Multinomial Logistic Regression, Neural Networks, Decision Trees, and Random Forests to Predict Children's Anemia Levels

Introduction

The purpose of this paper was to compare the accuracy of a multinomial logistic regression model to a neural network Model for predicting children's anemia levels. The dataset used in this paper is *Factors Affecting Children Anemia Level* and the primary purpose of this dataset is to investigate the relationship between predictor variables and a child's anemia level. A categorical response variable limits the possibilities for model creation. Four of the most commonly used methods for predicting categorical response variables are multinomial logistic regression, neural networks, decision trees, and random forests. Models were created using these four methods, and predictive accuracy was used to evaluate the stronger model.

Data

Factors Affecting Children Anemia Level comes from the 2018 Nigeria Demographic and Health Surveys and specifically was created to answer research questions about the different factors that may have a relationship with children's anemia level. This data has a categorical response variable of Anemia Level with four different levels: Not anemic, Mild, Moderate, and Severe. The relevant predictor variables are: Mother's Age, Mother's Residence, Mother's Education Level, Mother's Wealth Level, Mother's Number of Previous Births in the Past 5 Years, Child's Age, Hemoglobin Level, Mosquito Bed Net Status, and Mother's Smoking Status.

The data had to be cleaned for this paper. First the data was limited to 2000 rows in order to accommodate a reasonable sample size to run a neural network in R. Next, all rows with empty values for Anemia Level were removed. After this, Mother's Age was converted from an age range to an average age for each age range in order to be able to treat this variable as numeric. From here, the data was split into testing and training data.

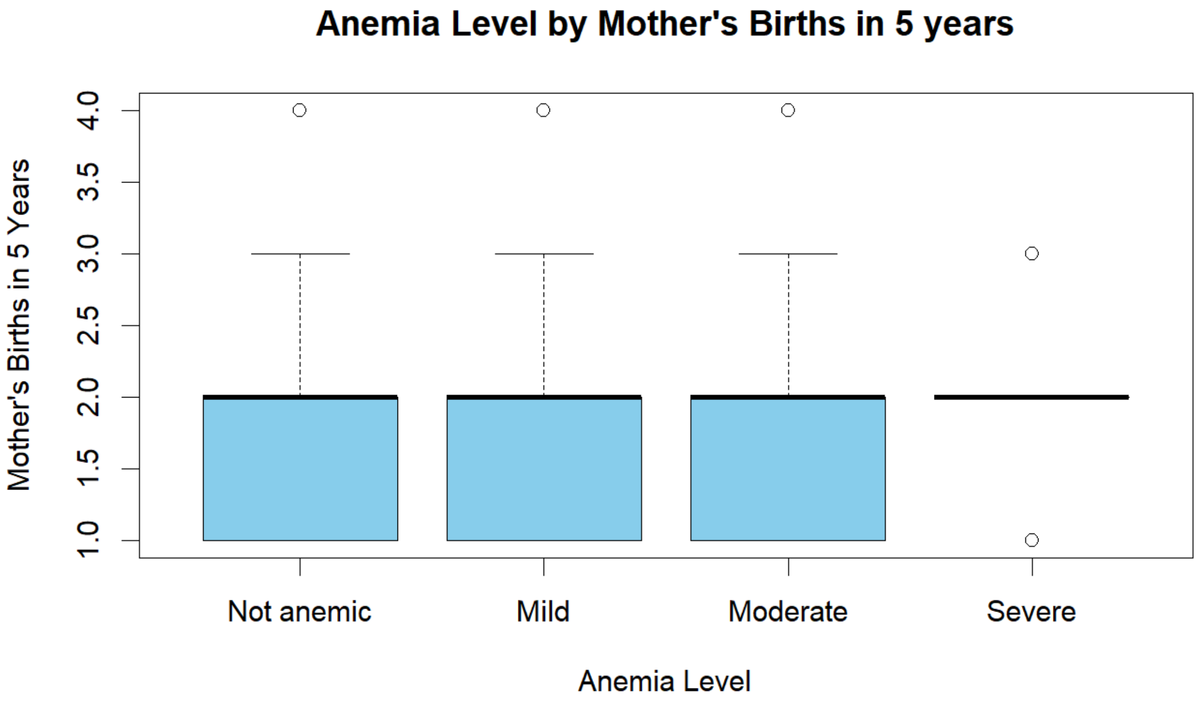
For the multinomial logistic regression, Anemia Level, Mother's Education Level, and Mother's Wealth Level were converted to ordinal categorical variables as they are ordinal categorical variables.

For the neural network model, these variables were left as standard categorical variables because neural networks cannot accommodate categorical predictor variables. Because of this, all categorical variables had to be transformed into dummy variables in order to be fed into the neural network model except for the response variable Anemia Level.

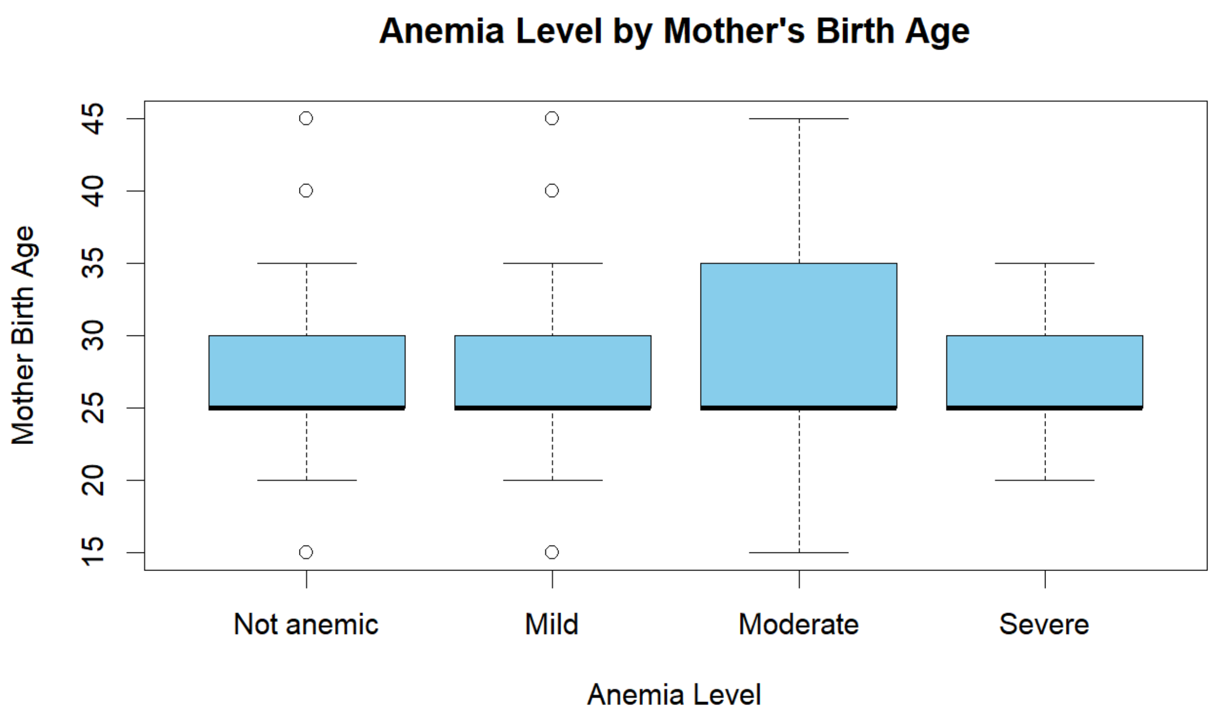
The decision tree and random forest models used the same transformations for multinomial logistic regression, as well as comforting anemia into a factor variable in order to make the decision tree and random forests classification models.

Diagnostics

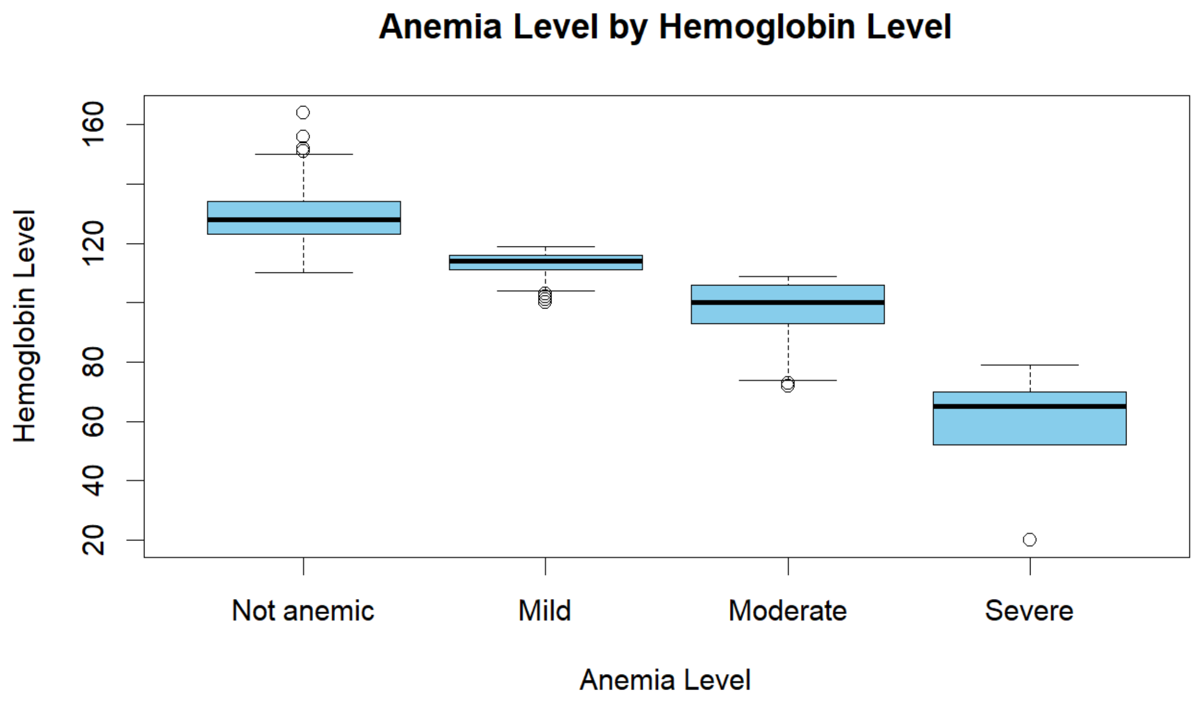
The relationship between predictor and response variables was investigated through plots. Anemia Level is an ordinal categorical variable, so for continuous numeric variables, a bar plot was used to look at the relationship. For categorical predictor variables, a bar plot with color indicating the categorical variable with a legend was used to look at the relationship.



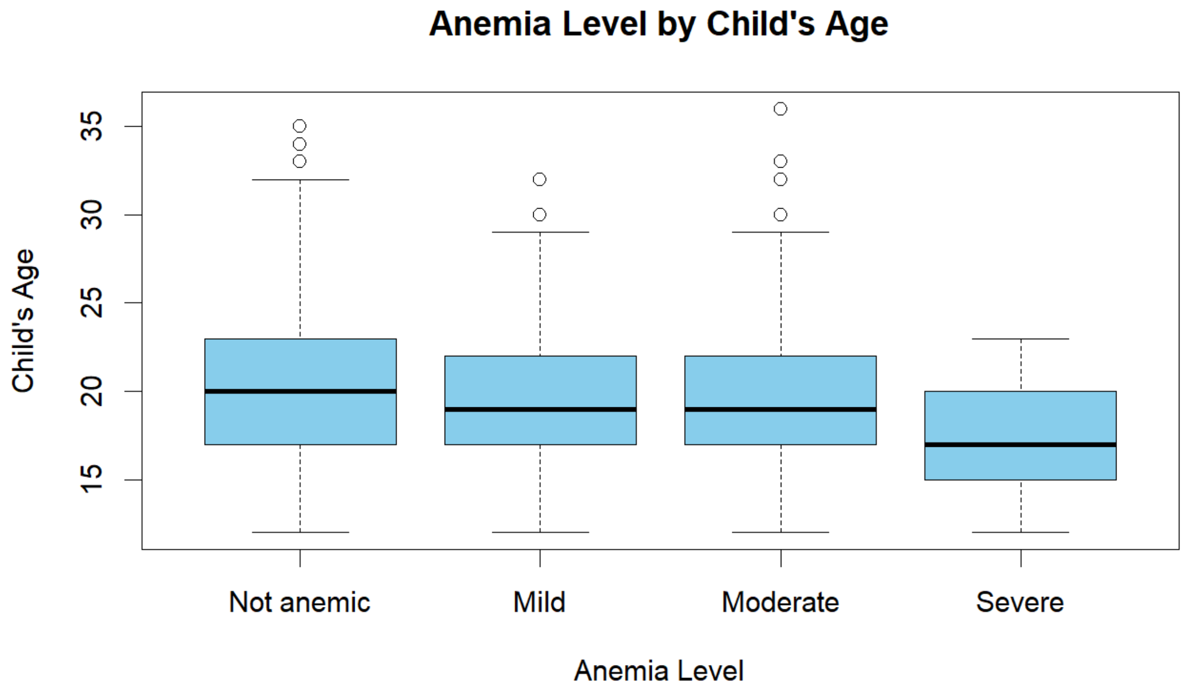
Mother’s Birth’s in 5 Years does not seem to have much of a relationship with Anemia Level based on this figure.



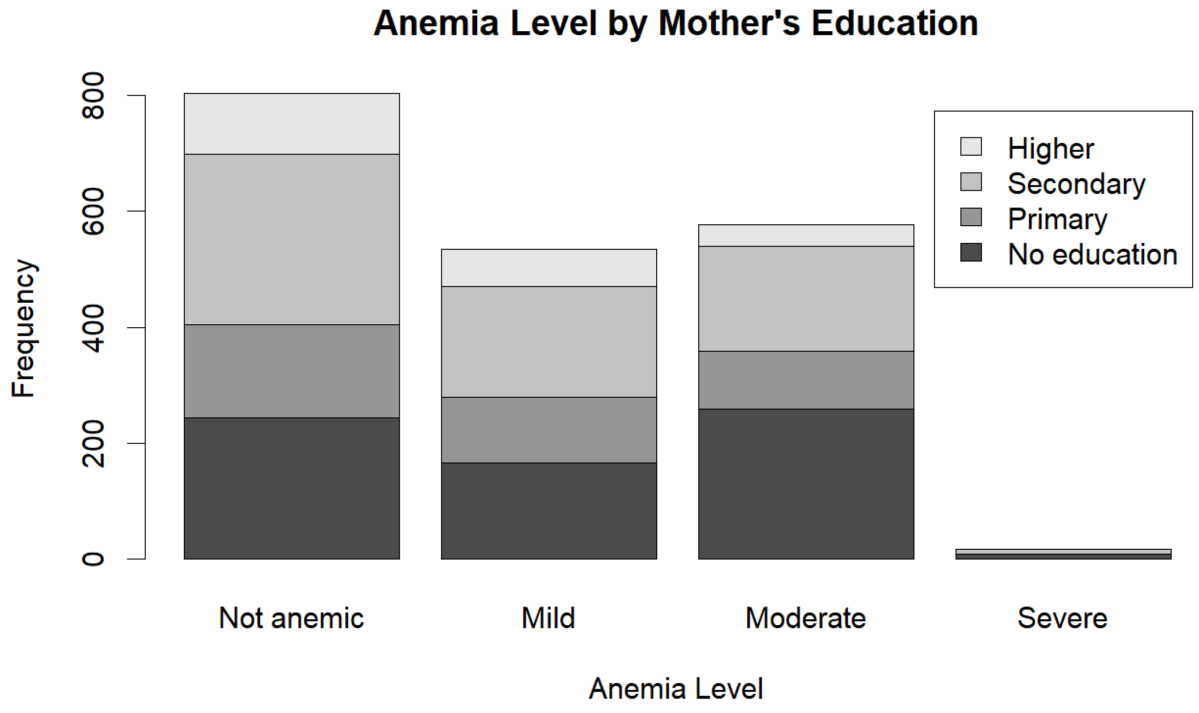
Mother’s birth age seems to possibly have a slightly positive relationship with Anemia Level.



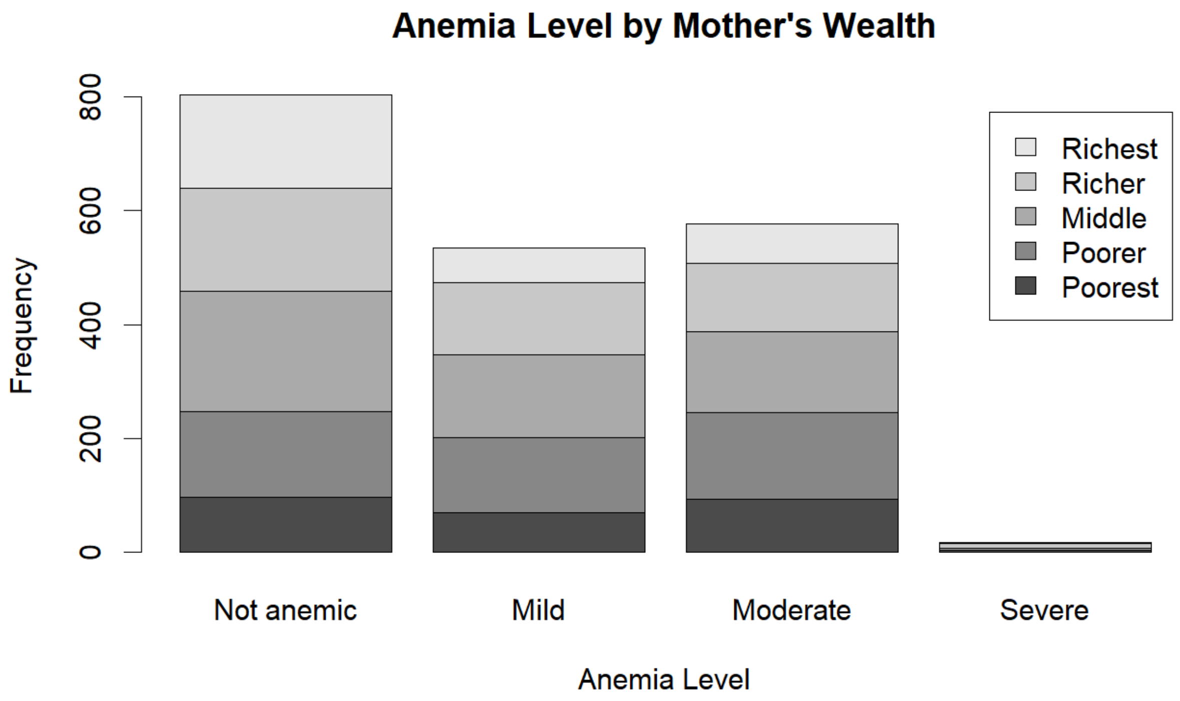
Hemoglobin Level and Anemia Level seem to have a fairly strong negative relationship. Hemoglobin Level will likely be a very strong predictor of Anemia Level because of this.



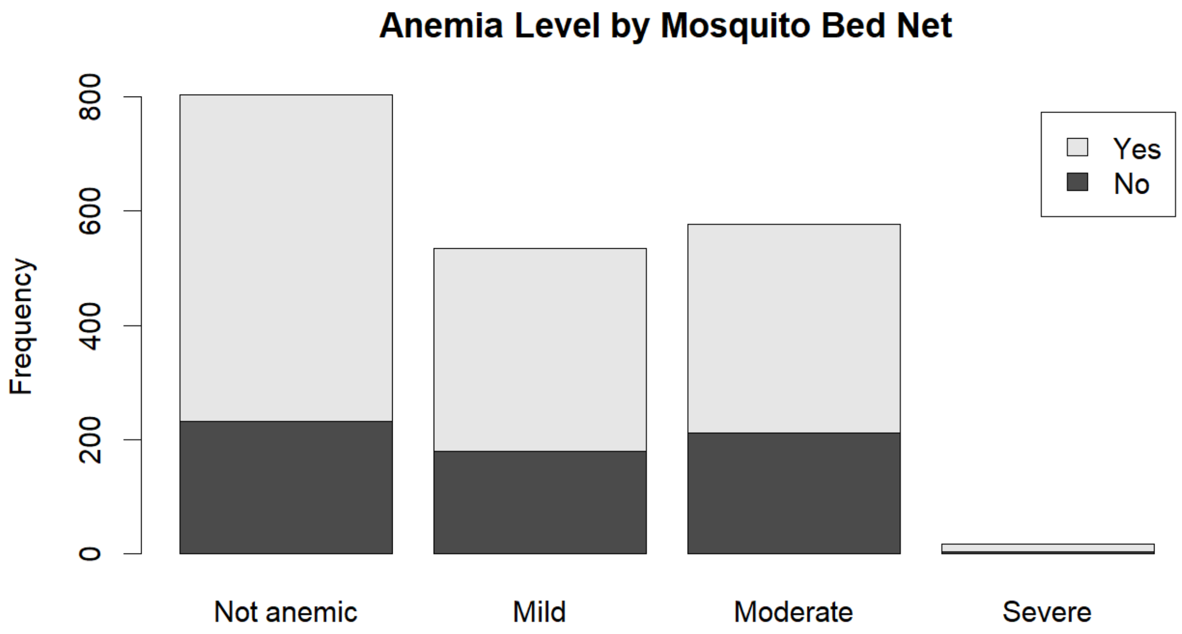
Child’s Age may have a small negative relationship with Anemia Level as shown in this graph.



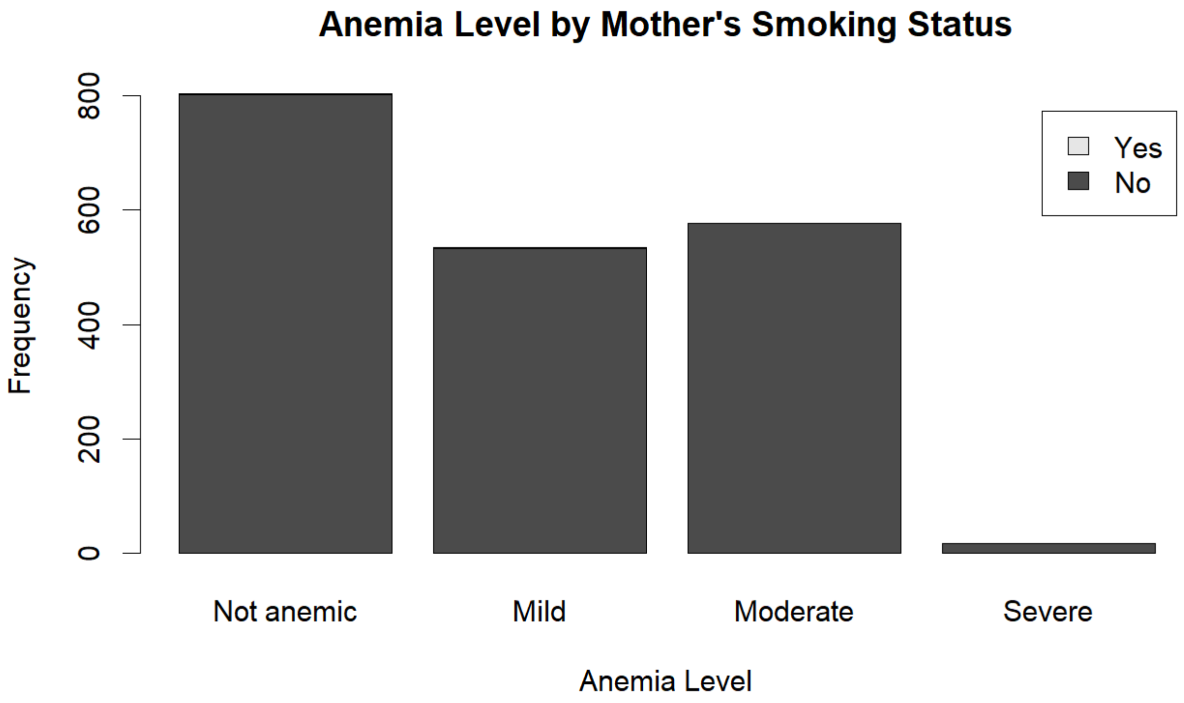
Mother’s Education Level seems to have a relationship with Anemia Level. The more severe Anemia Level bars have a higher portion of Mother’s Education Level being No education.



Mother’s Wealth seems to have a relationship with Anemia Level. Those with higher wealth levels seem to have less severe anemia. This would validate the expected relationship as those who have wealth can afford better healthcare.



This graph shows that a higher portion of those with less severe anemia levels have Mosquito Bed Nets than those with more severe anemia levels.



This graph does not show a strong relationship simply because there are very few smoking mothers. However, Mother's Smoking Status could still be a strong predictor.

Multinomial Logistic Model

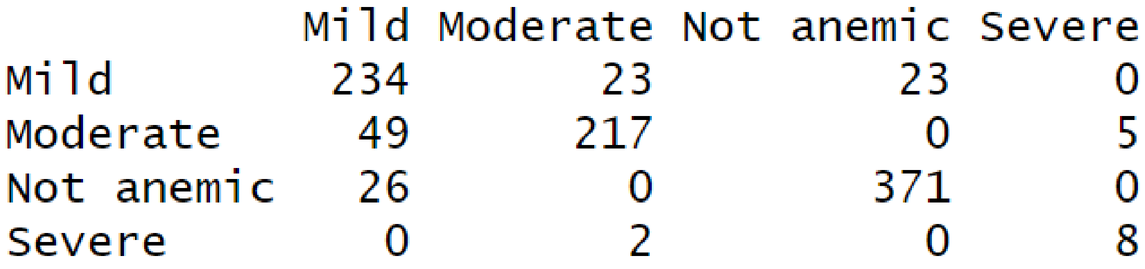
A multinomial logistic model was used to predict Anemia Level in this dataset because it is one of the few methods to predict a non-binary response categorical variable. A multinomial logistic model takes in a vector of data and returns a score that can be converted into a probability value. For prediction, the outcome probability value that is highest given the prediction data is selected as the prediction. There are several assumptions required for multinomial logistic regression. First, observations must be independent, we assume this is true and that the data was collected in a reasonably statistically consistent way. Next, the different outcomes must be mutually exclusive and fit all possible outcomes. This is true for the response variable Anemia Level. Next is that there are no outliers or highly influential points. There were no large outliers found in the diagnostic section of this paper.

The model for Multinomial Logistic Regression is as follows:

$$Pr(Y_i = k) = \frac{e^{\beta_k \cdot X_i}}{1 + \sum_{j=1}^{K-1} e^{\beta_k \cdot X_i}}, \; k \leq K$$

In this equation, k is the different possible values of the response variable Anemia Level, β_k is vector of regression coefficients responding to the outcome k and the score(X_i, k), and X_i is the vector of predictor variables. The score can be converted into a probability value which is used for prediction in this model.

Multiple multinomial logistic models were created with varying predictors using training data, and the model with the lowest AIC was chosen as the strongest model. This model used all predictors. The accuracy was then calculated by creating a confusion matrix of predicted response and actual response using testing data:



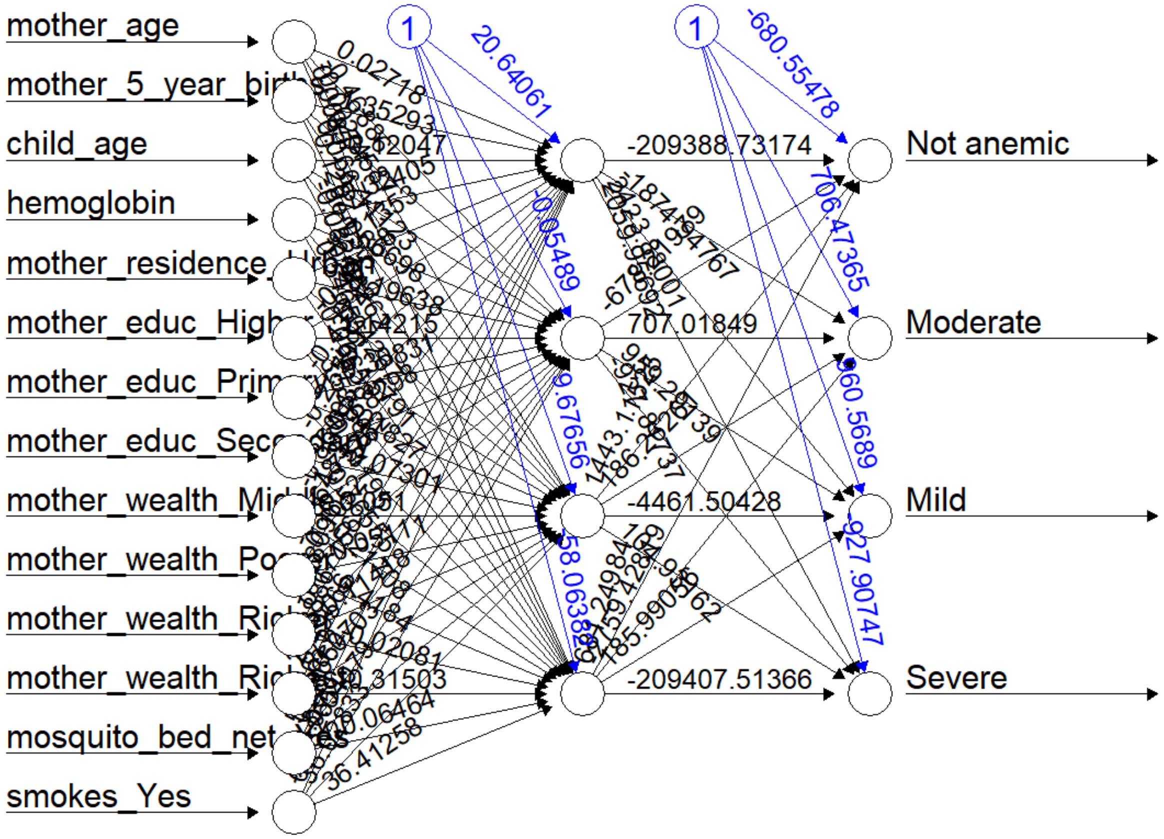
```
            Mild Moderate Not anemic Severe
Mild         234       23         23      0
Moderate      49      217          0      5
Not anemic    26        0        371      0
Severe         0        2          0      8
```

The multinomial logistic model was calculated to have an accuracy of 86.64%. This accuracy will be compared to the accuracy of the neural network model in the conclusion section.

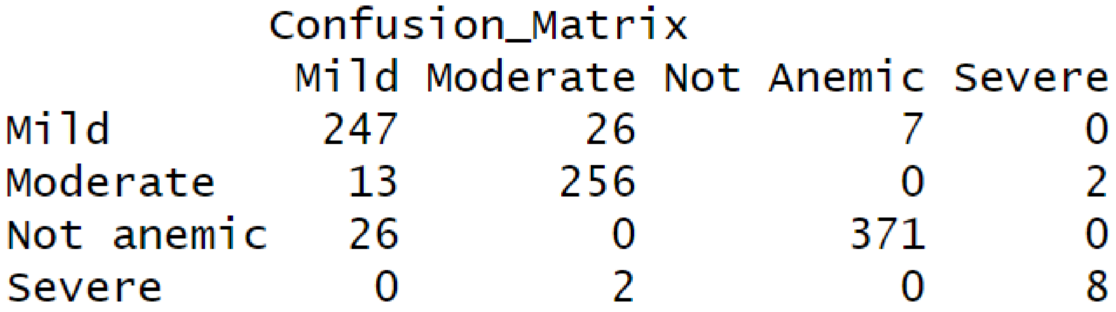
Neural Network Model

A neural network model was created using the training data to predict Anemia Level as it is another of the few ways to predict a non-binary categorical variable. A neural network takes in the predictor variables as the input layer, processes them through a hidden layer, and sends them to the output layer. In addition, neural networks do not rely on assumptions about the data, so there are no assumptions made about the data for this model.

For the input layer, to make an accurate comparison between the multinomial logistic model, the same predictor variables were used as the input layer of the neural network model. The output layer is simply the four different outcomes of Anemia Level. From here, many different models were created with varying hidden layers. This included varying the number of nodes in the hidden layer and varying the number of hidden layers. The model that had the highest accuracy used only one hidden layer with four nodes. The diagram for this neural network is crowded due to the large number of input layers:



The accuracy was then calculated by creating a confusion matrix of predicted response and actual response using testing data:

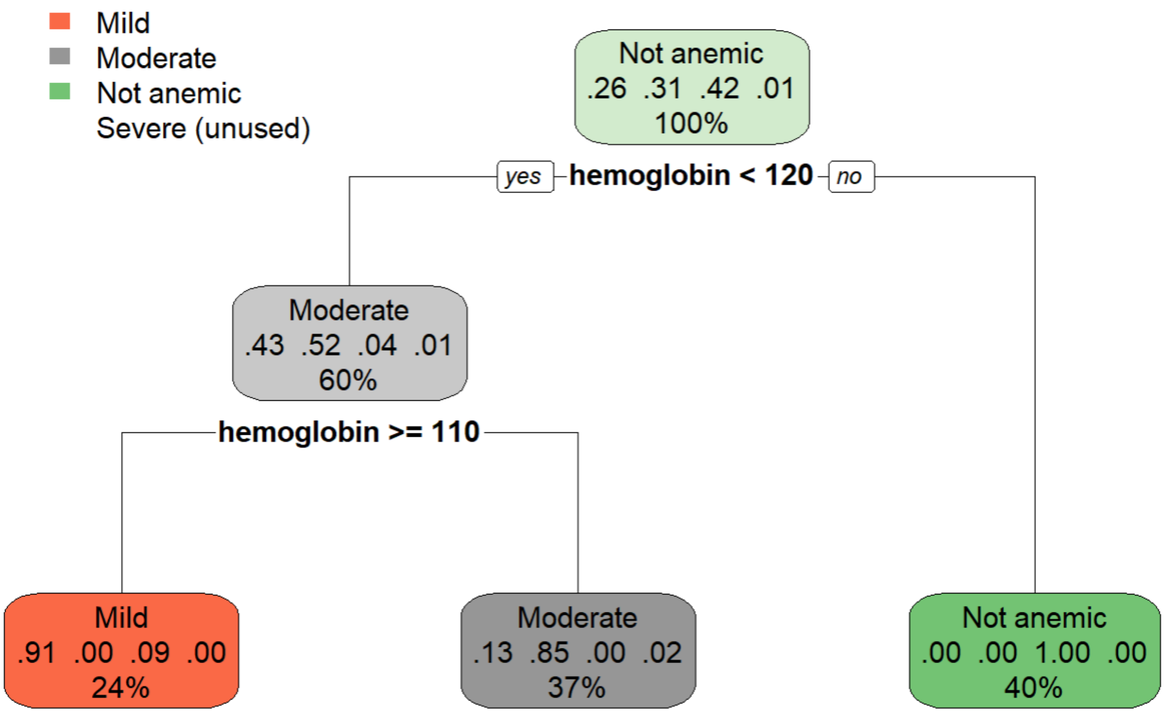

```
            Confusion_Matrix
             Mild Moderate Not Anemic Severe
  Mild        247       26          7      0
  Moderate     13      256          0      2
  Not anemic   26        0        371      0
  Severe        0        2          0      8
```

The accuracy of the neural network model is 91.75%.

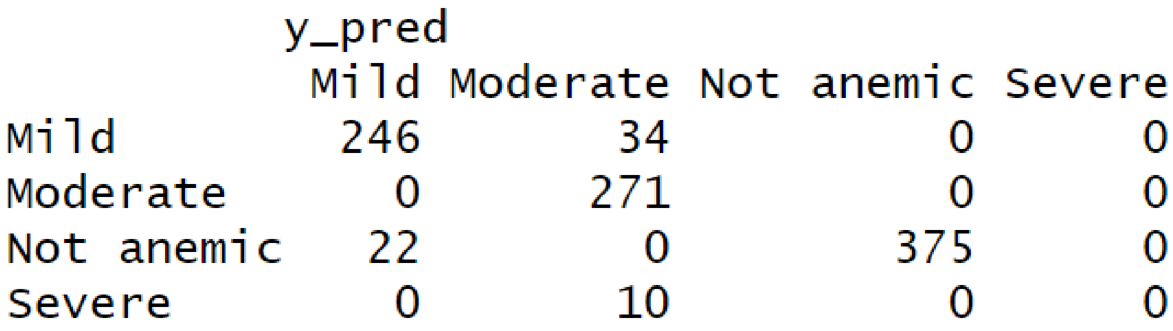
Decision Tree Model

A decision tree model was created using the training data to predict Anemia Level as it is another of the few ways to predict a non-binary categorical variable. Decision trees are a form of machine learning model that resembles a tree with different branches, where each branch is a decision made based on predictor variables. Many decision tree models with varying predictors were made to find one with the highest predictive accuracy.

A portion of the finalized decision tree model is visualized here:



The accuracy was then calculated by creating a confusion matrix of predicted response and actual response using testing data:



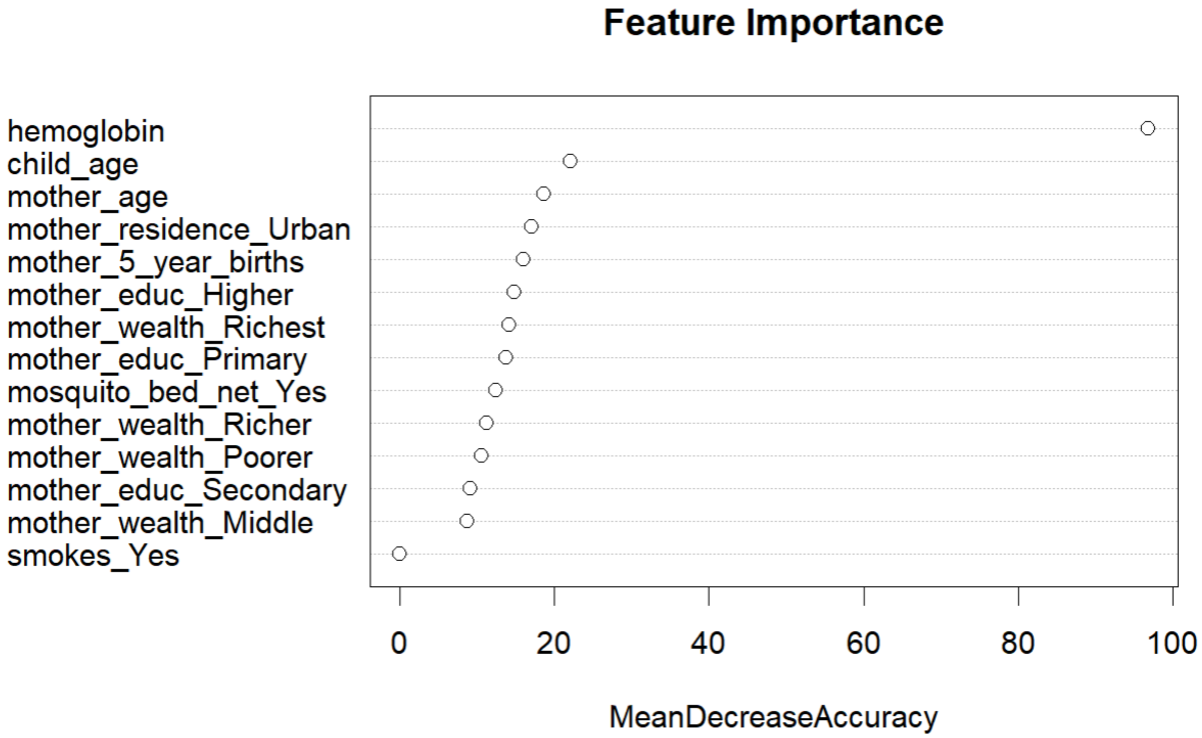
```
             y_pred
              Mild Moderate Not anemic Severe
  Mild         246       34          0      0
  Moderate       0      271          0      0
  Not anemic    22        0        375      0
  Severe         0       10          0      0
```

The accuracy for the decision tree model was calculated to be 93.11%.

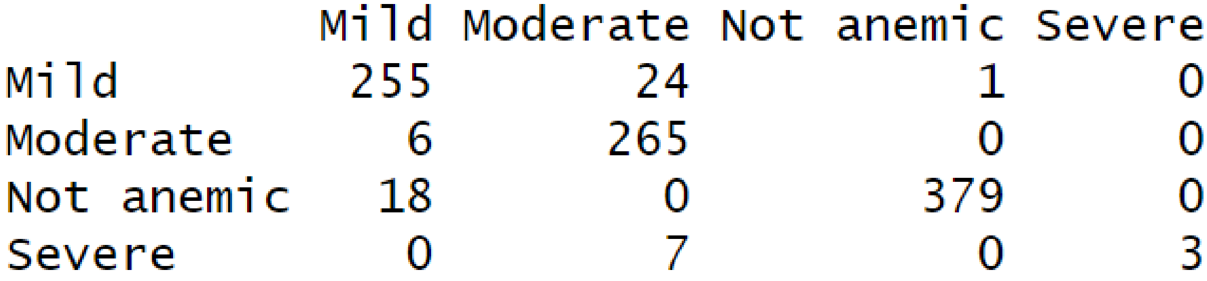
Random Forest Model

Random forests are non parametric models that average a set of uncorrelated decision trees to create an outcome with lower variability than the individual decision trees. Random forests control for overfitting as they use bootstrap aggregation, which creates simulated samples out of a single training dataset. Many random forest models with varying predictors and number of decision trees were made to find one with the highest predictive accuracy.

Feature importance was then calculated which shows the mean decrease in accuracy when removing each individual predictor. This shows the importance of each predictor.



The accuracy was then calculated by creating a confusion matrix of predicted response and actual response using testing data:



```
             Mild Moderate Not anemic Severe
Mild          255       24          1      0
Moderate        6      265          0      0
Not anemic     18        0        379      0
Severe          0        7          0      3
```

The accuracy for the random forest model was calculated to be 94.15%.

Conclusion

A multinomial logistic model, a neural network model, a decision tree model, and a random forest model were created in order to predict Anemia Level from a variety of predictor models. Comparing the accuracy of these models, the multinomial logistic model had an accuracy of 86.64%, the neural network model had an accuracy of 91.75%, the decision tree model had an accuracy of 93.11%, and the random forest model had an accuracy of 94.15%. All of these models performed very well at predicting anemia level, as this prediction response had four different response values. The random forest model had a higher accuracy so it could be considered the most strong model. However, the multinomial logistic model has the benefit of being much more interpretable. This model is created through regression techniques, whereas the random forest model is created through a highly complex random forest of decision trees. One negative of using random forests is that they are less explainable.

Overall, all models had fairly high accuracy and their own set of strengths and weaknesses in terms of interpretability and complexity.

Next Steps

Overall, all models had fairly strong predictive accuracy. However, there is still much room for improvement. Different types of neural networks such as convolutional neural networks and modular neural networks. In addition, datasets with more predictive variables could be used with great effect.

Sources

Adeola Adesina. (2023). <i>Factors Affecting Children Anemia Level</i> [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/6801499

Awan, A. A. (2023, February 6). *Building Neural Network (NN) models in R*. DataCamp. https://www.datacamp.com/tutorial/neural-network-models-r

Datasciencebeginners. (2020, May 27). *Multinomial logistic regression with R: R-bloggers*. R. https://www.r-bloggers.com/2020/05/multinomial-logistic-regression-with-r/

Decision tree. CORP-MIDS1 (MDS). (2023, December 14). https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

Disci, S. (2021, July 1). *Feature importance in random forest: R-bloggers*. R. https://www.r-bloggers.com/2021/07/feature-importance-in-random-forest/

Finnstats. (2021b, April 19). *Decision trees in R: R-bloggers*. R. https://www.r-bloggers.com/2021/04/decision-trees-in-r/

Finnstats. (2021, April 13). *Random Forest in R: R-bloggers*. R. https://www.r-bloggers.com/2021/04/random-forest-in-r/

Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Version 1.7.1. URL: https://github.com/jacobkap/fastDummies, https://jacobkap.github.io/fastDummies/.