Homework.1

Eleonora Giuliani 247161

2024-03-29

Breastfeeding intentions among pregnant mothers

The scientific community recognizes the advantages of breastfeeding, as it provides numerous heath benefits for newborns. Nevertheless, there is a significant variability in breastfeeding choices across populations. Nowdays, mothers can consider various alternative feeding practices, including bottle feeding. Understanding the reasons behind these decisions and the factors that influence them is crucial for developing targeted interventions to promote breastfeeding.

To illuminate the determinants influencing breastfeeding decisions, a scientific study was conducted in a UK hospital. A cohort of 135 expectant mothers was surveyed regarding their preferences for their forthcoming baby. Additional information about the mothers was examined: advancement of the pregnancy (pregnancy), how the mothers were fed as babies (howfed), how the mother's friend fed their babies (howfedfr), if they have a partner (partner), their age (age), the age at which they left full-time education (educat), their ethnic group (ethnic) and if they have ever smoked (smokebf) or if they have stopped smoking (smokenow).

These factors are analyzed in this study to discern the specific determinants impacting maternal decisions.

Description of the dataset

As previously mentioned, the variables included in the dataset are 10, specifically: breast, pregnancy, howfed, howfedfr, age, educat, smokebfand, smokenow. Before properly starting the analysis, we shall briefly examine the present dataset, beginning verifying the correctness of the opening. It is evident that the majority of variables are categorical, with the exception of age and education. Specifically, the response variable "breast" exhibits two categories: "bottle" and "breast". The first category includes the cases "breastfeeding", "try to breastfeed" and "mixed breast- and bottle-feeding", while the second category corresponds to "exclusive bottle-feeding".

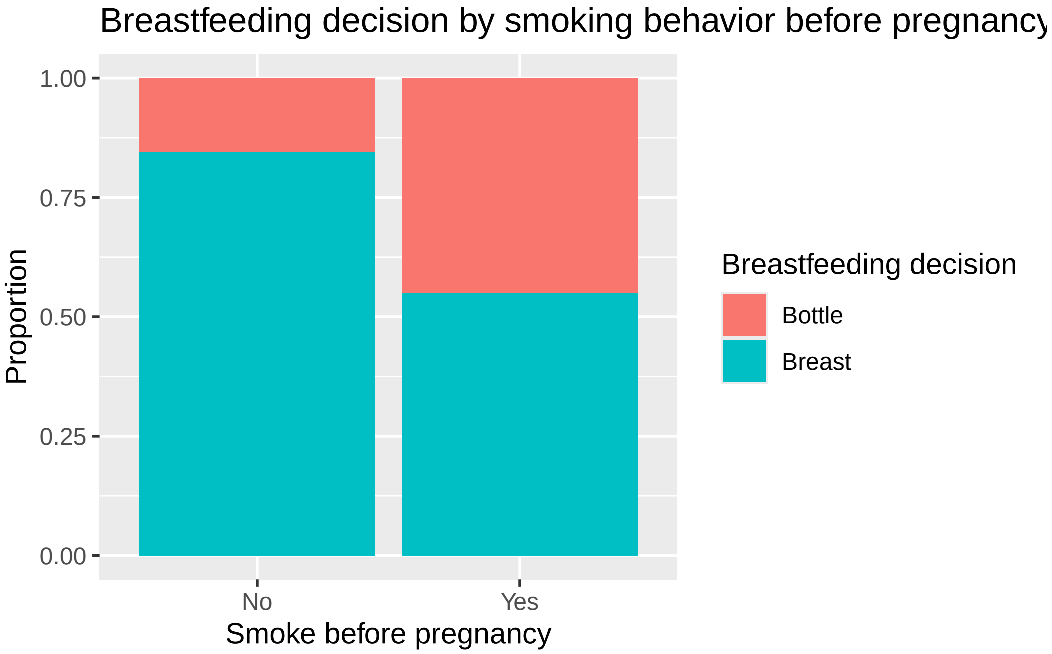
```
  breast pregnancy howfed howfedfr partner smokenow smokebf age educat
1 Breast Beginning Breast   Breast Partner       No      No  24     19
2 Breast Beginning Bottle   Breast Partner       No      No  27     18
3 Bottle Beginning Breast   Breast Partner       No      No  39     16
4 Bottle Beginning Breast   Breast Partner      Yes     Yes  29     16
5 Breast Beginning Breast   Breast Partner       No      No  21     21
6 Bottle Beginning Breast   Bottle Partner       No      No  NA     28
     ethnic
1 Non-white
2     White
3     White
4     White
5     White
6     White
```

```
    breast           pregnancy            howfed            howfedfr        
 Length:139         Length:139         Length:139         Length:139        
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                            
                                                                            
                                                                            
                                                                            
   partner            smokenow           smokebf                age       
 Length:139         Length:139         Length:139         Min.   :17.00  
 Class :character   Class :character   Class :character   1st Qu.:25.00  
 Mode  :character   Mode  :character   Mode  :character   Median :28.00  
                                                          Mean   :28.26  
                                                          3rd Qu.:32.00  
                                                          Max.   :40.00  
                                                          NA's   :2      
     educat         ethnic         
 Min.   :14.00   Length:139        
 1st Qu.:16.00   Class :character  
 Median :17.00   Mode  :character  
 Mean   :18.15                     
 3rd Qu.:19.00                     
 Max.   :38.00                     
 NA's   :2                         
```

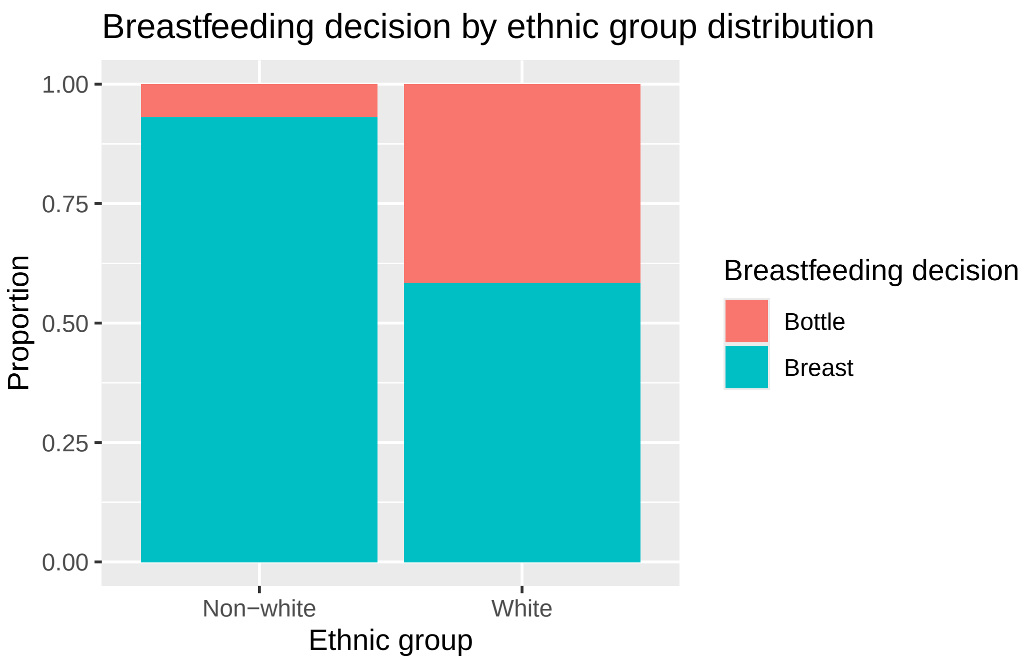
Upon inspection of the summary output, it's evident that both "age" and "educat" variables show two NA values each, meaning the presence of two missing values. Since the number of observations not available is relatively small (total NA's = 4), it is reasonable to consider removing rows corresponding to these values.

To access the discriminative power of predictors, we generated plots using the ggplot2 library. These plots allowed us to compare the variability of the response variable "breast" based on some predictor variables. It is evident that, among all the predictors, three specific variables play a significant role: whether the mother has stopped smoking, whether she currently smokes and the ethnicity.

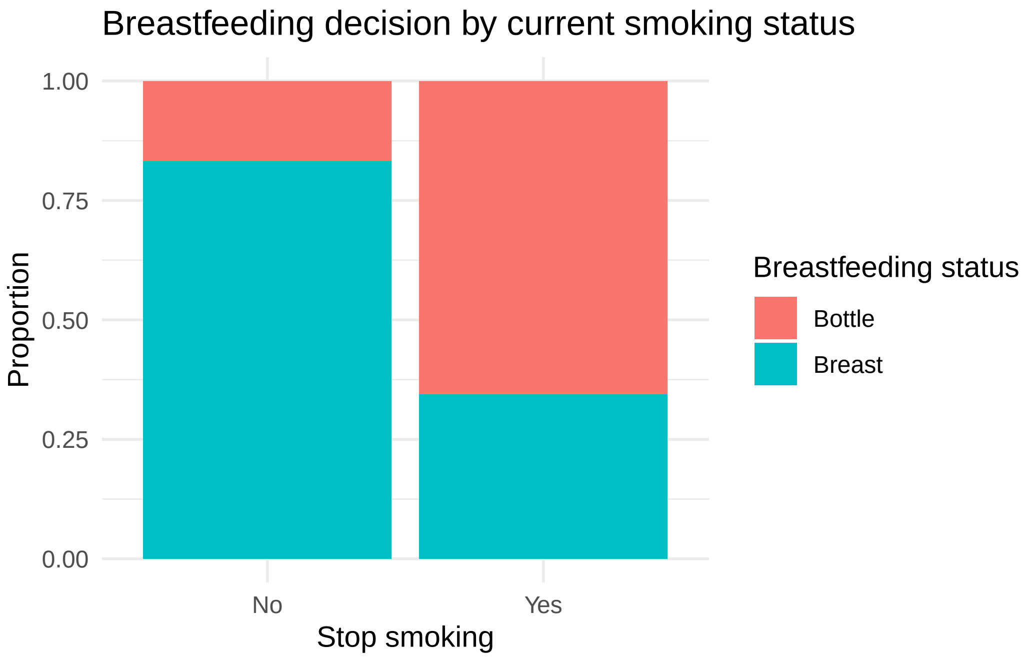
From the graph below, it is evident that mothers who never smoked are more likely to choose breastfeeding, with a proportion of over 75% compared to mothers who smoked before pregnancy, with a proportion of about 50%.



As anticipated, ethnic background has a significant impact. Analyzing the ggplot, it is evident that non-white mothers tend to choose breastfeeding over bottle feeding by more than 90%, whereas white mothers consider breastfeeding only around 50% of the time.



The last plot, instead, highlights the role of current smoking status. More than 80% of mothers who declare they haven't stopped smoking opt for breastfeeding, while only about 20% opt for bottle feeding. Differently, among mothers who have stopped smoking, only about 30% prefer breastfeeding, while the majority prefer bottle feeding.



Before dividing data into training and test sets, it's important to transform categorical variables into dummy variables. This process involves converting each categorical variable into a binary variables(0 and 1). In this way, the predictive model can understand and utilize these variables.

To verify the correctness of the step:

```
head(data)
```

```
  breast pregnancy howfed howfedfr partner smokenow smokebf age educat ethnic
1 Breast         1      1        1       1        1       1  24     19      1
2 Breast         1      0        1       1        1       1  27     18      0
3 Bottle         1      1        1       1        1       1  39     16      0
4 Bottle         1      1        1       1        0       0  29     16      0
5 Breast         1      1        1       1        1       1  21     21      0
7 Breast         1      1        1       1        1       1  27     19      1
```

We proceeded, with the splitting of the dataset. This step is crucial because it allows us to assess how well the model predicts actual data. In fact, by splitting the dataset into testing and training sets, we can train the model on training subset(70% of he data) and evaluete its performance on the testing subset (30% of the data). To ensure reproducibility, we set the seed to a specific value, in particular 98 .

```
library(caret)
# set the seed
set.seed(98)
train_size <- 0.7
train_index <- caret::createDataPartition(data$breast, p = 0.7, list = FALSE)
# select train and test
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

Before fitting the data into any model, it is necessary to transform the response variable "breast" into a binary numerical variable. In this transformation, the value "1" represents the breastfeeding category, while "0" represents the bottle feeding category.

After that, we fit a generalized linear model using the train data. The response variable is represented by "breast" and all the other variables serve as predictors. We utilized the binomial family for the process.

```
glm.fits <- glm(breast ~ pregnancy + howfed + howfedfr + partner + age + educat +
    ethnic + smokenow + smokebf, data = train_data, family = binomial)
```

Analyzing the summary, specifically the statistical significance of the coefficients, we can deduce that, while “pregnancy”, “howfed”, “partner”, “age” and “educat” are not statistically significant (as indicated by their large p-values), “howfedfr”, “ethnic”, “smokenow” and “smokebf” have a significant effect on the likelihood of the “breast” variable. Additionally, “smokenow” and the intercept term (“breast”) appear to have the most significant impact on the model’s results.

```
Call:
glm(formula = breast ~ pregnancy + howfed + howfedfr + partner +
    age + educat + ethnic + smokenow + smokebf, family = binomial,
    data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.36999    2.93933  -2.848  0.00441 **
pregnancy   -1.27938    0.79392  -1.611  0.10708
howfed       0.66444    0.76079   0.873  0.38247
howfedfr     1.30990    0.77478   1.691  0.09090 .
partner      0.76738    0.90619   0.847  0.39710
age          0.15877    0.08591   1.848  0.06459 .
educat       0.04128    0.14241   0.290  0.77192
ethnic       2.49541    0.99648   2.504  0.01227 *
smokenow     5.75744    1.75387   3.283  0.00103 **
smokebf     -3.08384    1.54270  -1.999  0.04561 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.144  on 95  degrees of freedom
Residual deviance:  55.143  on 86  degrees of freedom
AIC: 75.143

Number of Fisher Scoring iterations: 6
```

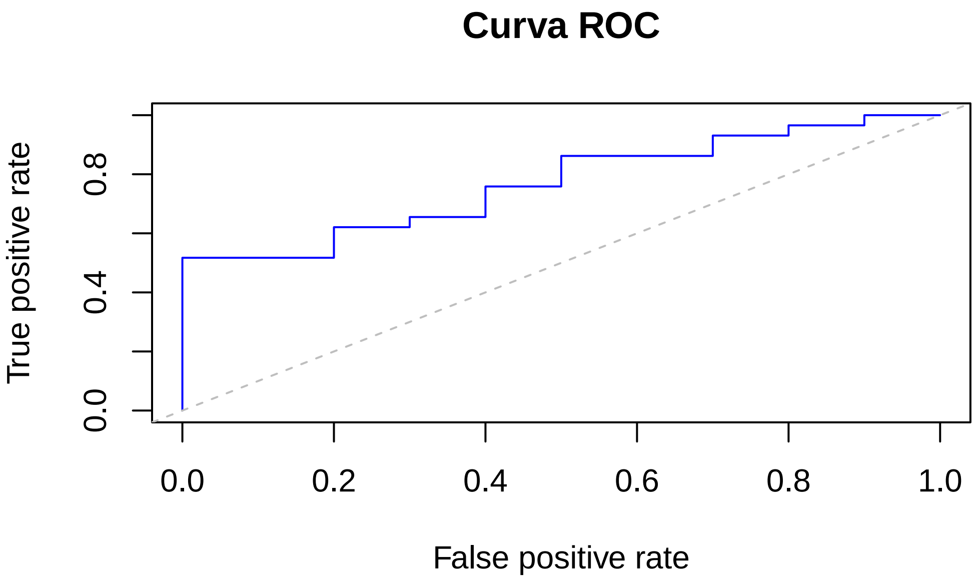
We proceed with the evaluation of the glm model by computing the accuracy of the predictions made by the model on test data.

The accuracy is approximately 66.67% and in this case it might be considered reasonable. It performs better than random guessing and so it captures some pattern in the data.

```
glm.pred  0  1
  bottle  6  9
  breast  4 20
```

```
[1] 0.6666667
```

To further analyze the model, we utilized the ROC curve (Receiver Operating Characteristic curve) and computed the AUC (Area Under the Curve). The resulting AUC value shows that the model has a good predictive capability, as it exceeds 0.5.



```
[1] 0.7689655
```

After separating the predictor variables from the target variable, we proceeded to fit a k-nearest neighbors classification model(KNN), using the class library. We experimented with various values of k, ultimately selecting a value of 9. Following this, We built the model using the train data.

```
k <- 9
knn_model <- knn(train_knn, test_knn, train_data$breast, k = k)
```

The accuracy of the KNN model, computed using the confusion matrix, reveals that approximately 79.49% of the predictions were correct out of the total.

```
knn_model  0  1
        0  4  2
        1  6 27
```

```
[1] 0.7948718
```

As for the previous model, we computed the AUC value of the ROC curve.

```
[1] 0.6655172
```

Conclusion

Analyzing the results obtained from the two predictive models and observing the ggplots, we can draw several conclusions. From the ggplots, confirmed then by the glm model, we identified significant variables that impact the choice of breastfeeding over bottle-feeding. Particularly, the ethnicity of the mother and her smoking habits impact the decision. Subsequently, a comparison between the performance of the two predictive models, GLM and k-NN, was conducted. In order to evaluate the best model, we used accuracy and AUC values. Considering accuracy, the k-NN model performs better, achieving a higher score (79.49% compared to 66.67%). However, based on the AUC value, the GLM model seems to perform better exhibiting n higher score(76.90% compared to 66.55%). Deciding between the two models depends on the trade-off between discriminative power and accuracy.

These findings aid in understanding the determinants that influence the decision process regarding breastfeeding.