

Homework 01

Damiano Orlandi

2024-03-25

Introduction

This study investigates factors affecting the decision of pregnant women to breastfeed their children at a UK hospital, using a dataset of 135 expectant mothers. The analysis aims to identify significant predictors of breastfeeding intention and evaluate the effectiveness of logistic regression and k-nearest neighbors (K-NN) classification in predicting breastfeeding choices.

Methods

Data Description

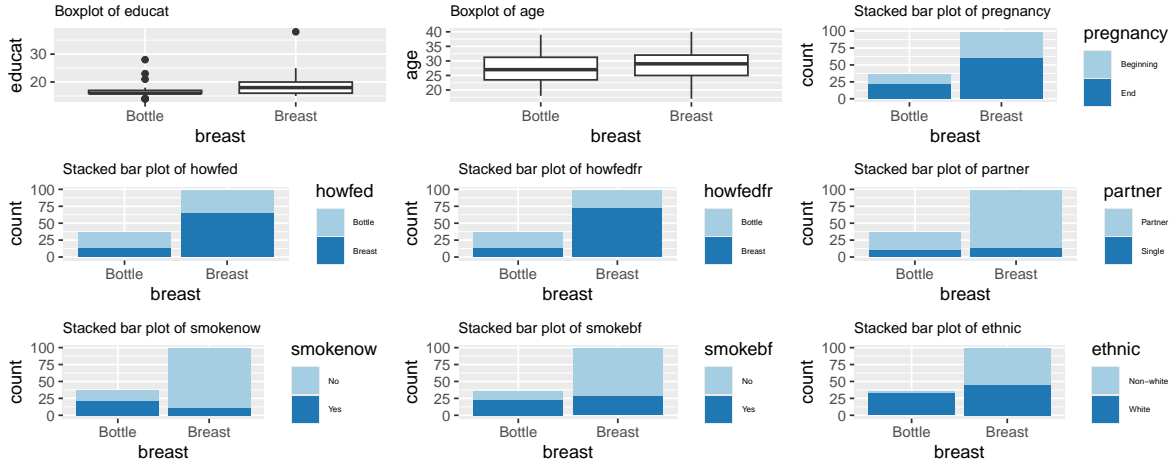
The data set comprises responses from 135 expectant mothers on their intended feeding method, categorized into breastfeeding (including “try to breastfeed” and “mixed breast-and bottle-feeding”) and exclusive bottle-feeding. Predictors include pregnancy advancement(**pregnancy**), maternal (**howfed**) and maternal friends’ infant feeding methods (**howfedfr**), partnership status (**partner**), age (**age**), when they left full-time education (**educat**), ethnic group (**ethnic**), and smoking behavior (**smokebf**: if they ever smoked before, **smokenow**: if they are smokers). All the columns contain categorical values except for “**educat**” and “**age**”, which are of type integer.

breast	pregnancy	howfed	howfedfr	partner	smokenow
"character"	"character"	"character"	"character"	"character"	"character"
smokebf	age	educat	ethnic		
"character"	"integer"	"integer"	"character"		

Data Exploration

The data set initially contained 139 entries with 10 attributes, with missing values identified in the “**age**” and “**educat**” variables, leading to the exclusion of affected rows. Given the minor proportion of missing data, this exclusion minimally impacts the sample size, avoiding potential

bias from imputation methods due to non-random missingness (using mean or median values). Additionally, character-type columns were recast as factors for logistic regression and later as numeric for KNN analysis. To assess each predictor's discriminatory capability, various plots were generated. Box plots of maternal age and education level indicated potential influences on feeding choices, while stacked bar plots evaluated other factors, highlighting ethnicity, current smoking status, and friends' feeding methods as significant.



Data Partition

The dataset was divided into training and test set, using the suggested function `caret::createDataPartition` in order to maintain the class imbalance. A seed was set for reproducibility, and 80% of the data was allocated for training, with the remainder for testing.

```
set.seed(18)

indexTrain <- caret::createDataPartition(y = data$breast, p = 0.8, list = FALSE)
df_train <- data[indexTrain,]
df_test <- data[-indexTrain,]
```

GLM

Analysis through GLM revealed correlations between predictors and feeding choices: “**howfedfrBreast**” (p-value of 0.00640), “**smokenowYes**” (p-value of 0.00511), “**ethnicWhite**” indicating significant associations at high confidence levels (99% for howfedfr-Breast and smokenowYes, while 95% for ethnicWhite). Other factors showed insignificant statistical relevance, suggesting limited influence on feeding decisions.

```
Call:
glm(formula = breast ~ ., family = binomial, data = df_train)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.65276	2.76438	-0.236	0.81333
pregnancyEnd	0.76285	0.62756	1.216	0.22414
howfedBreast	0.22736	0.65858	0.345	0.72992
howfedfrBreast	1.69979	0.62342	2.727	0.00640 **
partnerSingle	-0.95902	0.77583	-1.236	0.21641
smokenowYes	-2.92357	1.04404	-2.800	0.00511 **
smokebfYes	1.46618	1.02871	1.425	0.15408
age	0.03021	0.05525	0.547	0.58450
educat	0.09093	0.12577	0.723	0.46970
ethnicWhite	-2.38160	0.98588	-2.416	0.01570 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 126.286 on 108 degrees of freedom
Residual deviance: 77.729 on 99 degrees of freedom
AIC: 97.729

Number of Fisher Scoring iterations: 6

12	18	23	27	29	33	35
0.94417982	0.95229794	0.95197876	0.72455249	0.90924951	0.98690950	0.09012260
50	52	56	61	67	68	75
0.18413601	0.98962717	0.83731603	0.74841432	0.98954080	0.93141095	0.98589958
95	96	98	99	107	112	116
0.83648305	0.99460119	0.09265529	0.37251899	0.76733128	0.90499233	0.98664280
127	129	131	132	134		
0.99014519	0.49990243	0.97271160	0.69374477	0.93235785		

K-NN

The K-nearest neighbors (KNN) method is a simple yet effective algorithm used for both classification and regression tasks. It predicts the class or value of a new data point based on the majority class or average value of its 'k' nearest neighbors in the training dataset. Since KNN's performance heavily depends on the choice of 'k' and the distance metric used for calculating similarity, I ranged the values from 1 to 20 to have a better overview on the error rate. A utility function facilitated the assessment of the error rate for each tested 'k' value, enabling the measurement of accuracy for each selected 'k'. Subsequently, the error rate was plotted against the chosen 'k' values, allowing for graphical observation of the 'k' value with the lowest error rate.

In determining the appropriate ‘k’ value, consideration was given to the bias-variance trade-off. Given the modest size of the dataset, emphasis was placed on prioritizing a model with higher bias to facilitate accurate predictions on unseen data.

k=17 was selected for the final KNN model to minimize error rate and prevent overfitting.

```
set.seed(18)

cols_to_convert2 <- c("breast", "pregnancy", "howfed", "howfedfr",
                     "partner", "smokenow", "smokebf", "ethnic")
data[cols_to_convert2] <- lapply(data[cols_to_convert2], as.integer)

calc_error_rate <- function(predicted.value, true.value) {
  mean(true.value != predicted.value)
}

errors_tr <- errors_ts <- c()
k_vec <- c(seq(1:20), 30, 40, 50)
for (k in k_vec) {

  pred_tr <- knn(x_train, x_train, y_train, k = k)
  pred_ts <- knn(x_train, x_test, y_train, k = k)
  err_tr <- calc_error_rate(pred_tr, y_train)
  err_ts <- calc_error_rate(pred_ts, y_test)
  errors_tr <- append(errors_tr, err_tr)
  errors_ts <- append(errors_ts, err_ts)
}

plot(1, type="n", xlim = c(1,70), ylim= c(0, 0.6), log = "x", xlab = "K",
     ylab = "Error Rate")
lines(k_vec, errors_tr, type = "b", col="blue")
lines(k_vec, errors_ts, type = "b", col = "black")
legend("topright", legend = c("Train error", "Test error"), col = c("blue", "black"),
     pch = 19)
```

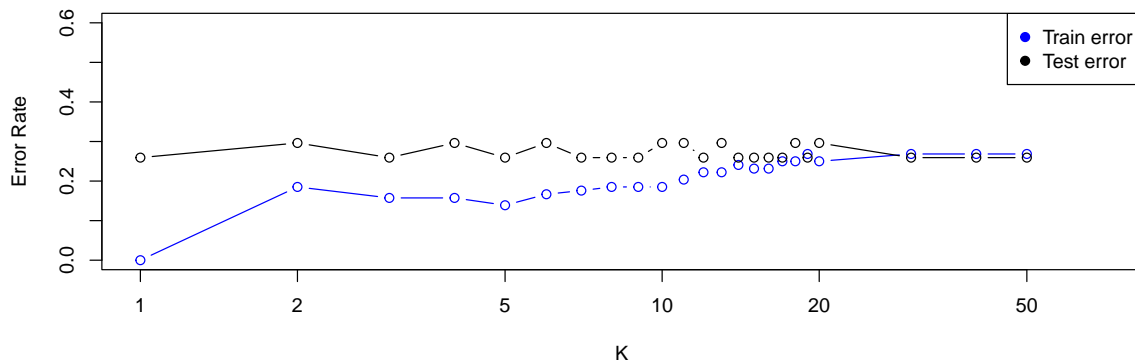


Figure 1: Error rate vs. K.

```
knn1 <- knn(x_train, x_test, y_train, k=17)
knn_conf_mat <- table(knn1, y_test)
knn_conf_mat
```

```
      y_test
knn1  1  2
     1  0  0
     2  7 20
```

Performance Evaluation of GLM and KNN

I considered the confusion matrix the more suitable tool to identify the best statistical model since it gives us information about precision and recall. The first one calculates the proportion of correctly predicted positive instances among all instances predicted while recall calculates the proportion of correctly predicted positive instances among all actual positive instances. We can notice that the GLM model outperforms the KNN model in terms of “Recall.” This suggests that the GLM model exhibits greater accuracy in modeling the “Bottle” instances. Conversely, the KNN model demonstrates slightly higher precision, potentially attributed to its inability to identify true positive instances as effectively as the GLM model.

```
      y_test_logit
glm_pre Bottle Breast
     0      4      1
     1      3     18
```

```
      y_test
knn1  1  2
```

1 0 0
2 7 20

Table 1: Summary of Model Performance

Model	Precision	Recall
GLM	0.9473684	0.8571429
KNN	1.0000000	0.7407407

Overall Conclusions and Limitations

The study's limitations include its small dataset size and class imbalance, potentially affecting generalizability and model performance. Despite these challenges, GLM demonstrated greater effectiveness in classifying feeding preferences among expectant mothers, making it the recommended approach for predictive analysis in this context.