7]. The first wavelength of 450 nm is at the peak of the bilirubin absorption spectrum. A second wavelength of 550 nm is used as a control. Hemoglobin absorbs light in roughly equal amounts for both wavelengths. From these two measurements, the contribution of bilirubin can be calculated [11].

As mentioned before, abnormally high bilirubin levels can be damaging to the newborn infant. Treatments to reduce these levels include phototherapy and blood exchange transfusion. There is, however, no definite threshold to determine when a treatment is needed. These thresholds are dependent on the postnatal age, the gestational age and birth weight, and can differ per hospital and per country. The guidelines for pediatricians in the Netherlands [12] show different charts, based on gestational age and birth weight. The infants are labeled by a risk level (standard or high), which are determined by factors such as APGAR score and sepsis. For patients with a higher risk, a lower threshold for treatment is used.

Figure 2 shows the bilirubin thresholds for infants with a gestational age below 35 weeks and birth weight between 1500 and 2000 grams. For different birth weights and gestational ages, different charts are used, since their thresholds differ [13]. The chart for term infants can be found in appendix A, figure 10. The stationary value of the threshold (maximum value, which is reached after the first few days) for preterm infants of all birth weights is shown in appendix B, table 7, and for term infants in table 8.

The guidelines state that in order to use a TcB measurement as a TSB value in the charts, a margin of 50 µmol L^{-1} must be added: $TSB = TcB + 50$. It also states that TcB should not be measured during or after phototherapy. When a threshold is reached using a TcB measurement, the bilirubin levels must be verified using a TSB determination from a blood sample.

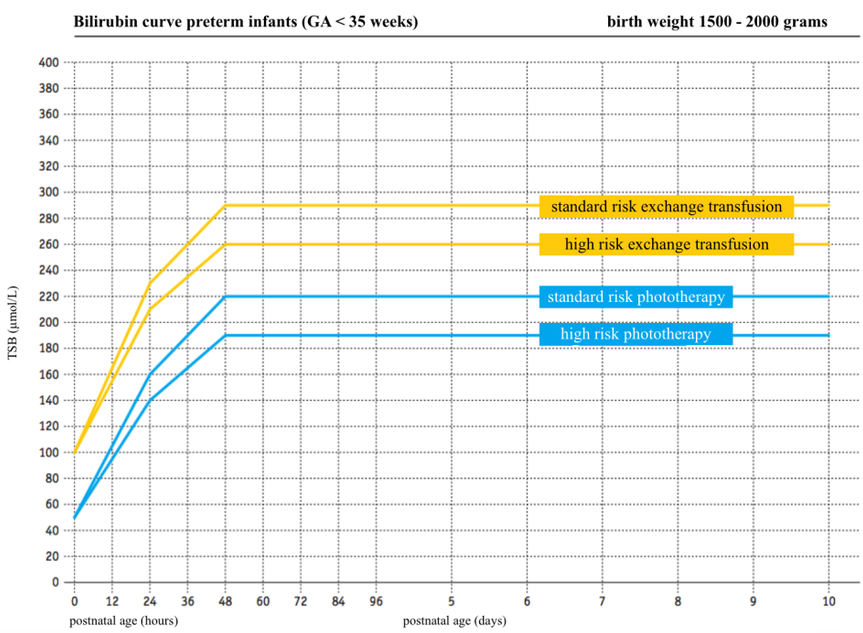


Figure 2: Chart showing the phototherapy and exchange transfusion thresholds for preterm infants, based on postnatal age for two levels of risk, adapted from Dutch pediatric guidelines [12]

2.2 Linear regression model

2.2.1 Model function

A linear regression model makes a prediction with the postulation that the response variable is linearly proportional to all predictor variables, as in equation 1:

$$R \sim 1 + X_1 + X_2 + \ldots + X_{N-1} + X_N \tag{1}$$

with R being the response variable, and X_n being the n^{th} predictor variable.

This results in a model function (equation 2) which calculates the predicted value f based on the independent variables $x_1, x_2, \ldots, x_{N-1}, x_N$. To achieve this, every variable is multiplied by a constant coefficient $c_1, c_2, \ldots, c_{N-1}, c_N$. Note that the coefficient c_0 does not belong to a predictor variable, but rather to the constant 1. This acts as a vertical intercept value.

$$f(x) = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_{N-1} x_{N-1} + c_N x_N \tag{2}$$

$$e_m = r_m - f_m \tag{3}$$

Equation 3 shows that the error e_m for the m^{th} measurement can be seen as the difference between the true response variable r_m and the predicted value f_m. Achieving a zero error e_m can easily be accomplished for a single measurement m. It is more interesting, however, to calculate the Root Mean Squared Error (RMSE) of all the errors of all measurements. The RMSE gives an indication of the accuracy of a model as a whole. Equation 4 shows how the RMSE is calculated based on the response values r and predicted values f, for M amount of measurements:

$$RMSE = \sqrt{\left(\frac{\sum_{m=1}^{M}(r_m - f_m)^2}{M}\right)} \tag{4}$$

The coefficients $c_0, c_1, ..., c_{N-1}, c_N$ are tuned during training in such a way that the RMSE is minimized.

The entire model is stored in the values of the coefficients. The coefficients represent the weight associated with that predictor variable.

2.2.2 Categorical values

At first sight, it may seem that this model function (equation 2) only works for continuous predictor variables. However, it is possible to associate categorical values to these predictors. Suppose we want to use the variable `C-section` as a predictor. This is a binary variable, as it can take the values `Yes` and `No`. We could then set the predictor variable x_1 as `Yes`, that is, $x_1 = 1$ when a `C-section` has taken place, as in equation 5:

$$x_1 = \begin{cases} 1 & \text{C-section} = Yes \\ 0 & \text{C-section} = No \end{cases} \tag{5}$$

In doing so, the corresponding coefficient c_1 adds a value to the model prediction when the categorical variable is in one state, and is equal to zero when the categorical variable is in another state.

Similarly, this method can be extended to a case where the categorical variable can take more than only two values, as in the binary case. Here, multiple intermediate predictor variables are created. Consider the variable `Fetal position`, which can take the values `Head`, `Breech` and `Transverse lie`. The corresponding case definitions for the predictor variables x_1 and x_2 are:

$$x_1 = \begin{cases} 1 & \text{Fetal position} = Head \\ 0 & otherwise \end{cases} \qquad x_2 = \begin{cases} 1 & \text{Fetal position} = Breech \\ 0 & otherwise \end{cases} \tag{6}$$

When both x_1 and x_2 are equal to 0, it means that the `Fetal position` variable has the value `Transverse lie`. `Transverse lie` is therefore defined as the control variable, as it does not contribute a value to the model. `Head` and `Breech` can both add or subtract a value to the model, depending on their corresponding coefficient.

More generally, a categorical variable can be split into $n - 1$ predictor variables, for n different values the variable can take.

2.3 Decision tree model (regression tree)

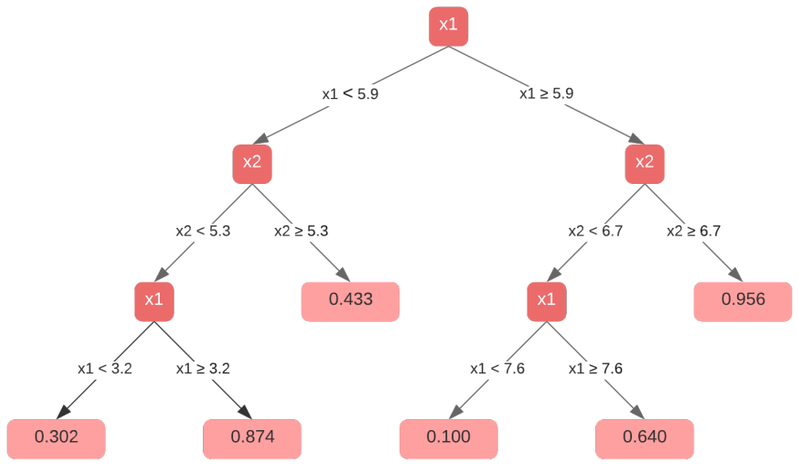
A decision tree is a model which fits data into different categories based on their predictors. It achieves this by considering conditional statements consecutively. A decision tree can therefore be seen as a combination of nested if-statements. The response variable can be both categorical (classification tree) or a continuous variable (regression tree).

A graphical representation of an example decision tree is shown in figure 3a. In the tree structure, the branches represent the decisions that are being evaluated, and the leaves represent the response labels.

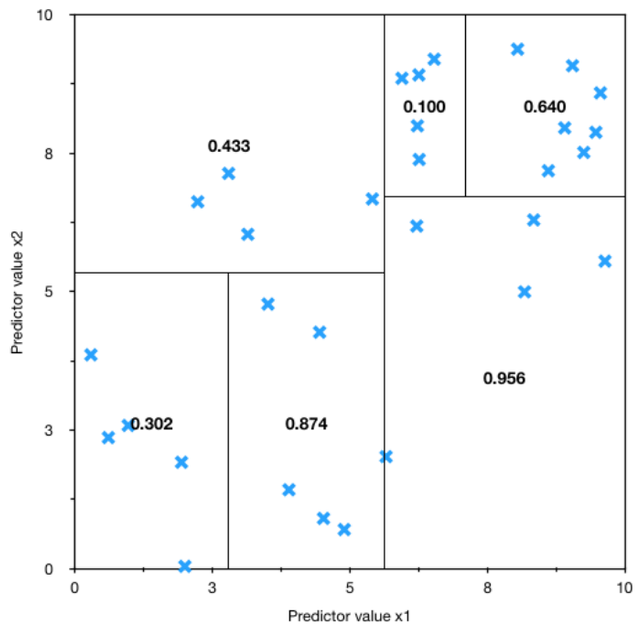
Figure 3b shows a plot with two predictor variables x1 and x2. The decision tree model is overlaid over the predictor data. For every branch node (decision split) in the tree, a straight line is added to the plot which partitions the data. The average response value per partition is displayed, which corresponds to the predicted value for that node. Note that in this example, there are only two predictor variables. This means that the plot can easily be displayed graphically. For higher order models, this is not the case.

Nodes in a graph are numbered in ascending order from the top row to the bottom row, and left to right within a row. The statements belonging to every node in the example decision tree are the following:

1. if $x1 < 5.9$ then node 2 else node 3
2. if $x2 < 5.3$ then node 4 else node 5
3. if $x2 < 6.7$ then node 6 else node 7
4. if $x1 < 3.2$ then node 8 else node 9
5. prediction = 0.433
6. if $x1 < 7.6$ then node 10 else node 11
7. prediction = 0.956
8. prediction = 0.302
9. prediction = 0.874
10. prediction = 0.100
11. prediction = 0.640



(a) Graphical representation decision tree



(b) Partitioned data according to the decision tree

Figure 3: An example regression tree

In order to apply this model onto a new set of data to predict the response value, the above scheme or graphical representation in figure 3a can be followed for every measurement. Additionally, the data can be plotted in the same plot as figure 3b and the value of the corresponding partition can be read off.

It is possible that a predictor variable is unknown for a certain measurement. A split can then not be made, since the conditional statement cannot be evaluated. Therefore there exists a third option that the branch can lead into, namely the weighted average of the values of the leaves under it.

Categorical variables can be used in a regression tree in the same way as the method described in the linear regression model.

2.3.1 Hyperparameters

While training a model, the machine determines which statements are evaluated at every node. Therefore these do not need to tuned by the user. There are, however, so-called hyperparameters which can be chosen by the user. These can, for example, control the depth of the regression tree. Which hyperparameters are tuned exactly will be explained later.

3 Method

All the data processing and model creation were done in MATLAB version R2020a, with an academic use licence. One toolbox is required to run the code, namely *Statistics and Machine Learning Toolbox*. The main function to train a linear regression model is `fitlm()` [14]. This function uses QR decomposition to optimise the coefficients of the model. QR decomposition uses linear algebra to perform a linear least squares algorithm. A matrix A is decomposed into the product $A = QR$, with Q an orthogonal matrix and R a right triangle matrix, hence the name QR decomposition.

Similarly, a regression tree is trained using `fitrtree()` [15]. This function uses standard CART [16] to determine a split based on maximising the reduction in MSE (mean squared error) per node. In other words, it finds the predictor that most effectively splits the data at a node. The function standard CART refers to a standard Classification And Regression Trees algorithm. For both models, new response values can be predicted using the `predict()` function [17].

3.1 Training, validation and test set

It is important to be able to determine the effectiveness of a model, after it has been created. This is done by testing the model on a different dataset than the one used to train the model. Therefore, throughout the process of making the models, the data has been split into different sets repeatedly. Generally, the data is split into a training set used to train (build) the model. Next, a validation set is used to assess the performance of the model based on the current hyperparameters, and is used to improve those hyperparameters. Once satisfied with the model, the final model can be tested using the test data. The process of distributing the data over the different sets is described below, and can be seen visually in figure 4. The width of the dataset is proportional to the number of patients it contains. The steps indicated in the figure correspond to the steps in the paragraphs below.

It is important to ensure that the accuracy of both the linear model and decision tree model are evaluated in the same way, so that they can be compared. Therefore, the following steps of distributing the data are performed only once, and the same distribution is used for both models.

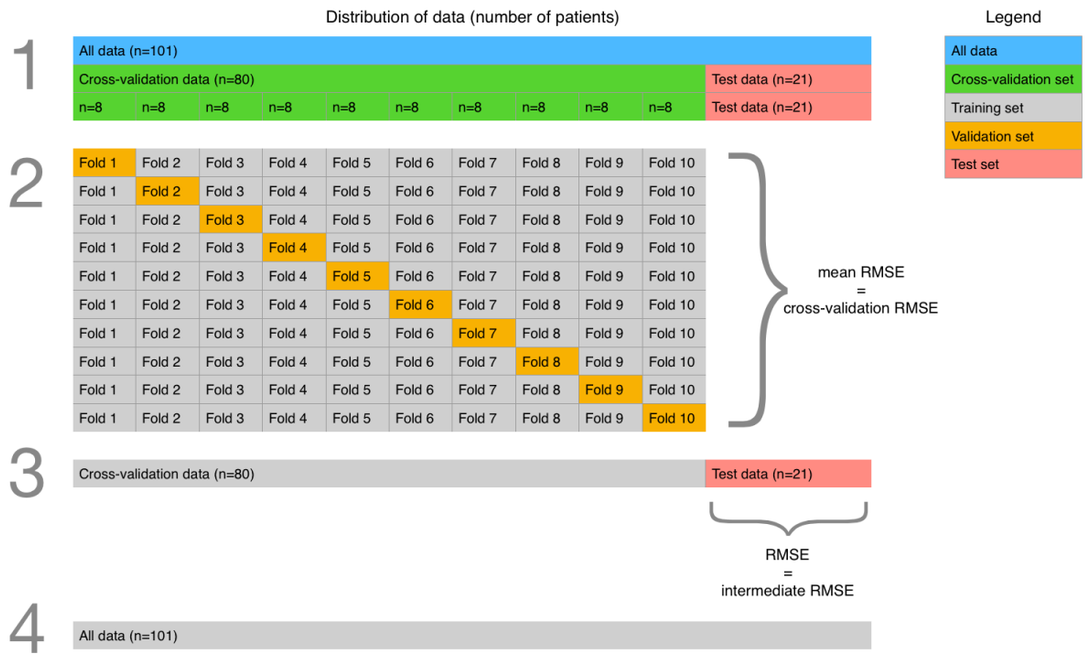


Figure 4: The distribution of the data, used for the k-fold cross-validation

3.1.1 Test set

Step 1: Firstly, a part of the complete dataset is held out, which can be used as a test set in the end to test the model. Here, 20% of the total data is held apart. This leaves 80% of the data to be used for both the training and validation. To optimize the use of the data, the data is split into training and validation sets after every training run of the model. This is called cross-validation. The specific cross-validation technique used is k-fold cross-validation.

3.1.2 k-fold cross-validation set

The cross-validation dataset is separated into k different subgroups. Here, $k = 10$ is chosen. The dataset is large enough to be split into 10 folds, with every fold still being representative of the entire dataset. The folds contain roughly the same amount of measurements, from the same amount of patients ($n = 8$). This is to make the folds as uniform as possible and decrease deviation. Moreover, the measurement data is split per patient, so the data from the same patient cannot be shuffled into multiple folds. This is to prevent the model from recognizing a patient based on its characteristics.

3.2 Hyperparameter tuning

Hyperparameters are parameters that can be set by the user, prior to training the machine learning model. The internal model parameters will be set during the training. The hyperparameters can be tuned after every cross-validation run. Note that for a linear regression model, there are no hyperparameters to be tuned. It is, however, used to add and remove predictor variables to create an accurate as possible model.

Step 2: For every training run of the model, one fold is kept apart, which will be used for validation. For every subsequent iteration of the training, a different fold will be used as a validation set, thus leading to different model performances. The performance is shown as the root mean square (equation 4). The mean RMSE will be saved as a cross-validation RMSE and gives an indication of the performance for the specific hyperparameters. After k runs, every fold will have been used as a validation set.

Step 3: After the optimal hyperparameters have been found (with the smallest cross-validation RMSE), the performance of the model can be determined. This is done by applying the model on the test set that has been set aside. The model is trained by using all the data in the cross-validation set, so without distinguishing between the folds. The performance on the test set is again determined using the RMSE, and saved as an intermediate RMSE.

This final model could be seen as an end product, but its performance depends somewhat on the distribution that has been made in the very first step. To counter this, all above steps 1 to 3 have been repeated, thus with different training, validation and test sets. The optimal hyperparameters are again determined, together with a new intermediate RMSE. Steps 1 to 3 are repeated five times, and their results recorded.

Step 4: Lastly, the opportunity remains to build the final model using all the available data. This model cannot be tested, however, since there is no test set available. Its accuracy can be derived from the intermediate RMSEs in step 3.

The pseudo-code for the process of training and testing a model is shown in algorithm 1:

```
for 5 times do
    randomly divide studies over 10 folds and one test set;
end
for each permutation of folds and test set do
    hold test set apart;
    for each combination of hyperparameters do
        for each fold in k=10 do
            train model using the 9 other folds;
            validate model using the fold;
            calculate and remember RMSE;
        end
        set cross-validation_RMSE as: average RMSE for these hyperparameters;
    end
    remember hyperparameters that yield lowest cross-validation_RMSE;
    train model using all k=10 folds and optimal hyperparameters;
    test model using test set that was held apart;
    determine and remember intermediate_RMSE;
    remember prediction of test set;
end
set final_model_RMSE as: maximum intermediate_RMSE of all test sets;
plot predictions of all test sets;
randomly divide studies over 10 folds (without creating test set);
for each combination of hyperparameters do
    for each fold in k=10 do
        train model using the 9 other folds;
        validate model using the fold;
        calculate and remember RMSE;
    end
    set cross-validation_RMSE as: average RMSE for these hyperparameters;
end
remember hyperparameters that yield lowest cross-validation_RMSE;
train final model using all k=10 folds and optimal hyperparameters;
final model cannot be tested;
```

Algorithm 1: Pseudo-algorithm of creating and testing a model

4 Results

4.1 Inspection of data

The data used in this study was previously obtained by the medical staff of the neonatal intensive care unit (NICU) of Isala Hospital in Zwolle, the Netherlands. A Draeger Jaudice Meter JM-105 was used for all the measurements. A part of the specification of this device can be found in the appendix, table 9. The measurements took place from December 2017, to August 2019. From the 101 patients ($n = 101$), an average of 17 times data was taken, over an observation period of 6 days.

Neonates with a gestational age below 37 weeks are considered preterm. All patients included in this study are preterm: the median gestational age is 30.5 weeks (range: 28.0 to 35.7 weeks). The TcB meter used in this study is not recommended to be used on preterm infants, since the error is too large for low gestational ages. The models created are trained solely on data of preterm babies.

The majority ($n = 54$) has a Caucasian ethnicity, and for a large group ($n = 41$), the ethnicity is unknown. The full analysis of the patient characteristics can be found in table 12 in the appendix.

A patient is observed over the course of its first couple of days of life. There are multiple measurement moments per day. Some (invariant) data about the patient characteristics are entered once when the baby is born, while other (variant) data is entered during the measurement moments. However, not all variant data is entered for every observation. It is therefore possible that the data for a TcB is known, but not the corresponding TSB value, or vice versa. There are 5 TcB measurement body locations: forehead, sternum, hip bone, tibia and ankle. Per location, a maximum of three consecutive measurements takes place, to increase accuracy. The mean value of these measurements is taken, after outliers have been removed. Ultimately, a total of 3071 TcB-TSB pairs are created. These pairs have been plotted in figure 5, per body location along with the line of best fit and the root mean square error between the TcB and TSB value. A straight line of slope 1 has been added for reference. A separate plot per body location can be found in appendix A, figures 11a to 11e. Visually we can see that the forehead, sternum and hip bone have a line of best fit relatively close to the optimal line, and have a smaller RMSE than the tibia and ankle measurements. The TcB measurements for the tibia and ankle are frequently below the TSB value, as seen by the lower gradient of the line of best fit. This is in agreement with the theory.

The specification of the TcB device states that the accuracy decreases after phototherapy has taken place. Also, the accuracy is lower for infants with a gestational age below 35 weeks. According to the measurement data, the forehead is the body location with the lowest error (RMSE: 30.2 µmol L^{-1}). This lies between the error for no phototherapy (± 27.4 µmol L^{-1}) and after phototherapy (± 39.0 µmol L^{-1}). This is plausible, since the data includes measurements of both before and after phototherapy. The hip bone measurements also lay between the specified errors. However, the sternum, tibia and ankle measurements have an error that is far greater than the specified error.

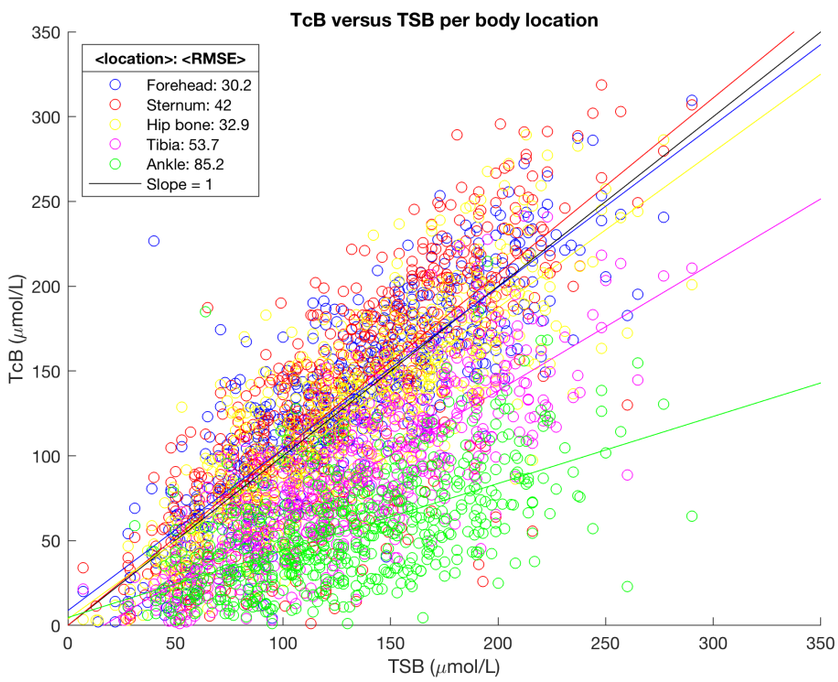


Figure 5: All TcB-TSB pairs plotted per body measurement location

4.2 Linear regression model

The data has been divided over a training, validation and test set, as per section 3.1. Per iteration of a cross-validation round, predictor variables are manually added and removed from the model to see what the influence on the cross-validation RMSE is. The goal is to minimize the RMSE. The generic equation 4 is rewritten to equation 7, to apply it to a regression model. Equation 7 shows how the RMSE is calculated based on the response values (TSB_{actual}) and predicted values ($TSB_{predicted}$).

$$RMSE = \sqrt{\left(\frac{\sum_{n=1}^{N}(TSB_{actual}(n) - TSB_{predicted}(n))^2}{N}\right)} \tag{7}$$

A linear regression model requires a value for all its predictor variables, otherwise the equation cannot predict a response value. However, as mentioned before, not all measurement moments include values for all the predictor variables. So when adding a new predictor variable to the model, the amount of useful measurements that contribute to the model decreases. Therefore, a tradeoff needs to be considered between lowering the RMSE, and having less measurements to train the model.

Equation 8 shows the final model:

$$TSB \sim 1 + TcB_1 + \ldots + TcB_5 + PatC_1 + \ldots + PatC_8 \tag{8}$$

with TcB_1 through TcB_5 being the TcB values of the 5 body locations, and $PatC_1$ through $PatC_8$ being the following 8 patient characteristics:

Continuous:

- Postnatal age (hours)
- Maternal age (years)
- Birth weight (grams)
- Gravidity

Categorical:

- C-section (yes/ no)
- IVH (yes/ no)
- Sepsis (yes/ no)
- Feeding (formula/ breastfeeding/ both)

The model incorporates a mix of continuous variables along with binary and higher order categorical variables.

A technique to help determine if a certain predictor variable has a good contribution to a model, is testing the following hypothesis:

Hypothesis: *The coefficient for the specified predictor variable is not equal to 0.*

The hypothesis can be accepted when the corresponding p-value is higher than a 95% significance level. When a coefficient is equal to 0, the associated variable does not contribute anything to the model. Accepting the hypothesis is therefore desired.

In other words, a high p-value for a predictor variable means the variable is significant to the model. The p-values of every predictor for an example model have been plotted in appendix A figure 12. Not all predictors reach the 95% significance mark, but are still useful for the model. The p-values change depending on which fold is being used to validate, and might be useful in another training data. The variable *Feeding* for example, consists of two subvariables *Formula* and *Both*. Removing the least performing variable *Both*, would also remove a well-functioning variable *Formula*.

Once satisfied with the predictor variables to use, the model can be trained using all 10 folds, and tested using the test set (step 3 from figure 4). This yields intermediate results (predictions and RMSE), which can be seen in figure 6. As mentioned before, steps 1 to 3 are repeated five times, thus yielding five intermediate models. The RMSE for these models ranges from 18.0 to 21.9 $\mu mol\,L^{-1}$. The mean is 20.1 $\mu mol\,L^{-1}$.

For this specific application, it is important to know how often the TSB is over- or underestimated, and to what extent. You would rather perform too many blood samples to determine the TSB in a lab, than miss an opportunity to reduce jaundice with treatment. In 49.8% of the cases, the TSB value is underestimated in the linear models. The error between the actual TSB and the predicted TSB (equation 3) is plotted as a histogram in appendix A figure 14a for every model. It gives an indication how the error is distributed. A positive error means an underestimation.

Lastly, a final model is built using the same predictor variables as before, while using all the available data. This is step 4 of the schema. The resulting coefficients can be found in figure 7 and appendix B table 10. The accuracy of this model is estimated using the accuracy for the five intermediate models. It is important for the healthcare application to employ a large enough error on the TSB prediction. Therefore the error of the final model has been estimated to be equal to the worst-performing intermediate model. The RMSE of that model is 21.9 $\mu mol\,L^{-1}$.

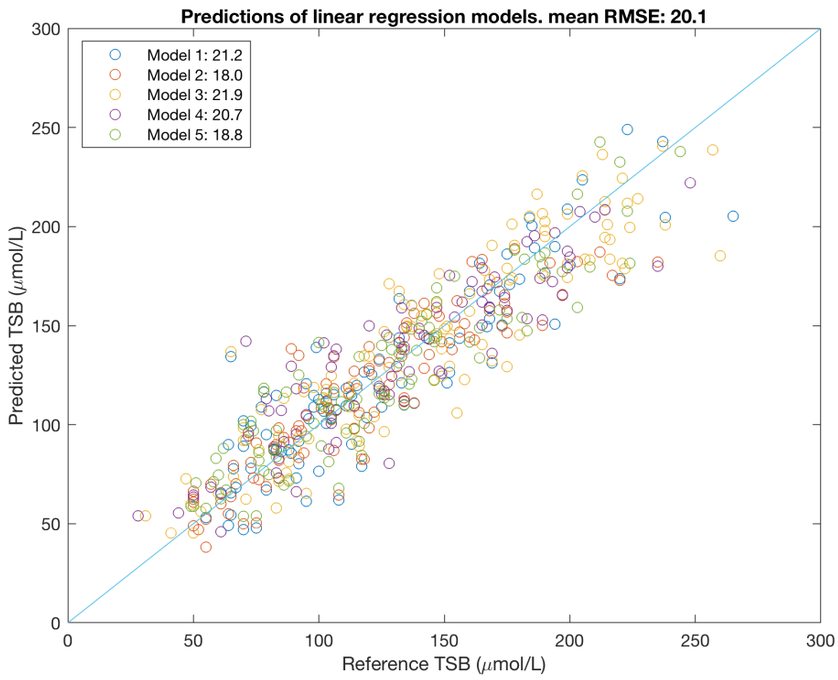


Figure 6: Predictions of the 5 intermediate linear models and their mean RMSE

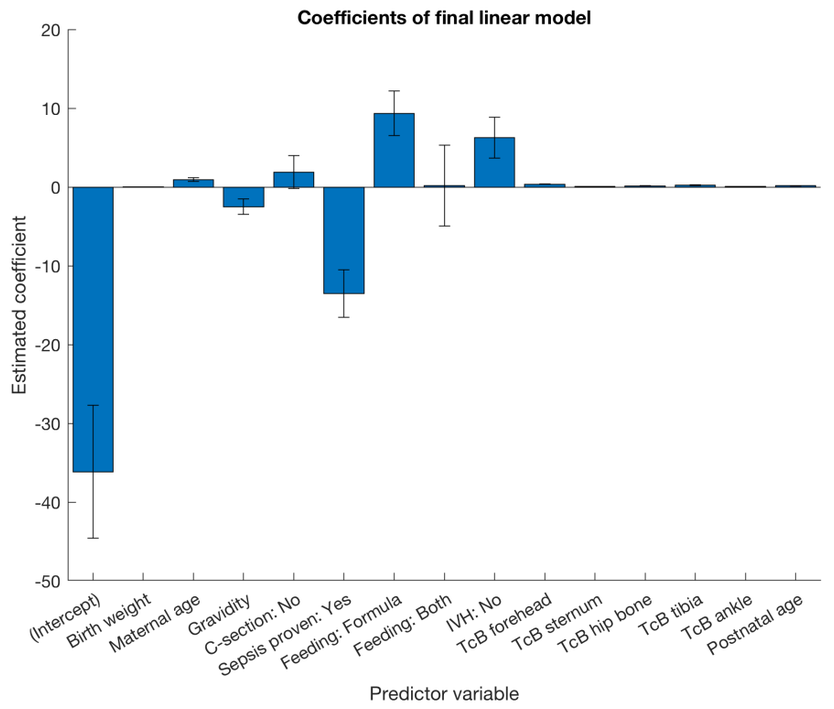


Figure 7: Final linear regression model coefficients with standard error

Note that this model does not provide information about the causality of the coefficient values. For example, it cannot be concluded that providing formula as the method of feeding increases the TSB of an infant by $9.34\,\mu\text{mol}\,\text{L}^{-1}$.

According to the Dutch pediatric guidelines, a TSB must be determined in a lab, when the $TcB+50$ reaches above the phototherapy threshold for an infant. It can therefore be said that collecting a blood sample is prevented, when the threshold is reached using the TcB device, but is not reached when applying the machine learning model. Table 1 shows the number of times the TcB measurement solely reaches the threshold, including the true positive, false positive, true negative, false negative. The TcB measurement on the forehead is used, since it is recommended by the manufacturer and the pediatric guidelines, and has the smallest error. The threshold used depends on the birth weight of the infant, as in appendix B, table 7. Table 1 and 2 show the predictions when using only the TcB device, and when applying the linear model, respectively. It can be seen that the number of false positives decreases drastically. Out of the 490 measurements, the TcB device required 243 blood samples to be

taken. The linear regression model reduced this number by 140, to 103. Out of these 140 prevented blood samples, 132 were appropriately marked as such, but 8 were wrongly prevented, as in table 3.

	True	False
Threshold reached	42	201
Threshold not reached	244	3

Table 1: TcB device prediction

	True	False
Threshold reached	34	69
Threshold not reached	376	11

Table 2: Linear model prediction

	True	False
Prevented blood samples	132	8

Table 3: The number of (rightly and wrongfully) prevented blood samples when using the linear model

A prevented blood sample occurs when the TcB value plus its margin reaches the threshold, but the model and its margin remains under the threshold, as in the following equation:

$$(TcB + 50) > \text{threshold AND } (TSB_{model} + error_{model}) < \text{threshold} \tag{9}$$

It is rightly predicted (true) when $TSB < threshold$, and wrongly (false) when $TSB > threshold$.

4.3 Decision tree model

The second type of model that is created, is a regression tree. It has been realized in the same way as the linear model, as described in section 3.1. During step 2, the hyperparameters can be tuned. The hyperparameters [15] that have been optimized during the development of this model are:

- `MinLeafSize`: Minimum number of leaf node observations (integer)
- `MaxNumSplits`: Maximal number of decision splits (integer)
- `Surrogate`: Surrogate decision splits flag (on/ off)
- `MergeLeaves`: Leaf merge flag (on/ off)

The first parameter `MinLeafSize` controls the tree depth. It is defined as the minimum amount of observations that make up one leaf node value. In terms of the visual representation of figure 3b, it is the amount of observations per partition. Increasing this parameter decreases the complexity of the tree. The second term is `MaxNumSplits`, which also controls the tree depth. It determines the maximum number of branch nodes in the tree model.

Setting the parameter `Surrogate` to "on", allows the model to adapt to missing data. This is done by selecting a next best variable (*surrogate*) to perform a decision split when the main variable is unknown. This is useful for sparse datasets, but requires more computational time and memory.

Lastly, the `MergeLeaves` parameter can be set to "on" to merge two leaf nodes which originate from the same parent branch node, when the sum of both MSEs is larger than the MSE of the parent node.

The hyperparameters are optimized by performing cross-validation for every single combination of hyperparameters (step 2). The hyperparameters which lead to the lowest RMSE are used to train an intermediate model (step 3). These steps are repeated to build a total of five intermediate models. The hyperparameters for every intermediate model, as well as the final model, can be found in appendix table 11. These models are then tested using their respective test sets. The results of these predictions are plotted in figure 8. The mean RMSE is 25.6 µmol L^{-1}, with a range of 19.5 to 30.4 µmol L^{-1}.

The predictions of a decision tree are discrete values. These are discretely distributed over the vertical axis. Therefore, there are as many horizontal lines in the prediction figure, as there are leaf nodes for that model.

On average, the decision tree models underestimate the TSB for 45.7% of the measurements. The distribution of the error can be found in figure appendix A 14b.

Lastly, the final model for the regression tree can be trained, using all available data (step 4). Before this can be done, however, the optimal hyperparameters should be determined. To do this, the complete dataset is again divided into $k = 10$ folds. The process of hyperparameter tuning (step 2) is applied to these folds. The combination of hyperparameters that yield the lowest RMSE, determine the optimal parameters. These are then used to train the final model with all available data. This model cannot be tested, since no test set has been held apart. The final model is graphically represented in figure 9, with the statements of each node in appendix B.

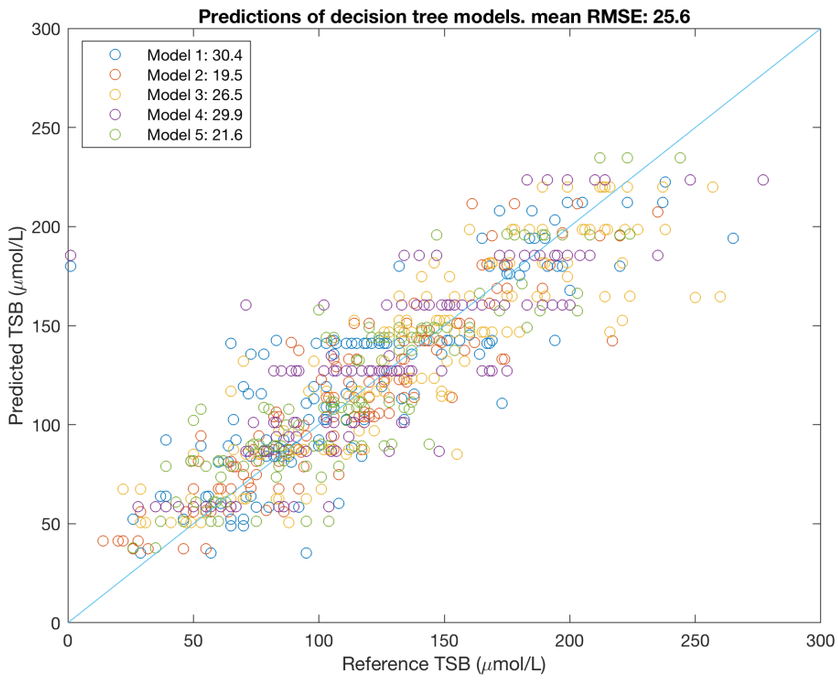


Figure 8: Predictions of the 5 intermediate decision tree models and their mean RMSE

During training, the machine learning determines which variables are best suited for use as a predictor variable. Apart from the 5 TcB body locations, 4 patient characteristics are used for the final model. These are as follows:

Continuous:

- Postnatal age (hours)
- Maternal age (years)
- Birth weight (grams)

Categorical:

- Feeding (formula/ breastfeeding/ both)

The importance of the predictor variables are estimated using the total accuracy each predictor brings to a model. The predictor importance of the final model can be found in appendix A figure 13. Since the TcB forehead is the first split that is performed, it has the highest contribution to the model, and thus has the highest predictor importance. The variables with a predictor importance of zero do not contribute to the model, and cannot be found in the split decisions.

The accuracy of this model is again determined using the accuracy of the worst-performing intermediate model. The RMSE of the final model can therefore estimated to be $30.4\,\mu\text{mol}\,\text{L}^{-1}$.

The same method of analysing how many blood samples are prevented, has been applied to this decision tree model. For a total of 636 measurements, 309 times the traditional TcB method called for a blood sample. Using the decision tree model, 181 blood samples are rightfully prevented, and 12 are wrongfully marked as not necessary.

	True	False
Threshold reached	55	254
Threshold not reached	324	3

Table 4: TcB device prediction

	True	False
Threshold reached	43	76
Threshold not reached	502	15

Table 5: Decision tree model prediction

	True	False
Prevented blood samples	181	12

Table 6: The number of (rightly and wrongfully) prevented blood samples when using the decision tree model

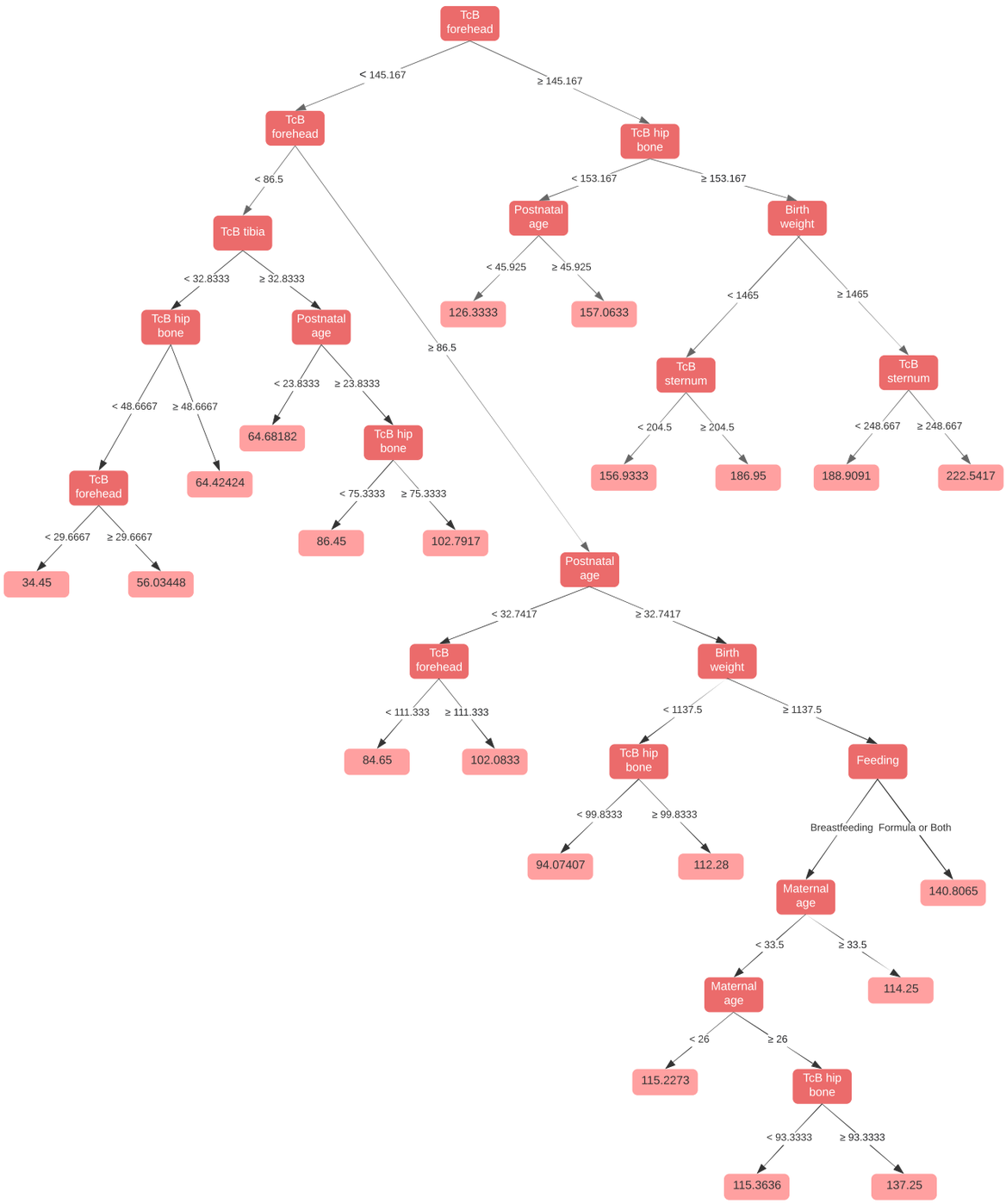


Figure 9: Graphical representation of the final decision regression tree

5 Discussion

During the process of making the models, five intermediate models were trained per model type. These were then tested with a test set to determine the accuracy. The research question for this assignment was to find one model that can be used to predict the TSB values. The consideration had to be made to either use one of those intermediate models as the final product, or to create a new final model. Training a final model has the advantage that it is trained using all of the available data. The downside is that its accuracy can only be estimated. It is desired to rather have a too high TSB prediction, than to miss the opportunity to treat the jaundice. Therefore the error margin for the final model can better be estimated too large than too tight. The choice was made to use the highest RMSE of the intermediate models as the RMSE of the final model. Perhaps better estimations of the accuracy could have been done, such as incorporating the mean and deviation in the intermediate models. It must be noted that there is a rather large spread between the RMSEs of the decision tree models.

Perhaps it would have been best not to create a new final model at all, but use one of the intermediate models as the final product. The question then still remains as to what the accuracy of that model is. Another assumption was that the method of hyperparameter optimisation also worked for the final model. The complete dataset was divided into 10 folds to perform cross-validation, just as how it was done for the intermediate models. The hyperparameters did not differ much from the ones used in the intermediate models.

The accuracy of both types of models can be compared to each other, because they are determined in the same way. They have undergone the same statistical analysis. The intermediate models were built using the same datasets: the partitioning over the train, validation and test set was identical for both models.

All the data was entered manually by the medical staff. Therefore the data is susceptible to human error. During the data processing, the most obvious outliers have been removed or fixed. For example, TSB values of 0 and 1 have been removed altogether, and the year 1018 has been changed to 2018. It is likely that not all mistakes have been removed, which will decrease the accuracy of the models.

Future research could be done to explore exactly what the limitations of both linear en decision tree models are. Perhaps they could compliment each other, by combining both models into one. Also other regression models might provide better predictions than the two used in this assignment. Moreover, the current research questions focused on predicting the actual TSB value based on the TcB measurements. Perhaps a more interesting research would have been to predict the need for treatment based on the TcB measurements. This is more applicable to the healthcare application. This would mean a classification model is trained instead of a regression model. For example, three predictions could be made: safe (low TSB), risk (blood test required), and high (jaundice treatment needed). The division of these zones could depend on the gestational age, postnatal age and birth weight, much like figure 2. However, the thresholds used by the different healthcare systems vary, so this model will probably not be able to be used worldwide.

One of the statistical analyses performed was similar to the above mentioned remark, as the predicted TSB value was used to determine if phototherapy is needed. This was then compared to the prediction done by using only the TcB value of the forehead. The limitations of this statistical analysis is that the threshold employed was only the stationary value of the phototherapy curve, based on the birth weight. It did not depend on the postnatal age, even though the guidelines specify this. Also, all infants were considered to be in the 'standard risk', even though some of them would likely have been a 'high risk'. This could not be incorporated, since that data was not part of the provided dataset. This statistical analysis was only used to illustrate the effects of using a model, and was not the desired output of the model. Moreover, the same data that was used to train the model, was used to calculate the prediction of the requirement for a blood sample. By using the models, the amount of blood samples that need to be taken is greatly reduced, but this also introduced some cases where a sample is not taken even though it should have. By increasing the error margin on the model, this number of false negative predictions is reduced, but less blood samples are prevented. A trade-off needs to be made to determine the desired error margin.

The theory shows that ethnicity is an important factor in the accuracy of the TcB measurements. Unfortunately, the ethnicity for a large group of the patients in this research is unknown. Moreover, there are few patients that have an ethnicity other than Caucasian. Therefore, this factor could not be used as an effective predictor variable in this study. The models would likely have been more accurate if this data would have been able to be used.

The specifications of the jaundice meter used state that there is a difference in accuracy between patients who have undergone phototherapy, and those who have not. This research did not take this factor into account when training the models. It would be interesting to see if a model can make more accurate predictions based on when and for how long a patient has undergone phototherapy.

6 Conclusion

During this assignment, two different machine learning models have been created. Both are regression models, which are part of supervised learning. The first is a linear regression model, which fits the predictors to a linearly proportional reference value. The predictors are TcB measurements of five different body locations and eight patient characteristics. The reference value is the actual TSB value. Five intermediate models have been trained using cross-validation, and tested. A final model is created using all available data. The root mean square of the final model is $21.9\,\mu\text{mol}\,\text{L}^{-1}$. When applying on the same training data, the linear model reduces the number of false positive blood sample indications from 201 to 69, when compared to using only a TcB device on the forehead.

The specifications of the TcB measurement device states that the accuracy for preterm infants is $27.4\,\mu\text{mol}\,\text{L}^{-1}$, and $39.0\,\mu\text{mol}\,\text{L}^{-1}$ after phototherapy. Without using a model to improve the accuracy, the best performing TcB measurement location is the forehead. The corresponding RMSE is $30.2\,\mu\text{mol}\,\text{L}^{-1}$. This is above the specified error for patients without phototherapy, but below the specified error with phototherapy. The linear regression model improves this error drastically, to $21.9\,\mu\text{mol}\,\text{L}^{-1}$.

Similarly, a decision tree model has been trained. It uses five TcB locations and four patient characteristics as predictor variables. The hyperparameters have been optimised using 10-fold cross-validation. The final model has an RMSE of $30.4\,\mu\text{mol}\,\text{L}^{-1}$. This is marginally worse than using only the TcB of the forehead. Nonetheless, the number of false positive indications for a blood sample is reduced from 254 to 76.

To answer the research questions, the linear regression model can be accepted as a method of improving the accuracy of TcB measurements in preterm infants. The RMSE is reduced from $30.2\,\mu\text{mol}\,\text{L}^{-1}$ (TcB forehead) to $21.9\,\mu\text{mol}\,\text{L}^{-1}$. The patient characteristics that are used in this model are: postnatal age, maternal age, C-section, feeding, gravidity, IVH, sepsis proven and birth weight.

References

[1] Charles I Okwundu et al. "Transcutaneous bilirubinometry versus total serum bilirubin measurement for newborns". In: *Cochrane Database of Systematic Reviews* (May 2017). DOI: 10.1002/14651858.cd012660. URL: https://doi.org/10.1002/14651858.cd012660.

[2] Michael Kaplan and Cathy Hammerman. "Hereditary Contribution to Neonatal Hyperbilirubinemia". In: *Fetal and Neonatal Physiology*. Elsevier, 2017, 933–942.e3. DOI: 10.1016/b978-0-323-35214-7.00097-4. URL: https://doi.org/10.1016/b978-0-323-35214-7.00097-4.

[3] Bolajoko O. Olusanya, Stephanie Teeple, and Nicholas J. Kassebaum. "The Contribution of Neonatal Jaundice to Global Child Mortality: Findings From the GBD 2016 Study". In: *Pediatrics* 141.2 (Jan. 2018), e20171471. DOI: 10.1542/peds.2017-1471. URL: https://doi.org/10.1542/peds.2017-1471.

[4] Ali Ahmed Raba, Anne O'Sullivan, and Jan Miletin. "Prediction of the need for phototherapy during hospital stay in preterm infants by transcutaneous bilirubinometry". In: *Early Human Development* 146 (July 2020), p. 105029. DOI: 10.1016/j.earlhumdev.2020.105029. URL: https://doi.org/10.1016/j.earlhumdev.2020.105029.

[5] Stephanie McClean et al. "Cost savings with transcutaneous screening versus total serum bilirubin measurement for newborn jaundice in hospital and community settings: a cost-minimization analysis". In: *CMAJ Open* 6.3 (2018), E285–E291. DOI: 10.9778/cmajo.20170158. URL: https://doi.org/10.9778/cmajo.20170158.

[6] E T Schmidt et al. "Evaluation of transcutaneous bilirubinometry in preterm neonates". In: *Journal of Perinatology* 29.8 (Mar. 2009), pp. 564–569. DOI: 10.1038/jp.2009.38. URL: https://doi.org/10.1038/jp.2009.38.

[7] Marlijn D. van Erk et al. "How skin anatomy influences transcutaneous bilirubin determinations: an in vitro evaluation". In: *Pediatric Research* 86.4 (June 2019), pp. 471–477. DOI: 10.1038/s41390-019-0471-z. URL: https://doi.org/10.1038/s41390-019-0471-z.

[8] S Samiee-Zafarghandy et al. "Influence of skin colour on diagnostic accuracy of the jaundice meter JM 103 in newborns". In: *Archives of Disease in Childhood - Fetal and Neonatal Edition* 99.6 (July 2014), F480–F484. DOI: 10.1136/archdischild-2013-305699. URL: https://doi.org/10.1136/archdischild-2013-305699.

[9] Tanja Karen, Hans Ulrich Bucher, and Jean-Claude Fauchère. "Comparison of a new transcutaneous bilirubinometer (Bilimed®) with serum bilirubin measurements in preterm and full-term infants". In: *BMC Pediatrics* 9.1 (Nov. 2009). DOI: 10.1186/1471-2431-9-70. URL: https://doi.org/10.1186/1471-2431-9-70.

[10] Aidan Wilson. *A Brief Introduction to Supervised Learning*. Accessed November 10, 2020. 2019. URL: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590.

[11] Alida J. Dam-Vervloet et al. "Inter-device reproducibility of transcutaneous bilirubin meters". In: *Pediatric Research* (Sept. 2020). DOI: 10.1038/s41390-020-01118-6. URL: https://doi.org/10.1038/s41390-020-01118-6.

[12] BabyZietGeel Richtlijn Hyperbilirubinemie. *Bilicurves prematuren < 35 wkn*. Accessed January 18, 2021. 2008. URL: http://babyzietgeel.nl/kinderarts/hulpmiddelen/diagnostiek/bilicurves_prematuren.php.

[13] Costantino Romagnoli et al. "Italian guidelines for management and treatment of hyperbilirubinaemia of newborn infants ≥ 35 weeks' gestational age". In: *Italian Journal of Pediatrics* 40.1 (Jan. 2014). DOI: 10.1186/1824-7288-40-11. URL: https://doi.org/10.1186/1824-7288-40-11.

[14] MathWorks. *Fit linear regression model*. Accessed November 10, 2020. 2020. URL: https://nl.mathworks.com/help/stats/fitlm.html.

[15] MathWorks. *Fit binary decision tree for regression*. Accessed November 10, 2020. 2020. URL: https://nl.mathworks.com/help/stats/fitrtree.html.

[16] Leo Breiman et al. *Classification And Regression Trees*. Routledge, Oct. 2017. DOI: 10.1201/9781315139470. URL: https://doi.org/10.1201/9781315139470.

[17] MathWorks. *Predict responses of linear regression model*. Accessed November 10, 2020. 2020. URL: https://nl.mathworks.com/help/stats/linearmodel.predict.html.

[18] BabyZietGeel Richtlijn Hyperbilirubinemie. *Bilicurve > 35 wkn*. Accessed January 18, 2021. 2008. URL: http://babyzietgeel.nl/kinderarts/hulpmiddelen/diagnostiek/bilicurve35wkn.php.

[19] Draegerwerk AG & Co. KGaA. *Draeger Jaundice Meter JM-105*. Accessed November 30, 2020. 2020. URL: https://www.draeger.com/en-us_us/Products/Jaundice-Meter-JM-105.

[20] MD Mary L. Gavin. *What Is the Apgar Score?* Accessed November 19, 2020. 2018. URL: https://kidshealth.org/en/parents/apgar.html.

A Figures

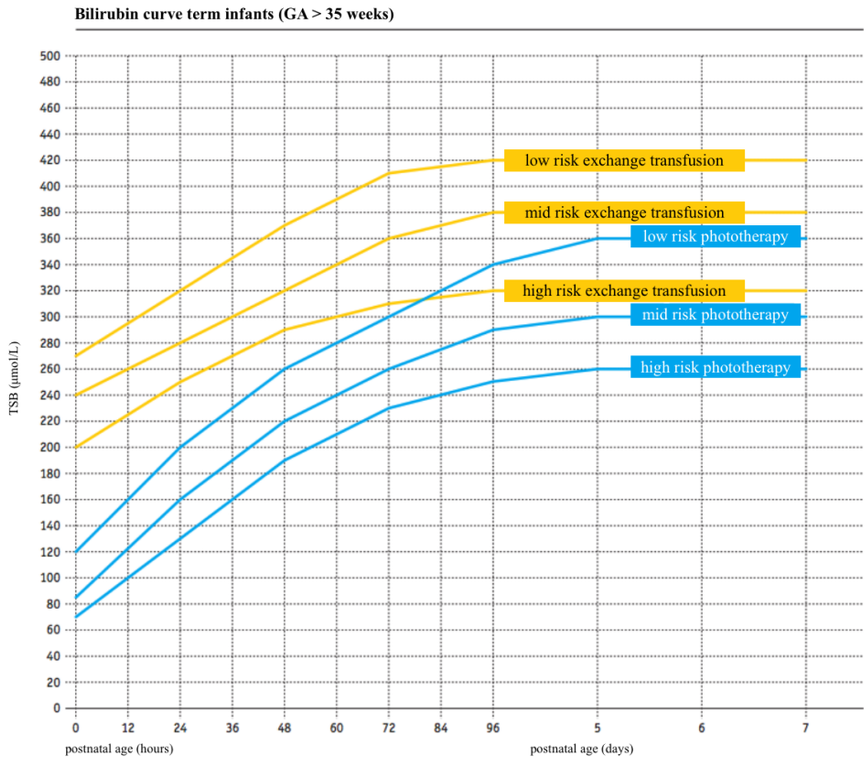
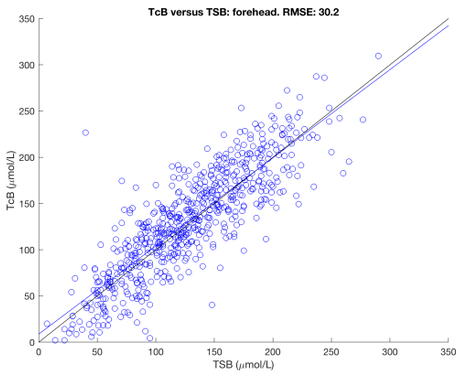
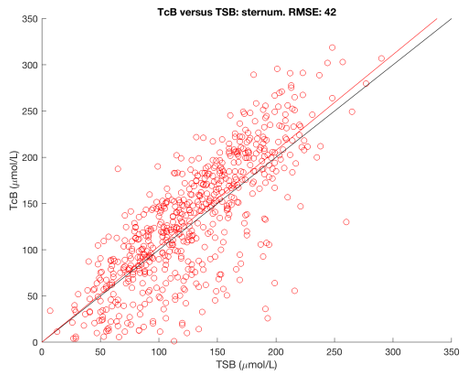


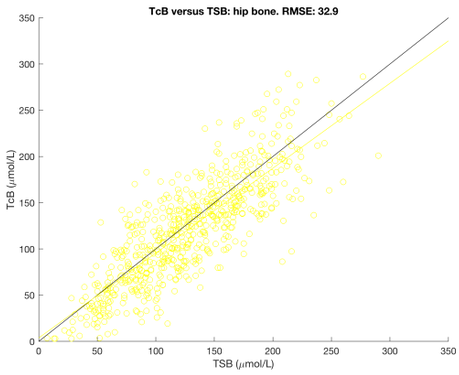
Figure 10: Chart showing the phototherapy and exchange transfusion thresholds for term infants, based on postnatal age for three levels of risk, adapted from Dutch pediatric guidelines [18]



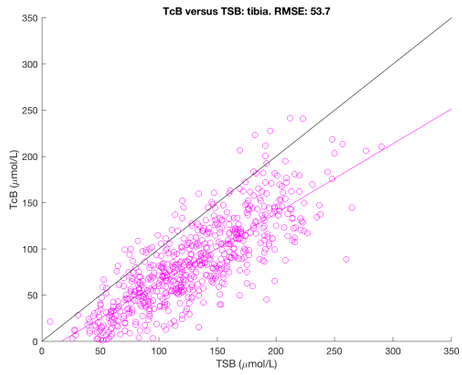
(a) Forehead



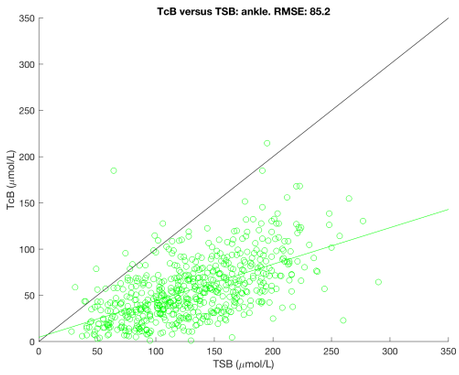
(b) Sternum



(c) Hip bone



(d) Tibia



(e) Ankle

Figure 11: TcB-TSB plot for every body measurement location

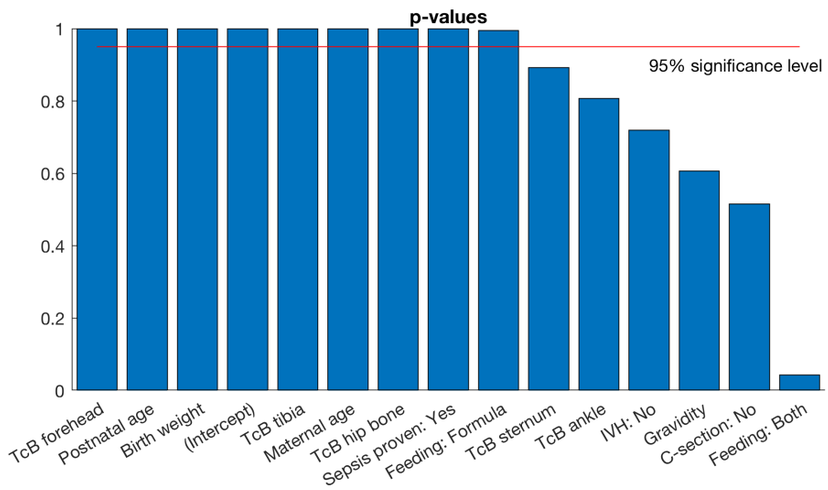


Figure 12: p-values testing the hypothesis, with a 95% significance level

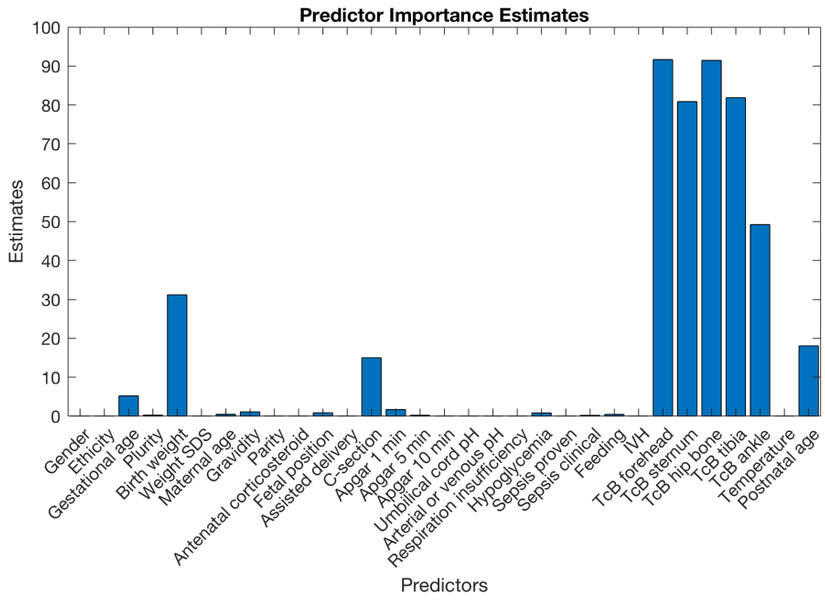
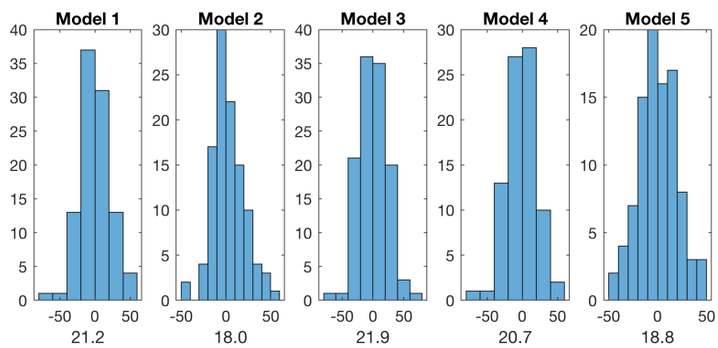
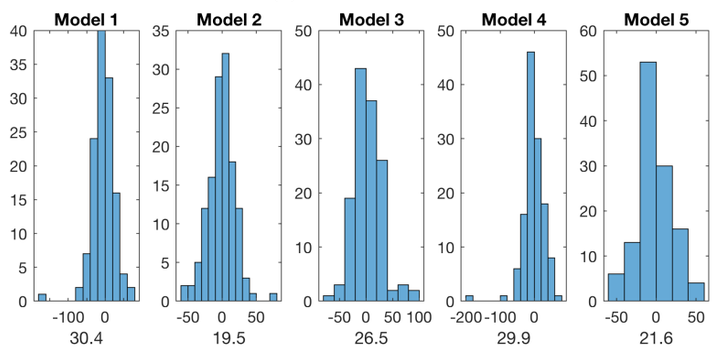


Figure 13: Predictor importance for the final decision tree model



(a) Linear models



(b) Decision tree models

Figure 14: Histogram of error in TSB prediction, with RMSE value below the graph

B Tables

Birth weight (gr)	Standard risk	High risk
<1000	100	100
1000 - 1249	150	100
1250 - 1499	190	150
1500 - 2000	220	190
>2000	240	220

Table 7: Stationary value of bilirubin thresholds for preterm infants ($\mu mol\,L^{-1}$), from Dutch pediatric guidelines [12]

	Low risk	Standard risk	High risk
Term infant	360	300	260

Table 8: Stationary value of bilirubin thresholds for term infants ($\mu mol\,L^{-1}$), from Dutch pediatric guidelines [18]

Measurement range	0 $\mu mol\,L^{-1}$ to 340 $\mu mol\,L^{-1}$	
Accuracy	$\pm$ 27.4 $\mu mol\,L^{-1}$	$\geq$ 24 weeks gestation
	$\pm$ 25.5 $\mu mol\,L^{-1}$	>35 weeks gestation
Accuracy after phototherapy	$\pm$ 39.0 $\mu mol\,L^{-1}$	$\geq$ 24 weeks gestation
	$\pm$ 38.0 $\mu mol\,L^{-1}$	>35 weeks gestation

Table 9: Draeger Jaundice Meter JM-105 specifications [19]

	Coefficient	Standard Error	p-value
(Intercept)	-36.14	8.46	1.00
Postnatal age (hours)	0.18	0.02	1.00
Maternal age (years)	0.95	0.25	0.99
C-section: No	1.89	2.08	0.64
Feeding: Both	0.18	5.12	0.03
Feeding: Formula	9.34	2.84	0.99
Gravidity	-2.50	0.98	0.99
IVH: No	6.28	2.58	0.98
Sepsis proven: Yes	-13.53	3.01	1.00
TcB ankle ($\mu mol\,L^{-1}$)	0.06	0.04	0.89
TcB forehead ($\mu mol\,L^{-1}$)	0.36	0.03	1.00
TcB hip bone ($\mu mol\,L^{-1}$)	0.16	0.04	1.00
TcB sternum ($\mu mol\,L^{-1}$)	0.06	0.03	0.97
TcB tibia ($\mu mol\,L^{-1}$)	0.23	0.05	1.00
Birth weight (grams)	0.02	0.00	1.00

Table 10: Final linear regression model coefficients, with the corresponding standard error and p-value

	Model 1	Model 2	Model 3	Model 4	Model 5	Final model
MaxNumSplits	37	37	19	7	26	19
MinLeafSize	1	4	20	4	10	20
Surrogate	on	on	off	on	on	on
MergeLeaves	on	on	on	on	on	on

Table 11: Hyperparameters of decision tree models

Final decision tree model

1. if $avgVH < 145.167$ then node 2 elseif $avgVH >= 145.167$ then node 3 else 127.225
2. if $avgVH < 86.5$ then node 4 elseif $avgVH >= 86.5$ then node 5 else 96.9974
3. if $avgHB < 153.167$ then node 6 elseif $avgHB >= 153.167$ then node 7 else 171.802
4. if $avgTIBIA < 32.8333$ then node 8 elseif $avgTIBIA >= 32.8333$ then node 9 else 67.9662
5. if $age_{hrs} < 32.7417$ then node 10 elseif $age_{hrs} >= 32.7417$ then node 11 else 115.597
6. if $age_{hrs} < 45.925$ then node 12 elseif $age_{hrs} >= 45.925$ then node 13 else 149.236
7. if $weight_{gr} < 1465$ then node 14 elseif $weight_{gr} >= 1465$ then node 15 else 187.642
8. if $avgHB < 48.6667$ then node 16 elseif $avgHB >= 48.6667$ then node 17 else 54.1463
9. if $age_{hrs} < 23.8333$ then node 18 elseif $age_{hrs} >= 23.8333$ then node 19 else 85.1364
10. if $avgVH < 111.333$ then node 20 elseif $avgVH >= 111.333$ then node 21 else 94.1591
11. if $weight_{gr} < 1137.5$ then node 22 elseif $weight_{gr} >= 1137.5$ then node 23 else 120.642
12. fit = 126.333
13. fit = 157.063
14. if $avgST < 204.5$ then node 24 elseif $avgST >= 204.5$ then node 25 else 168.94
15. if $avgST < 248.667$ then node 26 elseif $avgST >= 248.667$ then node 27 else 196.901
16. if $avgVH < 29.6667$ then node 28 elseif $avgVH >= 29.6667$ then node 29 else 47.2245
17. fit = 64.4242
18. fit = 64.6818
19. if $avgHB < 75.3333$ then node 30 elseif $avgHB >= 75.3333$ then node 31 else 95.3636
20. fit = 84.65
21. fit = 102.083
22. if $avgHB < 99.8333$ then node 32 elseif $avgHB >= 99.8333$ then node 33 else 102.827
23. if $Feeding = Breastfeeding$ then node 34 elseif $FeedinginFormulaMixofabove$ then node 35 else 127.504
24. fit = 156.933
25. fit = 186.95
26. fit = 188.909
27. fit = 222.542
28. fit = 34.45
29. fit = 56.0345
30. fit = 86.45
31. fit = 102.792
32. fit = 94.0741
33. fit = 112.28
34. if $age_{mother} < 33.5$ then node 36 elseif $age_{mother} >= 33.5$ then node 37 else 123.538
35. fit = 140.806
36. if $age_{mother} < 26$ then node 38 elseif $age_{mother} >= 26$ then node 39 else 125.75
37. fit = 114.25
38. fit = 115.227
39. if $avgHB < 93.3333$ then node 40 elseif $avgHB >= 93.3333$ then node 41 else 129.484
40. fit = 115.364
41. fit = 137.25

	Mean (standard deviation)	*Median (range)*
Gender (n)		
Male	60	
Female	41	
Ethnicity (n)		
African	1	
Azian	1	
Caucasian	54	
Latin-American	1	
Turkish	1	
Other	2	
Unknown	41	
Plurity (n)		
Singleton	62	
Twin	32	
Triplet	6	
Unknown	1	
Sepsis (n)		
Proven	8	
Clinical	27	
No	59	
Unknown	7	
Feeding (n)		
Breastfeeding	81	
Formula	14	
Both	3	
Unknown	3	
C-section (n)		
Yes	57	
No	43	
Unknown	1	
IVH (n)		
Yes	13	
No	84	
Unknown	4	
Birth weight (grams)	1518 (447.5)	1450 (675-3280)
Gestational age (weeks)	30.71 (1.71)	30.5 (28.0-35.7)
Maternal age (years)	29.98 (4.19)[A]	29 (21-39)[A]
Gravidity	1.86 (1.19)	2 (1-6)
Parity	0.53 (0.70)	0 (0-3)
Measurement moments per patient	17.0 (6.7)	16 (6-38)
Observation period per patient (days)	6.3 (2.3)	6.3 (2.0-12.9)

[A]: 1 unknown value

Table 12: Patient characteristics

Predictor variable	*Units/ values*	*Description*
Continuous		
Gestational age	weeks	age of the pregnancy
Postnatal age	hours	age infant since birth
Birth weight	grams	weight of infant at time of birth
Weight SDS		birth weight Standard Deviation Score: number of standard deviations from mean birth weight
Maternal age	years	age of mother at time of birth
Gravidity		number of times the mother has been pregnant
Parity		number of times the mother has given birth to a fetus with gestational age >= 24 weeks
Apgar 1 min	0-10	Apgar score[B] 1 minute after birth
Apgar 5 min	0-10	Apgar score[B] 5 minute after birth
Apgar 10 min	0-10	Apgar score[B] 10 minute after birth
Umbilical cord pH	pH	
Temperature	Celcius	body temperature (if known: rectal, else via skin)
TcB forehead	μmol L^{-1}	mean of three TcB measurements on the forehead
TcB sternum	μmol L^{-1}	mean of three TcB measurements on the sternum
TcB hip bone	μmol L^{-1}	mean of three TcB measurements on the hip bone
TcB tibia	μmol L^{-1}	mean of three TcB measurements on the tibia
TcB ankle	μmol L^{-1}	mean of three TcB measurements on the ankle
Binary		
Gender	male/ female	
C-section	yes/ no	Patient underwent a Cesarean delivery
Hypoglycemia	yes/ no	Low blood sugar
Sepsis proven	yes/ no	Diagnosed sepsis
Sepsis clinical	yes/ no	Clinical expectation of a sepsis
IVH	yes/ no	Intraventricular hemorrhage: bleeding in the brain
Higher order		
Ethnicity	African/ Asian/ Caucasian/ Latin-American/ Turkish/ other/ unknown	
Fetal position	head/ breech/ transverse lie	
Assisted delivery	yes/ no/ unknown	Delivery assisted using ventouse or forceps
Feeding	breastfeeding/ formula/ both	
Plurity	singleton/ twin/ triplet	
Antenatal corticosteroid treatments	yes/ no/ incomplete	medication taken by mother to reduce risk of breathing difficulties in preterm babies
Respiratory insufficiency	No/ ventilation/ "NIPPV/CPAP/LF/HF"[C]	Ventilation support needed due to a respiratory insufficiency

[B]: APGAR is a method to indicate health of a newborn right after birth: [20]

Appearance (skin color)	0-2
Pulse (heart rate)	0-2
Grimace response (reflexes)	0-2
Activity (muscle tone)	0-2
Respiration (breathing rate and effort)	0-2

[C]: NIPPV/CPAP/LF/HF are forms of ventilation support

Table 13: Predictor variables