NATIONAL UNIVERSITY OF ENGINEERING

COLLEGE OF ECONOMICS AND STATISTICAL ENGINEERING AND SOCIAL SCIENCES

STATISTICAL ENGINEERING PROGRAM

EF012 PROJECT WORKSHOP

Design of a Statistical Processing System for the Prediction of Children Anemia Up to 5 Years Old Using a Logistics Regression Model.

by:

Mercedes M. CONGACHA FERNANDEZ

Professor

Ivan Victor SILVA GUILLEN

LIMA – PERU

2021

Abstract

The present work develops the application of the binary logistic regression model in order to identify characteristics and determine the predictive power of children with prevalence of anemia between the ages of 0 to 59 months in Peru. To identify the predictive power, precision, recall and AUC metrics were used; On the other hand, to determine the fit of the model and the significant variables in the model, the Wald test and the model fit tests were used. As a result, a precision of 68 %, a recall of 71 % and an AUC of 74.3 % were obtained.

Keywords – Anemia, Binary logistic regression, AUC, accuracy, recall

Resumen –

El presente trabajo desarrolla la aplicación del Modelo de regresión logística binaria con el fin de identificar factores y determinar el poder predictivo de niños que presentan prevalencia de anemia entre las edades de 0 a 59 meses en el Perú. Para la identificación del poder predictivo se empleó las métricas de exactitud, sensibilidad y AUC; por otro lado, para determinar el ajuste del modelo y las variables significativas en el modelo se empleó Prueba de Wald y pruebas de ajuste del modelo. Como resultado se obtuvo una medida de exactitud de 68 %, sensibilidad de 71 % y AUC de 74,3 % en la data balanceada. Y los factores más influyentes fue si el niño vive en Puno, si la madre del niño no presenta educación y el niño es del sexo masculino, hay una gran probabilidad de que el niño presente anemia.

Palabras clave – Anemia, Regresión logística binaria, AUC, sensibilidad.

General Index

Chapter 1

Problem Statement

1.1. Description of the problem situation

The presence of anemia raises much concern in all areas and levels of health, since its consequences have a negative impact on the development of girls and boys at a cognitive, motor, emotional and social level.

At the national level, during 2019 the highest levels of anemia in girls and boys from 6 to 35 months of age were registered in the Sierra (48.8%), followed by La Selva (44.6%), Rest of the Coast (37.5%) and Metropolitan Lima (30.4%) 1. For the 2019 I semester, anemia in children under 5 years of age was 35.6

1National Institute of Statistics and Informatics

The present work aims to provide a statistical model with predictive results, which will be calculated according to certain important variables of the profile, focusing on the high probabilities of suffering from anemia in children.

1.2. Problem Statement

1.3.1. General objective

Is it possible to identify factors that determine the behavior of anemia in children from 0 to 59 months of age in Peru, as well as determine if the Binary Logistic Regression model can identify positive predicted values (children with prevalence of anemia)?

1.3.2. Specific objectives

- What factors affect the prevalence of anemia between the ages of 0 to 59 months of age?
- Will it be possible to identify the intensity of the factors that affect the prevalence of anemia in children from 0 to 59 months of age?
- Which of the departments has the highest prevalence of anemia in children from 0 to 59 months of age?
- Does the rolling binary logistic regression model show a major predictive power of no rolling in the case of anemia in 59-month-old children?

1.3. Research objectives

1.3.1. General objective

Identify factors and determine the predictive power of children with prevalence of anemia between the ages of 0 to 59 months of age in Peru using a Binary Logistic Regression model.

1.3.2. Specific objectives

- Determine what factors affect the prevalence of anemia between the ages of 0 to 59 months of age.

- Determine the intensity of the factors that affect the prevalence of anemia in children from 0 to 59 months of age.
- Determine which of the departments has the highest prevalence of anemia in children from 0 to 59 months of age.
- Compare the predictive capacity of the binary logistic regression model without and with balancing.

The verification of its fulfillment of the specific objectives will be carried out as follows:

- In the first objective, we worked with the Wald test to identify which factors are significant and therefore explain the prevalence of anemia in children from 0 to 59 months of age.
- To determine the intensity or weight of the factor, the odd ratio measure was used.
- This would be determined by the odd ratio.
- To determine the predictive power of the model with or without balancing, the statistical tool AUC, accuracy and sensitivity will be used.

1.4. Research justification and limitations

1.4.1. Justification

It provides theoretical aspects of predictive models in the health area where there was really little knowledge in predicting and observing variables that influence cases of anemia in children. In the social aspect, the present work can indicate certain behavior of cases of anemia in children, having a follow-up could give a better quality of health.

1.4.2. Limitation

- The lack of studies of predictive models for the health area
- That some respondents do not have the same patience to fill out the questionnaire

1.5. Research background

Next, some descriptions are presented that look for characteristics that are associated with anemia levels with the application of state-of-the-art prediction models.

In 2019, San Eusebio Condorio, proposed to implement a data mining model to predict cases of anemia in pregnant women in the province of Ilo. The models that I implemented were the algorithms of Multilayer Perceptron, Naive Bayes and Decision Tree J48, these were trained on a historical database of 422 records of pregnant mothers with anemia from the Province of Ilo, the algorithm that reached the highest precision was the of Naive Bayes with 89%, followed by the decision tree J48 with 79% and finally the multilayer perceptron with 62%. The development of the project was based on the CRISP-DM methodology to develop each of the stages that led to the final result.

Another study carried out in sub-Saharan Africa, studies of analysis of risk factors for anemia in the SSA were carried out by using multivariate regression models where individual, maternal and domestic risk factors were considered, the hemoglobin test was also considered. As a result, it was obtained that demographic and socioeconomic factors, family structure, water, sanitation, growth, maternal health, and recent illnesses were significantly associated with the presence of childhood anemia. These groups of risk factors explain a significant fraction of anemia (ranging between 1.0% and 16.7%) at the population level.

Chapter 2

Conceptual Theoretical Framework

2.1. Conceptual Bases

2.1.1. Anemia – Definition

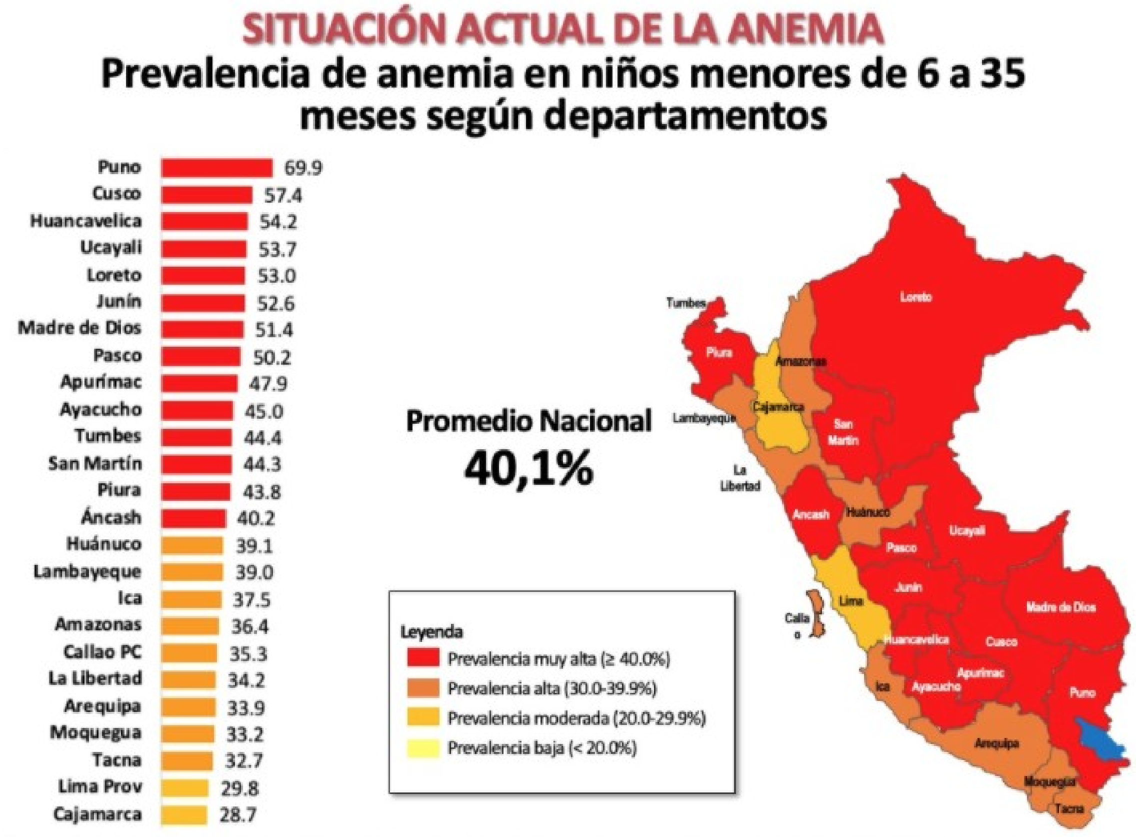
Anemia is a disorder in which the number of red blood cells (and therefore the oxygen-carrying capacity of the blood) is insufficient to meet the body's needs. Specific physiological needs vary based on age, gender, altitude above sea level at which the person lives, smoking, and different stages of pregnancy. Iron deficiency is thought to be the most common cause of anemia collectively, but can be caused by other nutritional deficiencies (including folate, vitamin B12, and vitamin A), acute and chronic inflammation, parasitic diseases, and Hereditary or acquired diseases that affect the synthesis of hemoglobin and the production or survival of red blood cells. The hemoglobin level alone cannot be used to diagnose iron deficiency (also called iron deficiency). However, it must be measured, even though not all anemias are caused by iron deficiency. The prevalence of anemia is an important health indicator and, when used with other determinations of nutritional status with respect to iron, the hemoglobin concentration can provide information on the severity of iron deficiency (1).

Population	No anemia	Anemia		
		Mild	Moderate	Serious
Children 6 to 59 months of age	110 or higher	100-109	70-99	less than 70

Figura 2.1: Hemoglobin concentrations to diagnose anemia at sea level (g/l)±

2.1.2. Current situation of anemia in Peru

Currently in Peru, 40.1% of children, from 6 to 35 months, suffer from anemia; in other words, we are talking about almost 700 thousand anemic children under 3 years of age out of 1.6 million nationwide. This alarming situation has made the current government aim to reduce to 19% by 2021, through the National Plan to combat anemia (MS).



Fuente: Instituto Nacional de Estadística e Informática (2020). Encuesta Demográfica y de Salud Familiar (ENDES) 2019.

Figure 2.2: Current situation of anemia cases in children in Peru

2.2. Theoretical Bases

2.2.1. Binary Logistic Regression Model

This model is an extension of the linear regression model. The logistic regression model, the dependent variable Y is categorical, where it seeks to determine the probability that any of the categories occurs in terms of a group of predictor variables X1, X2, ..., Xk, which can be categorical or numerical. Consider a binary regression model. The response variable yi will be assumed to be a Bernoulli random variable, whose probability distribution is as follows

$$Y_i \sim Ber(p_i)$$

$$p_i = F(x_i^T\beta) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

where $x_i = (1, x_{i1}, x_{i2}, ..., x_{ik})^T$ is a vector with the values of the k explanatory variables Y_i a binary variable such that $Y_i = 1$ occurs with probability p_i, $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ a vector of k regression coefficients. $F = (.)$ denotes a cumulative distribution function (fda).

When F is a cumulative distribution function of a symmetric distribution, the resulting link function is symmetric and has a symmetric form around $p_i = 0, 5$ (Bazán y Bayes, 2010).

a) Estimation of parameters in a binary logistic regression model

The estimation of the logistic regression parameters is carried out from a sample of X and Y, where the variables y_i are independent Bernoulli random variables,

$$f_i(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

and using the method of maximum likelihood, this will allow to maximize the probability; the likelihood function is given by:

$$L(y_1, y_2, ..., y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) \tag{2.1}$$

$$L = \prod_{i=1}^{n} p_i{}^{y_i}(1-p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} (1-p_i)e^{Ln(\frac{p_i}{1-p_i})^{y_i}}$$

$$L = \prod_{i=1}^{n} (1-p_i)e^{\sum_{i=1}^{n} y_i Ln(\frac{p_i}{1-p_i})}$$

The log-likelihood is obtained by calculating the logarithm of the likelihood function and is given by

$$l=\sum_{i=1}^{n}[y_i LnF(x_i^T\beta)+(1-y_i)(1-LnF(x_i^T\beta))]$$

$$l=\sum_{j=0}^{k}(\sum_{i=1}^{n}y_i x_{ij})\beta_j+\sum_{i=1}^{n}Ln(1-p_i)$$

$$l=\sum_{j=0}^{k}(\sum_{i=1}^{n}y_i x_{ij})\beta_j-\sum_{i=1}^{n}Ln(1+e^{\sum_{j=0}^{k}\beta_j x_{ij}})$$

Then the first derivative is applied to the log-likelihood function and defined as the following expression:

$$U(\beta)=\frac{dl}{d\beta_j}=\sum_{i=1}^{n}y_i x_{ij}-\sum_{i=1}^{n}x_{ij}[\frac{e^{\sum_{j=0}^{k}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{k}\beta_j x_{ij}}}]$$

It has:

$$\sum_{i=1}^{n}y_i x_{ij}-\sum_{i=1}^{n}x_{ij}\hat{p}_i=0$$

$$\sum_{i=1}^{n}(y_i x_{ij}-\hat{p}_i)=0$$

where:

$$\hat{p}_i=\frac{e^{\sum_{j=0}^{k}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{k}\beta_j x_{ij}}}$$

in matrix form:

$$x^T(y-p)=0$$

The information function:

$$H(\beta)=\frac{d^2 l}{d\beta_j d\beta_g}=\sum_{i=1}^{n}x_{ij}x_{ig}p_i(1-p_i)$$

One of the methods to find the solutions of a nonlinear equation is the Newton

Raphson method which consists of generating the sequence.

$$\{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}\}_{i=0}^{\infty}$$

from a given initial value x_0 dado. Where the function $f(x_i)$ is the function $U(\beta)$ and its derivative $f^{1}(x_i)$ which is represented by $H(\beta)$, then what we would have is

$$\{\beta_{i+1} = \beta_i - \frac{U(\beta_i)}{H(\beta_i)}\}_{i=0}^{\infty}$$

b) Interpretation of parameters of a logistic regression model

The betting ratio, odd, is the ratio between the probability that an event will happen and the probability that that event will not happen. The probability that the event does not take the value of 1 is:

$$1 - p_i = \frac{1}{1 + e^{x_i^T \beta}}$$

The betting ratio,

$$odd_i = \frac{p_i}{1 - p_i}$$

$$odd_i = e^{x_i^T \beta}$$

Indicate how the advantage ratio when observing Y = 1, changes before a unit increase in the variable X_i.

The Odd ratio for the variable X_i.

- If $\hat{\beta}_j > 0$ then $e^{\beta j} > 1$, the variable X_j increases the advantage ratio, that is, it indicates a greater probability for the value Y_1 to be successful. Or the result that Y = 1 as success, when $e^{\beta j}$ is more likely when the variable X_j increases by one.
- If $\hat{\beta}_j < 0$ then $e^{\beta j} < 1$, the variable X_j decreases or decreases the advantage ratio, that is, it indicates less probability for the value. Y_1 to be successful.

The logit is the logarithm of the Odds, using the natural logarithm of the betting ratio, a linear expression is obtained for the model:

$$ln(odd_i) = ln(\frac{p_i}{1-p_i}) = x_i^T \beta$$

$$ln(odd_i) = \beta_o + \sum_{j=1}^{k} \beta_j x_{ij}$$

c) Hypothesis testing for parameters

Significance test for the coefficients, *β*β_i, using the Wald statistic. It is a test used to individually assess whether any independent variable has a statistically significant influence on the dependent variable; through testing the hypothesis of its regression coefficient.

$$H_0 : \beta_j = 0$$
$$H_0 : \beta_j \neq 0$$

In this case the statistic is given by:

$$W = \frac{\hat{\beta}_j - \beta_j^{(0)}}{se(\hat{\beta}_j)} \quad (2.2)$$

and assuming that H_0 is true, the W statistic will have a distribution N (0, 1), according to Dunn and Smyth (2018).

d) Goodness-of-fit test of the model

The Deviance statistical test is a goodness-of-fit measure of generalized linear models

H_0: The model fits the data

H_1: The model does not fit the data

The test statistic,

D = 2ln(L(saturated model) − LnL(β^)) (2.3)

The Deviance statistic follows a Chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the saturated model (n) and the fitted model (p).

Statistical decision:

H_0 is rejected if $X^2_{\alpha,n-p} \leq D$, it is concluded that the logistic model does not fit the data.

Where:

Saturated model, it is a model that has exactly n parameters and fits perfectly to the sample data.

Observing the data flow, we can distinguish between unidirectional networks (feedforward) and recurring or feedback networks (feedback). While in unidirectional networks the information circulates in a single direction, in recurring or feedback networks the information can circulate between the different layers of neurons in any direction, even in the input-output direction.

2.3. Model metrics

Metrics provide information from the analysis model. The purpose is to find performance indicators in the analysis to predict.

2.3.1. Confusion Matrix

It is a tool that allows to visualize the performance of the model used in supervised learning. Each column of the matrix represents the number of predictions of each class, while each row represents the real classes, this means that it allows to see the types of successes and errors that the model presents after passing the learning process of the model. This is in order to obtain the efficiency of the model.

Metrics: Metrics:

- Sensitivity or recall: Percentage of positive cases detected correct

$$Sensibilidad = \frac{VP}{VP + FN}$$

Predicted values		
Values real	**No anemia (0)**	**anemia (1)**
No anemia	True negative (VN)	False positive (FP)
anemia	False negative (FN)	True positive (VP)

***Table 2.1**: Confusion matrix*

Accuracy: percentage of correct predictions

$$Accuracy = \frac{VN + VP}{Total}$$

Precision: Percentage of correct positive predictions

$$Precisión = \frac{VP}{VP + FP}$$

Specificity: Percentage of negative cases correctly detected (e.g.: the ability to identify cases of children without anemia among all children without anemia)

$$Especificidad = \frac{VN}{VN + FP}$$

2.3.2. ROC and AUC curve

It is the graphic representation of sensitivity versus specificity, for our case, take into account 0 is the category that refers to the fact that a child does not have anemia and 1 that it does. Measurement:

- AUC = 0,5 Null
- 0,7 ≤ AUC < 0,8 Acceptable
- 0,8 ≤ AUC < 0,9 Excellent
- 0,9 ≤ AUC Great

2.4. Chi Square test

H_0 : The variable X_i is independent of the variable of interest

H_1 : The variable X_i is not independent of the variable of interest

Statistical:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(observado_{ij} - esperado_{ij})^2}{esperado_{ij}}$$

where:

- $observed_{i\ j}$: is the observed frequency of the row i and column j.
- Expected: $expected_{i\ j} = \frac{marginal\ row * marginal\ column}{total}$

The null hypothesis is rejected if $\chi^2 > \chi^2$, where α is the level of significance or say that the p-value is less than 0.05.

Chapter 3

Methodology

3.1. ENDES survey

The National Demographic and Demographic Health Survey (ENDES) 2019, aims to promote updated information on demographic dynamics, the health status of mothers and children under five years of age, as well as provide information on characteristics associated with diseases non-transmissible and transmissible, as well as to provide access to diagnosis and treatment services, in order to provide information that allows estimating the indicators identified in the Budget Programs.

The ENDES survey contains 10 modules

- Home Features
- Characteristics of the House
- MEF Basic Data
- Birth History - Method Knowledge Chart
- Pregnancy, Childbirth, Puerperium and Lactation
- Immunization and Health
- Marriage - Fertility - Spouse and Wife
- AIDS awareness and condom use
- Maternal Mortality - Family Violence
- Weight and height - Anemia
- Child Discipline
- Health survey
- Social programs

The main module to study is Weight and Height - Anemia, it should be noted that other modules such as Household Characteristics and Social Programs will also be used.

3.1.1. Study population

Boys and girls under 59 months of age in Peru in the 2019 period.

3.1.2. Sample

Boys and girls under 59 months of age in the period 2019. The sample is characterized by being two-stage, it is selected from the housing register applying balanced sampling considering the variables of girls and boys under 5 years of age.

3.2. Variables Operationalization

3.2.1. Dependent variable

- Anemia level: This consists of four categories mild, moderate, severe and does not have anemia; which is categorized into two categories, has and does not have anemia.

3.2.2. Independent variable

- Birth place
- BMI
- Socioeconomic Type
- Sex
- Age in months
- Weight in kilograms
- Highest educational level of the mother
- Birth order number

Matrix of operational variables					
N	**Variable**	**Operational definition of the variable:**	**Technique and instrument of harvest**	**Dimensions**	**Measurement level**
1	HHID	Identification Individual questionnaire	character	character	
2	HC1	Age in months	Survey Questionnaire	Continuous	Of reason
3	HC2	Weight in kilograms (1 dec.)	Survey Questionnaire	Continuous	Of reason
4	HC3	Height in centimeters (1 dec.)	Survey Questionnaire	Continuous	Of reason
5	HC27	Sex	Survey Questionnaire	Man Woman	Nominal
6	HC61	Highest educational level of the mother	Survey Questionnaire	Without education Primary Secondary Higher	Nominal
7	HC64	Birth order number	Survey Questionnaire	Discreet	Of reason
8	HV024	Departments	Survey Questionnaire	25 regions	Nominal
9	HC57	Anemia level	Survey Questionnaire	Anemia No anemia	Nominal

Chapter 4

Analysis and Results

4.1. Previous knowledge

It is important to consider certain points regarding the variables and their treatment, in order to optimize results in the model. These are:

1. Two modules of the 2018 National Demographic and Demographic Health Survey (ENDES) were concatenated. The joined modules were 64, characteristics of the home, and module 74, Weight and Height - Anemia.
2. The variable of interest HC57, Level of anemia, which this variable is categorized into two categories: With anemia, 1, and Without anemia, 0.
3. It is important to know that there were variables with missing data, from which the 4-1-8 standard was eliminated (see annex). The variables with missing data were: HC57, HC61, HC64. The elimination of HC57 occurred since it is my response variable, and the other variables were eliminated from the missing data because no difference is found in the modeling analysis.

4.2. Descriptive analysis of the variables

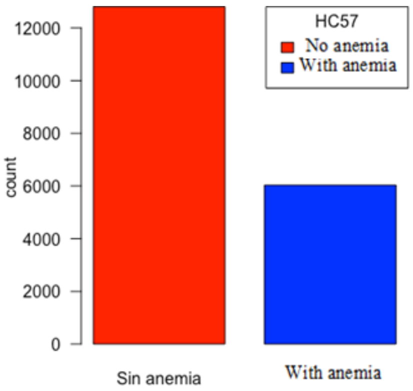


Figure 4.1*: Bar graph by case of anemia*

Figure 4.1 shows the behavior of the variable of interest, HC57, if a child from 0 to 59 months has or does not have anemia, the anemia group represents 33% of the cases. It is worth mentioning that the variable of interest is not balanced, for the analysis there will be no need to balance it since it is not a critical case (% success less than 5%).

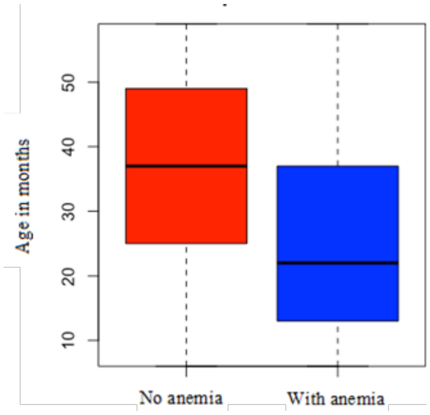


Figure 4.2: *Box diagram of ages by case of anemia*

Figure 4.2 shows that there is no presence of atypical values, we will note that there is not a great difference between the cases of anemia in the average ages, however, children with anemia have a higher mean than children who do not have anemia.

On the other hand, it is important to observe the relationship between the variable of interest and the other variables.

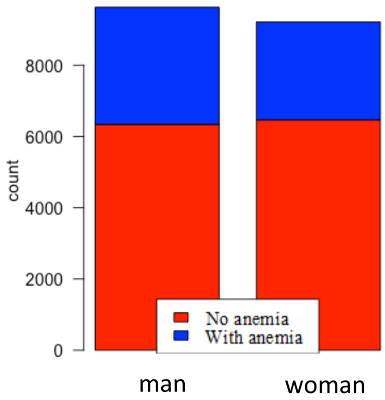


***Figure 4.3**: Bar graph of the sex variable by case of anemia*

In figure 4.3, for the case of the Sex variable, it can be observed that most cases of anemia in children from 0 to 59 months of age belong to the male sex.

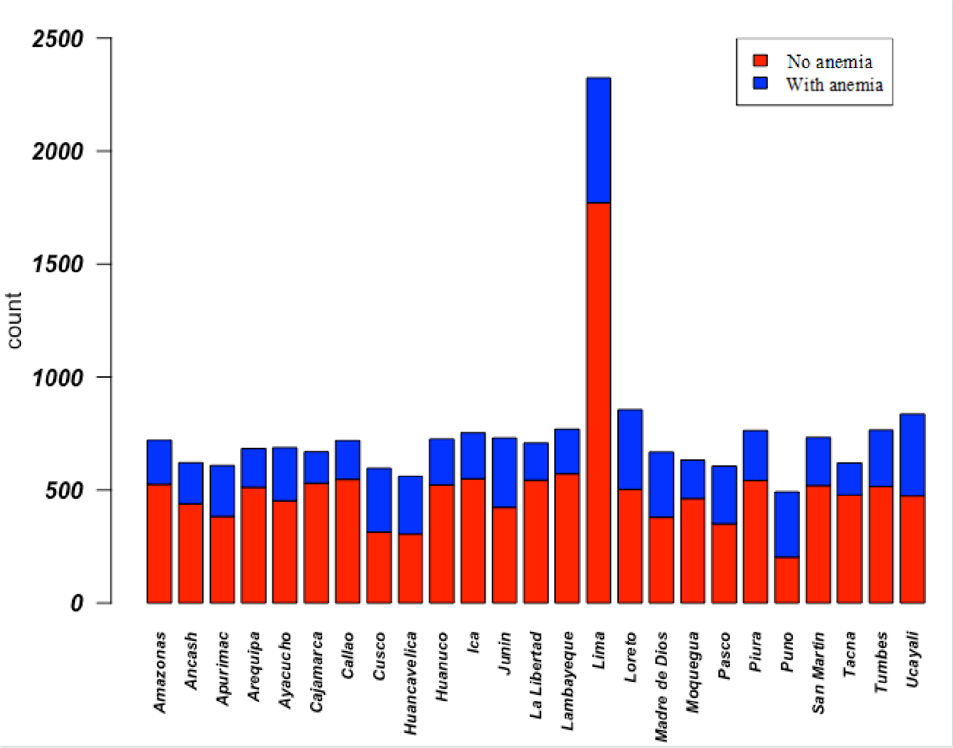


***Figure 4.4:** Bar graph of the department variable by variable of interest*

In figure 4.4, the variable of interest HC57 is analyzed with respect to the variable department, to get an idea of the behavior of cases of children with anemia by department.

The highest prevalence of anemia is observed in the departments of Puno with 69%, followed by Cusco with 57.4%, Huancavelica with 54%. on the other hand, the departments with the lowest prevalence of anemia is Cajamarca with 28.7%. Lima with 29.8%, Tacna with 32% and Moquegua with 33.2%.

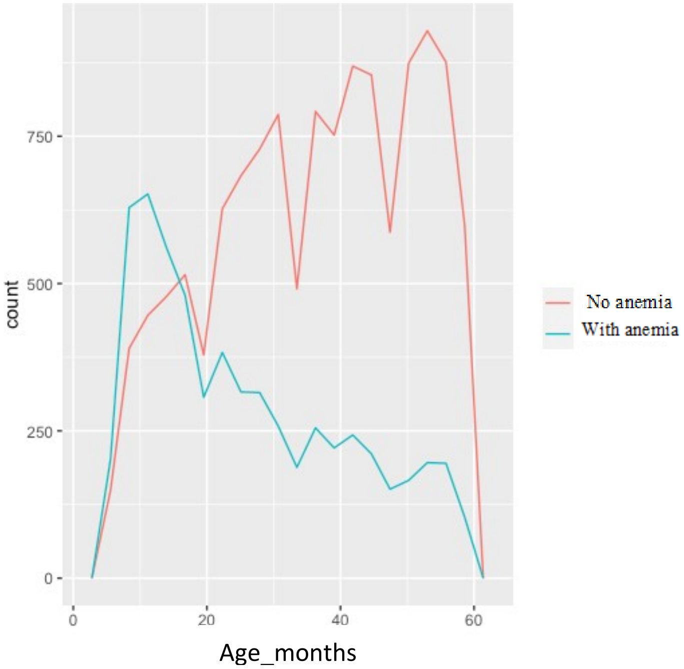


Figure 4.5: *Frequency polynomial graph*

In figure 4.5, the distribution of the variable Age in months is observed with respect to children who prevail with anemia and not. In this analysis, it is observed that children from 0 to 16 months have a greater presence of cases with anemia.

4.3. Model implementation

Before implementation, the relationship between the categorical explanatory variables and the variable of interest was tested using the Chi Square statistical test, in order to select the variables that present a relationship with the response variable. In a first step, the categorical explanatory variables were dummy.

4.3.1. Initial model

Once the descriptive analysis has been carried out, the implementation of the model for the cases of anemia in Peru in 2019 is carried out. The modeling will be carried out with all the variables already detailed in the table of variables operations. It is worth mentioning that some variables have been normalized, dummied.

Variables:

X1: Age in months (numeric)

X2: Weight in kilograms (numeric)

X3: Height in centimeters (numeric)

X4: Sex (Female: 1 and Male: 0)

X5: The educational level of the mother (Dummy)

X6: Birth order number (numeric)

X7: Department (dummy)

X8: Anemia level (with anemia: 1 and without anemia: 0)

Once the database has been prepared, we proceed to carry out the independence test of the explanatory variables, with respect to the variable of interest. It is possible to observe the variables that are dependent on the variable of interest, Level of anemia with the application of the Chi Square Test.

H_0 : The variable X_i is independent of the variable of interest.
H_1 : The variable X_i is not independent of the variable of interest.

This statistical tool will allow us to provide at first impression which variables can enter the model based on the dependence on the variable of interest.

Chi square correlation test					
N	**Variable**	**X-squared**	**df**	**p-value**	**Conclution**
1	Departamento_Amazonas	8.1683	1	0.004263	Rejection
2	Departamento_Ancash	2.139	1	0.1436	No Rejection
3	Departamento_Apurimac	7.2667	1	0.007024	Rejection
4	Departamento_Arequipa	15.379	1	0.00008797	Rejection
5	Departamento_Cajamarca	1.2665	1	0.2604	No Rejection
6	Departamento_Callao	40.067	1	0.0000000	Rejection
7	Departamento_Cusco	21.528	1	0.000003487	Rejection
8	Departamento_Huancavelica	64.954	1	0.000000000	Rejection
9	Departamento_Huanuco	48.192	1	0.000000000	Rejection
10	Departamento_Ica	5.276	1	0.02162	Rejection
11	Departamento_Junin	8.4081	1	0.003736	Rejection
12	Departamento_La Libertad	34.887	1	0.0000000	Rejection
13	Departamento_Lambayeque	25.893	1	0.0000003	Rejection
14	Departamento_Lima	14.659	1	0.0001288	Rejection
15	Departamento_Loreto	82.041	1	0.000000000	Rejection
16	Departamento_Madre de Dios	33.615,	1	0.00000000	Rejection
17	Departamento_Ica	40.374	1	0.00000000	Rejection
18	Departamento_Moquegua	6.943	1	0.008415	Rejection
19	Departamento_Pasco	29.028	1	0.00000	Rejection
20	Departamento_Piura	3.0874	1	0.0789	No Rejection
21	Departamento_Puno	166.22	1	0.00000000	Rejection
22	Departamento_San Martin	2.675	1	0.1019	No Rejection
23	Departamento_Tacna	23.87	1	0.000001031	Rejection
24	Departamento_Tumbes	0.10874	1	0.7416	No Rejection
25	Departamento_Ucayali	50.129	1	0.0000000	Rejection
26	Sexo_Hombre	40.737	1	0.00000000	Rejection
27	Sexo_Mujer	40.737	1	0.00000000	Rejection
28	nivel_edu_madre_Sin educacion	6.6283	1	0.01004	Rejection
29	nivel_edu_madre_Primaria	58.219	1	0.00000000	Rejection
30	nivel_edu_madre_Secundaria	1.3961	1	0.2374	No Rejection

It can be seen that in table 4.1 the variables that would not enter the model would be department of San Martín de Porras, department of San Martín de Tumbes, the mother's secondary educational level, department of Ancash, department of Cajamarca, these variables can be appreciate the p - value associated with the statistic are greater than 0.05, which implies that they do not reject the null hypothesis.

For the creation of the model, the data was partitioned in estimation and validation, or in which a sample was randomly drawn that represents 70% of records for the estimation and 30% for the validation.

Estimation data	
No anemia: 8942	With Anemia: 4221

Validation data	
No Anemia: 3831	With Anemia: 1809

Table 4.1*: Training and test data*

Based on the previous analysis of the Chi Square test, the Binary Logistic Regression model was implemented, where the following was obtained in the Chi Square test table, where the coefficient, the standard deviation of the error, is appreciated, the z-value and p-value, it should be noted that the z-value is the approximation of the Wald test, where from the p-value associated with the statistic that are less than 0.05 indicate that the associated parameter fits the model.

Summary of the significance of the model					
Factor	Coef	p-value	Odd	Ic 2.5	Ic 97.5
(Intercept)	5.12	0.00	167.49	4.78	5.46
‘nivel_edu_madre_Sin educacion‘	0.82	0.00	2.27	0.50	1.14
Departamento_Puno	0.80	0.00	2.23	0.55	1.05
nivel_edu_madre_Primaria	0.66	0.00	1.94	0.51	0.81
nivel_edu_madre_Secundaria	0.46	0.00	1.58	0.33	0.58
Sexo_Hombre	0.31	0.00	1.37	0.23	0.39
Altura_centi	-0.07	0.00	0.93	-0.07	-0.06
Departamento_Loreto	-0.18	0.06	0.83	-0.38	0.01
Departamento_Apurimac	-0.40	0.00	0.67	-0.63	-0.17
Departamento_Moquegua	-0.41	0.00	0.66	-0.64	-0.18
Departamento_Tumbes	-0.41	0.00	0.66	-0.62	-0.20
Departamento_Ayacucho	-0.47	0.00	0.63	-0.68	-0.25
Departamento_Ica	-0.56	0.00	0.57	-0.78	-0.34
Departamento_Ancash	-0.61	0.00	0.55	-0.84	-0.37
Departamento_Arequipa	-0.68	0.00	0.50	-0.92	-0.45
‘Departamento_San Martin‘	-0.76	0.00	0.47	-0.98	-0.54
Departamento_Tacna	-0.77	0.00	0.47	-1.02	-0.51
Departamento_Piura	-0.78	0.00	0.46	-1.00	-0.57
Departamento_Huanuco	-0.80	0.00	0.45	-1.01	-0.58
Departamento_Lambayeque	-0.80	0.00	0.45	-1.02	-0.58
Departamento_Lima	-0.84	0.00	0.43	-0.99	-0.69
Departamento_Amazonas	-0.95	0.00	0.39	-1.17	-0.73
Departamento_Callao	-0.99	0.00	0.37	-1.23	-0.75
‘Departamento_La Libertad‘	-1.15	0.00	0.32	-1.39	-0.91
Departamento_Cajamarca	-1.38	0.00	0.25	-1.63	-1.13

Once the Wald test has been indicated to observe the behavior of the adjustment of the variables in the model; we proceed to carry out the goodness of fit tests for the model

H_0: The model fits the data well

H_1: El model does not fit the data well

Tests of goodness of fit				
Test	**x-squared**	**gl**	**p-value**	**Conclution**
Deviation	2089.211	25	0.0	Rejection
Chi squared	32.021	1	0.0	Rejection
Hosmer Lemeshow	12.343	8	0.1366	No Rejection

Table 4.2: *Goodness-of-fit test table of the model*

Now we proceed to the analysis of the confusion matrix of the model. From table 4.2 the accuracy is 72.5% in the training and 72% in the test, this means that 100 cases of model 71 are being correctly assigned among children who have and do not have anemia. And if we look at the sensitivity metric with respect to my objective category, children with anemia, it is only predicting with 38% of cases with anemia.

Predictions in training data		
Observations	No anemia	With anemia
No anemia	7873	1069
With anemia	2548	1673

Predictions in the test data		
Observations	No anemia	With anemia
Sin anemia	3412	419
Con anemia	1136	673

Table 4.3*: Confusion matrix of training and validation data*

From table 4.3, the accuracy metric of 72.5% was obtained in the training data and 72.4% in the test data, this indicates that out of every 100 cases it predicted 72 cases of correct wood between both categories (With anemia and without anemia). Another of the metrics of interest in the study is specificity, where 40% in the training data and 37.3% in the test data, this means that for every 100 children with anemia the model correctly predicts 37 cases in average.

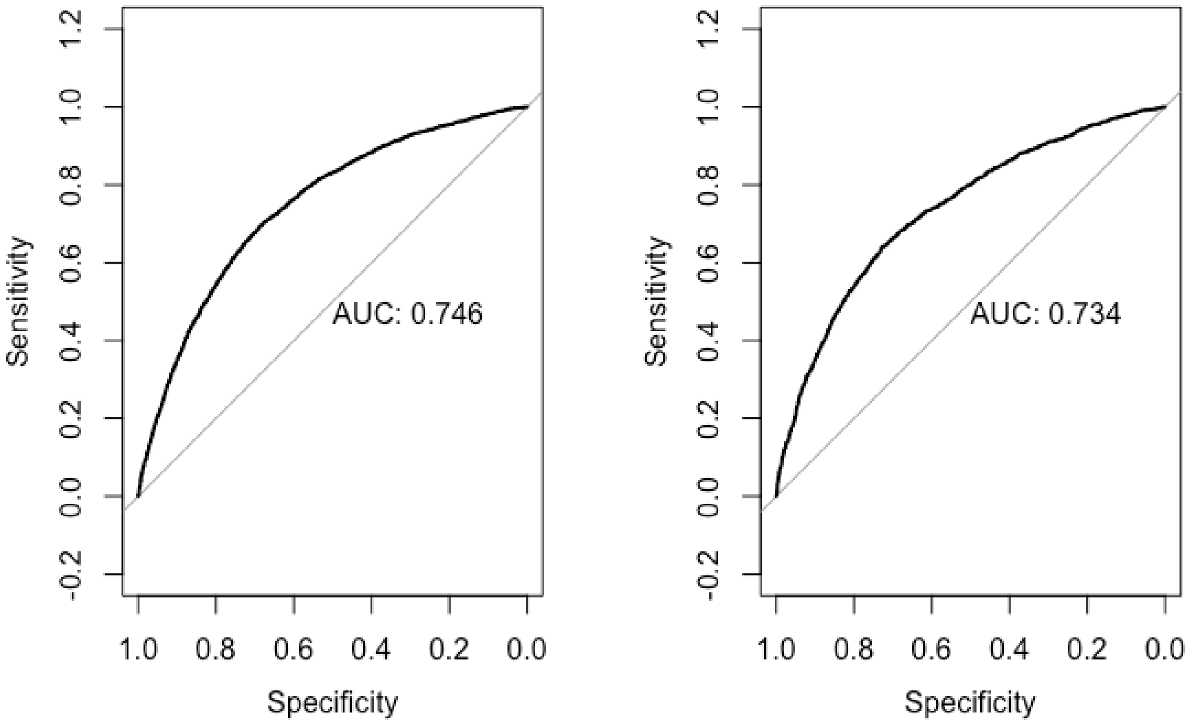


Figure 4.6: *ROC curve of training data and test data*

Figure 4.8 shows the model efficiency where it indicates under the AUC criterion the model is good at predicting both classes of the model.

Balancing logistic regression model

As a consequence of the results obtained from the unbalanced binary logistic regression modeling, where the specificity metric is 27%, we proceed to perform the binary logistic regression modeling with balanced data using the Smote technique.

Applying Smote to balance the data

Balanced dating Smote	
No anemia: 6375	With Anemia: 6376

Table 4.4: Balanced data

Variables significance of the binary logistic regression model						
Variable	Coef	P.Wald	pvalue	Odd	Ic2.5%	Ic 97.5%
(Intercept)	5.82	34.80	0	338.04	5.50	6.15
‘nivel_edu_madre_Sin educacion‘	0.97	5.79	0	2.64	0.64	1.30
Departamento_Puno	0.93	7.08	0	2.54	0.67	1.19
nivel_edu_madre_Primaria	0.62	8.57	0	1.85	0.48	0.76
nivel_edu_madre_Secundaria	0.47	7.80	0	1.60	0.35	0.59
Sexo_Hombre	0.33	8.37	0	1.39	0.25	0.41
Altura_centi	-0.07	-38.51	0	0.93	-0.07	-0.06
Departamento_Apurimac	-0.32	-2.84	0	0.73	-0.53	-0.10
Departamento_Moquegua	-0.33	-2.88	0	0.72	-0.55	-0.11
Departamento_Ayacucho	-0.39	-3.60	0	0.68	-0.60	-0.18
Departamento_Tumbes	-0.47	-4.73	0	0.62	-0.67	-0.28
Departamento_Ancash	-0.51	-4.57	0	0.60	-0.73	-0.29
Departamento_Ica	-0.57	-5.40	0	0.57	-0.77	-0.36
Departamento_Huanuco	-0.58	-5.45	0	0.56	-0.79	-0.37
Departamento_Arequipa	-0.65	-5.73	0	0.52	-0.87	-0.43
Departamento_Lambayeque	-0.68	-6.14	0	0.51	-0.90	-0.46
Departamento_Piura	-0.70	-6.64	0	0.49	-0.91	-0.50
‘Departamento_San Martin‘	-0.73	-7.05	0	0.48	-0.93	-0.53
Departamento_Tacna	-0.78	-6.43	0	0.46	-1.01	-0.54
Departamento_Amazonas	-0.82	-7.66	0	0.44	-1.02	-0.61
Departamento_Lima	-0.86	-12.38	0	0.42	-0.99	-0.72
Departamento_Callao	-0.90	-8.03	0	0.40	-1.13	-0.68
‘Departamento_La Libertad‘	-1.18	-10.41	0	0.31	-1.40	-0.95
Departamento_Cajamarca	-1.45	-12.78	0	0.23	-1.67	-1.23

Table 4.5: *Summary of significance and Odd ratio of the balanced model*

Table 4.5 shows the estimated parameter, p value associated with the Wald test, where it is observed that the factors are significant in the model with a significance level of 5%, the Odd ratio and the confidence interval at which belongs.

In table 4.6, the metric of accuracy in the balanced data is 68% and the metric of interest according to objective agreement of the prevalence of anemia, the metric of sensitivity which is 71%; this is with respect to testing on balanced data. In figure 4.10, show the efficiency of the model is acceptable with an AUC of 74.3%

Smote predictions			
Observations	No anemia	With anemia	Sensitivity
No anemia	4333	2075	67.6%
With anemia	1933	4476	70%

Predictions in training data			
Observations	No anemia	With anemia	Sensitivity
No anemia	6008	2907	67.3%
With anemia	1316	2957	69.2%

Predictions in the test data			
Observations	No anemia	With anemia	Sensitivity
No anemia	2573	1318	66.1%
With anemia	504	1256	71.1%

Table 4.6: *Confusion matrix of training and validation data in balancing*

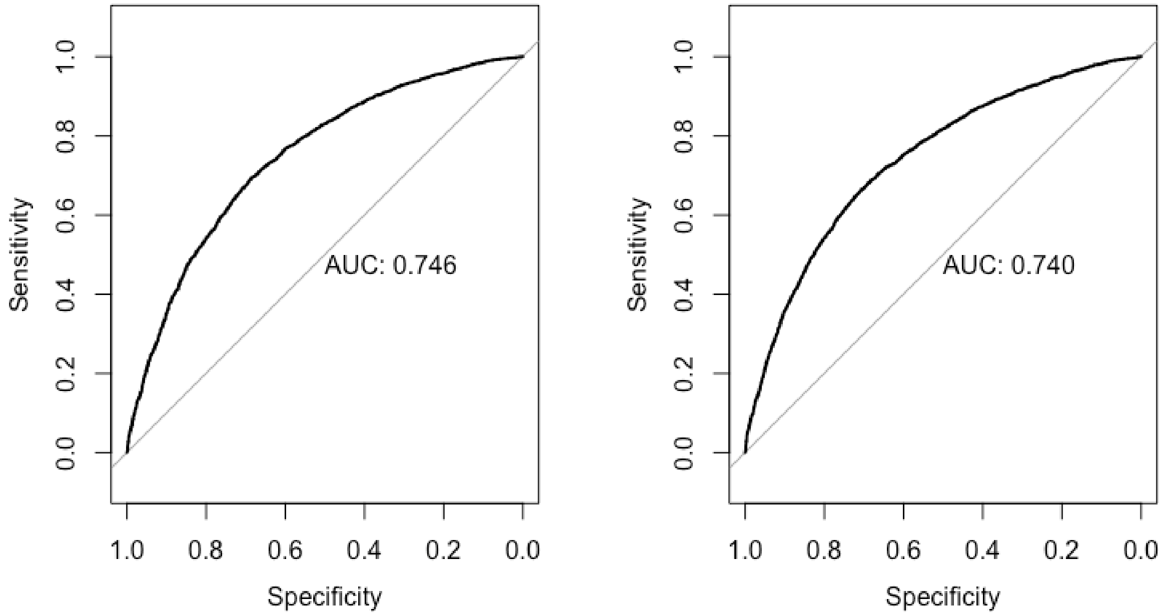


Figure 4.7: *ROC curve of balanced data and training data respectively*

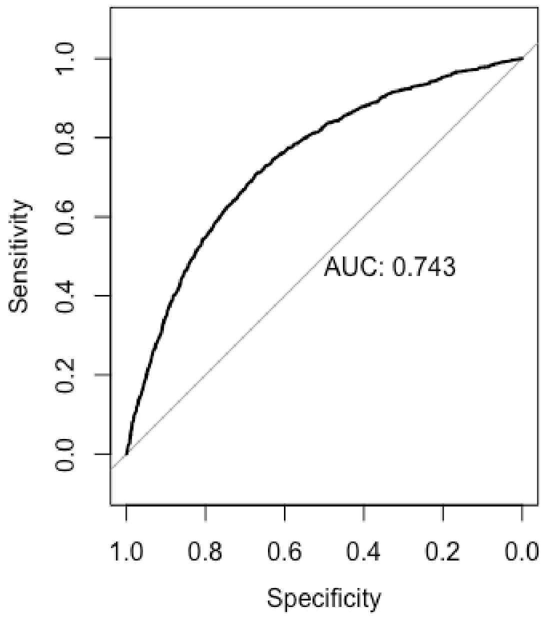


Figure 4.8: ROC curve of test data

Selected Model

Binary Logistic Regression Model	Training	Test
Accuracy	72.5%	72.4%
Sensitivity	40%	37.3%
AUC	74.6%	73.4%

***Table 4.7**: Summary of the metrics in the data without balancing*

Binary Logistic Regression Model	Smote	Training	Test
Accuracy	69%	68%	68%
Sensitivity	70%	69%	71%
AUC	74.6%	74%	74.3%

***Table 4.8:** Summary of metrics in balanced data*

Regarding the accuracy metric, there is a slight difference between the unbalanced and balanced models, in the AUC indicator there is no difference between the applied models.

And according to the sensitivity metric, which seeks which model best predicts children with anemia between the ages of 0 to 59 months of age, for this indicator the selected model would be the binary logistic regression model with balancing.

From table 4.5, from the fifth column known as Odd, interpretation. It should be mentioned for those values less than 1, it will be interpreted from its inverse:

- The result that the child from 0 to 59 months of age has anemia is 2.54 more likely if the child belongs to the department of Puno.
- The result that the child from 0 to 59 months of age has anemia is 2.64 more likely if the child's mother has no education.
- The result that the child from 0 to 59 months of age has anemia is 1.85 more likely if the mother of the child has elementary school as a maximum grade
- The result that the child from 0 to 59 months of age has anemia is 1.60 more likely if the mother of the child has a secondary education at the high school.
- The result that the child from 0 to 59 months of age has anemia is 1.39 more likely if the child is male.
- The result that the 0-59-month-old child has anemia is 1.07 less when the child's height increases by one.
- The result that the child from 0 to 59 months of age has anemia is 4.34 less likely if the child lives in the department of Cajamarca.

CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

1. There is a high probability that if the child from 0 to 59 months of age is male, belongs to the department of Puno, the mother of the child does not have an education or a maximum grade of primary or secondary education, the child has anemia
2. The intensity of greater weight to indicate that the child has anemia with a high probability is the factor if the child's mother does not have an education with an average 2.64 more likely that the child has anemia, then the factor that If the child belongs to the department of Puno, the child's mother does not present an education with.
3. According to the accuracy, sensitivity and AUC metrics, the rolling model has a better predictive power than the data without balancing, it is worth mentioning that our study is to predict children with anemia. The metric measurements were 74.6%, 74% and 74.3% respectively.

5.2. Recommendation

- The indicators are acceptable, but it should be mentioned that anemia is a health problem, which would seek a better predictive capacity in the indicators. It would be recommended to introduce other variables such as SES, among others; apply another classification model such as Neural Networks, Classification Trees among others.

BIBLIOGRAPHY

1. Moschovis PP, Wiens MO, Arlington L, Antsygina O, Hayden D, Dzik W, Kiwanuka JP, Christiani DC, Hibberd PL. Individual, maternal, and household risk factors for anemia among young children in sub-Saharan Africa: a cross-sectional study.

2. Assessing the iron status of populations: report of a joint World Health Organization / Centers for Disease Control and Prevention technical consultation on the assessing of iron status at the population level, 2nd ed., Geneva, World Health Organization, 2007.https://www.who.int/nutrition/publications/micronutrients/ anaemia_iron_deficiency/9789241596107.pdf

3. World Health Organization (2013): Hemoglobin concentrations to diagnose anemia and assess its severity. Vitamin and Mineral Nutrition Information System. WHO.

4. Pasricha, Sant-Rayn; Caruana, Sonia R.; Phuc, Tran Q.; Casey, Gerard J.; Jolley, Damien; Kingsland, Sally et al. (2008): Anemia, Iron Deficiency, Meat Consumption, and Hookworm Infection in Women of Reproductive Age in Northwest Vietnam. In the American Journal of Tropical Medicine and Hygiene 78 (3).

5. National Institute of Statistics and Informatics (Ed.) (2019): Demographic and Family Health Survey 2019. INEI. Lima Peru.

6. Herrera Gómez, Marcos (2008): An introduction to multilevel analysis: Is individual health demand affected by the family doctor? Economic Sciences Program. Zaragoza's University. Available online at http: //mpra.ub.uni- muenchen.de/35267/.

7. Porquecas.V and Chapilliquen.R, Factors associated with anemia in children from 6 to 36 months of age treated at the Leoncio Amaya Tume Essalud medical center - First semester 2019.

8. World Health Organization. Joint Declaration of the World Health Organization and the United Nations Children's Fund: Anemia at the center of attention, towards an integrated approach for effective control of anemia. Geneva; 2004

9. United Nations Children's Fund. UNICEF. Iron deficiency and anemia situation. Panama; 2006

10. Bazan,J .2008. A classification of asymmetric binary regression models. VolXXXI, N62.

11. Obregón, C.E (2018), Contribution of contextual individual risk factors to the increased risk of anemia in children under five years of age in Peru.

12. Medina. M, Viscount.O, Viscount.M and Minchon. B, (2013), Generalized linear models for prognosis of childhood anemia through associated factors

13. Salas.V, Kevin.E; Vilca.L, Juan. J, (2015), Predictive risk model for anemia in six-month-old infants attended by outpatient clinic during the period from July to December 2014 at the San Juan Bautista de Huaral Hospital – Peru

14. Bach.F, (2019), Socioeconomic factors and malnutrition of children under five years of age, Pisonaypata health post, Apurimac, 2017.

15. Executive Directorate of Food and Nutrition Surveillance (2019), Anemia in children under five years of age, National Institute of Health

16. Cerda. J and Cifuente. L. Use of ROC and AUC curves in clinical research.

17. Allison.P, Why I don't trust the Hosmer-Lemeshow test for logistic regression (2013)

18. Hosmer DW y Lemeshow S. (1980) A goodness-of-fit test for the multiple logistic regression model. "Communications in Statistics A10: 1043-1069.

APPENDIX

Appendix A: Standards and ISO

1. ISO 3534-1:1993

Statistics -- Vocabulary and symbols

Part 1: Probability and general statistical terms

General statistical terms and terms used in the calculation of probabilities

This standard defines general statistical terms, as well as terms used in the calculation of probabilities that can be applied in the elaboration of other technical standards.

In addition, it defines symbols for a number of these terms.

2. ISO/TR 10017:2003

Guidance on statistical techniques for ISO 9001:2001

ISO/TR 10017:2003 provides guidance on the selection of appropriate statistical techniques that may be useful to an organization in developing, implementing, maintaining and improving a quality management system in compliance with ISO 9001. This is done by examining those requirements of ISO 9001 that involve the use of quantitative data, and then identifying and describing the statistical techniques that can be useful when applied to such data.

3. International definition of variables

- **WASTING**

Low weight-for-height, or wasting, is defined as being more than two standard deviations lower than the World Health Organization (WHO) median child growth-for-height patterns. Wasting is the result of recent rapid weight loss or failure to gain weight.

- **UNDERWEIGHT**

El ow weight-for-age, or underweight, is defined as being shorter by more than two standard deviations than the median of the World Health Organization (WHO) child growth patterns.

- **STUNTING**

Stunting, or short height for age, is defined as being shorter by more than two standard deviations than the median of the World Health Organization (WHO) child growth patterns.

4. Data cleansing

- STANDARD 4-1-8: If non-imputed items are used in estimating totals or ratios (as in Rule 4-1-3 above), the risks of not using imputed data should be described. 1. Estimated totals using no imputed data implicitly imply a zero value for all missing data. These zero implicit imputations will mean that the totals estimates underestimate the actual population totals. Therefore, when reporting totals based on a non-imputed item, the response rate for that item should be noted in a footnote in the data table. 2. Proportions (averages) using no imputed data will implicitly impute the cell proportion for all missing data within the cell. This can cause inconsistencies in estimates between tables.
- STANDARD 4-1-3: For longitudinal data sets, two imputation approaches are acceptable: cross-wave imputations or cross-sectional imputations. Crossover wave Imputations can be used to fill in missing data for longitudinal analysis or cross-sectional imputations can be used. (Guideline 4-1-2C of this standard applies here, as well.)
- GUIDELINE 4-1-2C: The imputation procedures must be internally consistent, on theoretical and empirical considerations, appropriate for the analysis, and make use of the most relevant data available. If a multivariate analysis is anticipated, care must be taken to use imputations that minimize the attenuation of the underlying relationships. The Chief Statistician must review the imputation plans before implementation.

Appendix B: Limitations

- **Sampling**

When constructing a sampling plan, attention should be paid to issues such as sample size, sampling frequency, sample selection, the basis for subgroups, and various other aspects of sampling methodology.

A proper sampling requires that the sample be selected free of bias (that is, the sample is representative of the population from which it was drawn). Failure to do this will result in a poor estimate of population characteristics. In the case of acceptance sampling, unrepresentative samples may result in the unnecessary rejection of lots of acceptable quality, or the improper acceptance of lots of unacceptable quality.

Even with samples free of bias, the information derived from samples is subject to a certain degree of error. The magnitude of this error can be reduced by taking a larger sample size, but it cannot be eliminated. Depending on the specific question and the context of the sampling, the sample size required to achieve the desired level of confidence and precision may be too large to be of practical value.

- **Limited data availability.**

Availability of enough, relevant and reach data is a limitation in most of statistical processing systems. Students have to look for information from different sources, and analyze the data for determining their validity and representativeness. The availability of relevant data has an important effect on the precision and scope of the obtained results.

Appendix C: Script

```
l ibrary ( haven)

hogar <− read_sav ( " / Users / mercedes . congacha /Documents/SEMINARIO TESIS /

Desnutricion / Modulo64 /RECH0. SAV" )

dim( hogar ) head( hogar )


l ibrary ( Hmisc) Label ( hogar )

glimpse ( hogar )


l ibrary ( dplyr )

hogar_ <− dplyr : : select ( hogar , HHID, UBIGEO, HV024 , HV025 , HV026 ,NOMCCPP)

 colnames ( hogar_ ) <− c ( "HHID" , "Ubigeo" , "Departamento" , "Area_ resi " , "
 Lugar_Resi " , "Centro_pobla" )
 head( hogar )


 #=====================================================

 # Data processing

 anemia <− read_sav ( " / Users / mercedes . congacha /Documents/SEMINARIO TESIS /

 Desnutricion /Modulo74/RECH6. SAV" )

 dim( anemia)

 anemia_ <− dplyr : : select (anemia , HHID, HC1,HC2, HC3,HC27,HC57, HC61,HC64,

                                       #HC70, HC71, HC72, HC73

                                       )
 colnames(anemia_) <- c("HHID" ,"Edad_meses" ,"Peso_kg" ,"Altura_
 centi" ,"Sexo" ,"Nivel_anemia" ,"nivel_edu_madre" ,
                                  "orden_nacimiento"
                                  #"Talla/edad" ,"Peso/edad" ,"Peso/talla " ,"Imc_sd"
                                  )
 cor(anemia_)


 data<- merge( hogar_ ,
 anemia_ , by="HHID" )head(
 data )
 dim( data )


 library(DataExplorer)
plot_missing(data)


 # removing missing values from target
 data_1 <- data[ ! is.na(data$Nivel_anemia) , ]


 #valores perdidos
plot_missing(data_1)
 apply( is.na(data_1) ,2 ,sum)



 #================================================
# 1 Eliminating missing values


data_1 <- data_1 [ ! is.na ( data_1$Imc_sd ) , ]
```

```
data_1 <− data_ 1 [ ! i s . na ( data_1$orden_nacimiento ) , ] data_1 <− data_ 1 [ ! i s . na ( data_1$
nivel _edu_madre) , ] data_1 <− data_ 1 [ ! i s . na ( data_1$Peso_kg ) , ]

apply ( i s . na ( data_ 1 ) , 2 ,sum)

dim( data_ 1) head( data_ 1)


data_1$Departamento <− factor ( data_1$Departamento , labels = c ( ’Amazonas ’ , ’ Ancash ’ ,

        ’ Apurimac ’ , ’ Arequipa ’ , ’ Ayacucho ’ , ’ Cajamarca ’ , ’ Callao ’ , ’ Cusco ’ , ’ Huancavelica ’ ,
   ’Huanuco’ , ’ Ica ’ , ’ Junin ’ , ’ La Libertad ’ , ’ Lambayeque ’ ,’ Lima ’ , ’ Loreto ’ , ’Madre de Dios ’ ,
         ’Moquegua ’ , ’ Pasco ’ ,  ’ Piura ’ ,  ’Puno ’ , ’ San Martin ’ , ’ Tacna ’ ,  ’Tumbes ’ ,  ’ Ucayali ’ )

)


data_1$Area_ resi <− factor ( data_1$Area_ resi            , label=c ( "Urbano" , " Rural " ) )

data_1$Lugar_ Resi <− factor ( data_1$Lugar_ Resi , label=c ( "Ciudad" , " Peque a _ciudad" ,

                                                                        "Pueblo" , "Campo" ) )

data_1$Sexo <− factor ( data_1$Sexo , label=c ( "Hombre" , "Mujer" ) )

data_1$ nivel _edu_madre <− factor ( data_1$ nivel _edu_madre , label=c ( ’ Sin educacion ’ ,

                                                      ’ Primaria ’ , ’ Secundaria ’ , ’ Superior ’ ) )

head( data_ 1)

data_ 1 [ ’ Edad_nino ’ ] <− i f e l s e ( data_1$Edad_meses>48 , ’ edad_5 ’ ,

                                  i f e l s e ( data_1$Edad_meses >36 , ’ edad_4 ’ ,

                                        i f e l s e ( data_1$Edad_meses>24 , ’ edad_3 ’ ,

                                              i f e l s e ( data_1$Edad_meses>12 , ’ edad_2 ’ , ’
                                              edad_1 ’ ) ) ) )

plot_missing ( data_ 1)
```

```
data_1$ Nivel _anemia <− ifelse ( as . numeric ( data_1$ Nivel _anemia) == 4 , 0 , 1 )

table ( data_1$ Nivel _anemia )

data_1$ Nivel _anemia <− factor ( data_1$ Nivel _anemia , labels = c ( ’Con anemia ’ ,

library ( fastDummies)

colnames ( data_ 1)

data_ dico <− dummy_ cols ( data_ 1 , select _columns=c ( ’ Departamento ’ , ’ Sexo ’ ,

                                                        ’ nivel _edu_madre ’ , ’ Edad_nino ’ ) )

colnames ( data_ dico )


tabla _ final <− data_ dico [ , ! colnames ( data_ dico ) %in % c ( ’HHID’ , "Ubigeo" ,

                                                                      ’ Departamento ’ ,

                                                     ’ Lugar_ Resi ’ , ’ Centro_pobla ’ , ’ Sexo ’ ,

                                             ’ nivel _edu_madre ’ , ’ Area_ resi ’ ’ Edad_meses ’ ) ]

head( tabla _ final )

tabla _ final $ Nivel _anemia <− ifelse ( as . numeric ( tabla _ final $ Nivel _anemia) == 2 , 1 , 0
)

table ( tabla _ final $ Nivel _anemia)

colnames ( tabla _ final )

cor ( tabla _ final )

head( tabla _ final )

tabla _ final [ 10 ] colnames ( tabla _ final ) for ( i in 5 : 41 ) {

      print ( colnames ( tabla _ final [ i ] ) )
```

```
            print ( chisq . test ( tabla _ f i nal [ i ] , tabla _ f i nal $ Nivel _anemia ) )

}

chisq . test ( tabla _ f i nal $ Nivel _anemia , tabla _ f i nal $Departamento_Amazonas)


chisq . test ( tabla _ f i nal [ 15 ] , tabla _ f i nal $ Nivel _anemia)
( desnutricion1$ Nivel _anemia)


tabla _ f i nal $Peso_kg <− tabla _ f i nal $Peso_kg / 10

tabla _ f i nal $ Altura _ centi <− tabla _ f i nal $ Altura _ centi / 10


########################################################################################
############################# STATISTICAL MODELING #####################
##############################################################################

# Selection of training sample (70%) and Validation (30%)

l ibrary ( caret )

set . seed( 123 )

index <− create DataPartition ( tabla _ f i nal $ Nivel _anemia , p= 0 . 7 , l i s t =FALSE)

training <− tabla _ f i nal [ index , ] testing <−      tabla _ f i nal [ − index , ]


# Checking the structure of partitioned data

100 *prop . table ( table ( tabla _ f i nal $ Nivel _anemia ) )

100 *prop . table ( table ( training $ Nivel _anemia ) )

 100 *prop . table ( table ( testing $ Nivel _anemia ) )
```

```
head( training )

summary( tabla _ f i nal )

modelo_2 <− glm ( Nivel _anemia ~ Altura _ centi      + Departamento_Amazonas +

                         Departamento_Ancash + Departamento_Apurimac

                         + Departamento_Arequipa +

                         Departamento_Ayacucho + Departamento_Cajamarca

                         + Departamento_ Callao

                         + Departamento_Huanuco + Departamento_ Ica +

                         ‘ Departamento_La Libertad ‘ + Departamento_Lambayeque

                         + Departamento_Lima +

                         Departamento_ Loreto + Departamento_Moquegua

                         +

                         Departamento_ Piura + Departamento_Puno

                         + ‘ Departamento_San Martin ‘ +

                         Departamento_Tacna + Departamento_Tumbes + Sexo_Hombre +

                         ‘ nivel _edu_madre_ Sin educacion ‘ + nivel _edu_madre_ Primaria + nivel
                         _edu_madre_Secundaria+

                         Edad_nino_edad_ 3 , family = binomial , data=training )

summary( modelo_ 2)

# Advantage Ratio (Odd Ratio)

l ibrary (MASS)

OR <− exp ( coef ( modelo_ 2 ) ) # ODDS_RATIO calculation

Probabilidad <− 100 * (OR / (1 + OR) ) # Probability calculation
```

```
z <− summary( modelo_ 2) $coefficients [ , 1 ] /summary( modelo_ 2) $coefficients [ , 2 ]

z

p <− (1 − pnorm( abs ( z ) , 0 , 1 ) ) * 2     #We use pnorm() to estimate

la probabilidad , dado que es una f u n c i n de prob acumulada usamos 1−pnorm( ) p


# Advantage Ratio and Confidence Interval at 95%

a= cbind ( Coef=round ( modelo_2$coef , 2 ) , pvalue=round ( p, 2 )        , OR=round (OR, 2 ) ,

round ( confint . default ( modelo_ 2 ) , 2 ) )


head( tabla _ f i nal )

modelov1 = dplyr : : select ( tabla _ f inal , Altura _ centi , orden_nacimiento ,
Departamento_Amazonas ,

                                    Departamento_Ancash , Departamento_Apurimac ,
                                    Departamento_Arequipa ,

                                    Departamento_Ayacucho , Departamento_Cajamarca ,
                                    Departamento_ Callao

                                    , Departamento_Huanuco , Departamento_ Ica ,

                                    ‘ Departamento_La Libertad ‘    , Departamento_Lambayeque ,
                                    Departamento_Lima ,

                                    Departamento_ Loreto , Departamento_Moquegua       ,
                                    Departamento_ Piura , Departamento_Puno ,

                                    ‘ Departamento_San Martin ‘    ,

                                    Departamento_Tacna , Departamento_Tumbes , Sexo_Hombre , ‘
                                    nivel _edu_madre_ Sin educacion ‘        ,
                          nivel _edu_madre_ Primaria ,

                          nivel _edu_madre_Secundaria , Nivel _anemia)

head( modelov1 )
# write . csv ( modelov1 , f i l e ="Documents /CAPSTONE/ modelov1 . csv " , row. names = F)

# Variable Selection
l ibrary (MASS)
step <− stepAIC ( modelo_ 2 , direction="backward" , trace =FALSE)
step$anova

summary( modelo_ 2) $coefficients [ , 1 ]
summary( modelo_ 2) $coefficients [ , 2 ]
z <− summary( modelo_ 2) $coefficients [ , 1 ] /summary( modelo_ 2) $coefficients [ , 2 ] z
p <− (1 − pnorm( abs ( z ) , 0 , 1 ) ) * 2     # We use pnorm() to estimate the probability, d
p

# prediction:
predicciones <− i f e l s e ( modelo_2$ f i t t e d . values > 0. 49 , 1 , 0 )
matriz _confusion <− table ( modelo_2$model$ Nivel _anemia , predicciones ,
                                                 dnn=c ( ’ observaciones ’ , ’ predicciones ’ ) )
matriz _confusion
```

```
h <− sum( diag ( matriz _confusion ) )

h/sum( matriz _confusion )


l ibrary (ROCR)

pred1<− predict . glm ( modelo_ 2 , newdata = training , type= ’ response ’ ) pred <− ROCR: :
prediction ( pred1 , training $ Nivel _anemia)

perf <− performance ( pred , ’ tpr ’ , ’ fpr ’ )

plot ( perf )


auc_ train <− performance ( pred1 , measure = "auc" )@y. values [ [ 1 ] ]

cat ( "AUC" , auc_ train , "n" )

# prediction:


test _prob = predict ( modelo_ 2 , newdata = testing , type = " response" )

test _roc = roc ( testing $ Nivel _anemia ~ test _prob , plot = TRUE, print . auc = TRUE)


predicciones _1 <− i f e l s e ( test _prob> 0. 49 , 1 , 0 )

matriz _confusion_1 <− table ( testing $ Nivel _anemia , predicciones _ 1 ,

                                                        dnn=c ( ’ observaciones ’ , ’ predicciones ’ ) )

matriz _confusion_1
```

```
###############################################################################

# imputing values

hogar <− read_sav ( " / Users / mercedes . congacha /Documents/SEMINARIO TESIS /

Desnutricion / Modulo64 /RECH0. SAV" )

dim( hogar ) head( hogar )


l ibrary ( Hmisc) Label ( hogar )

glimpse ( hogar )


l ibrary ( dplyr )

hogar_ <− dplyr : : select ( hogar , HHID, UBIGEO, HV024 , HV025 , HV026 ,NOMCCPP) colnames (
hogar_ ) <− c ( "HHID" , "Ubigeo" , "Departamento" , " Area_ resi " , " Lugar_ Resi " , "
Centro_pobla" )


head( hogar )


#=======================================================

# Data processing

anemia <− read_sav ( " / Users / mercedes . congacha /Documents/SEMINARIO TESIS /
Desnutricion

dim( anemia)

anemia_ <− dplyr : : select ( anemia , HHID, HC1, HC2, HC3, HC27, HC57, HC61, HC64,

                                    #HC70, HC71, HC72, HC73

)
```

```
colnames ( anemia_ ) <− c ( "HHID" , "Edad_meses" , "Peso_kg" , " Altura _ centi " , "Sexo" , " Nivel
_anemia" , " nivel _edu_madre" ,
                                    "orden_nacimiento"
                                    #" Talla / edad" , " Peso / edad" , " Peso / talla " , " Imc_sd"
)
cor ( anemia_ )
data<− merge ( hogar_ , anemia_ , by="HHID" ) head( data )
dim( data )

l ibrary ( DataExplorer )
plot_missing ( data )

# removing missing values from target
data_1 <− data [ ! i s . na ( data$ Nivel _anemia ) , ]

# missing values
plot_missing ( data_ 1)
apply ( i s . na ( data_ 1 ) , 2 ,sum)

# deviation
dev <− modelo_2$deviance
nullDev <− modelo_2$null . deviance
```

```
modelChi <− nullDev − dev

modelChi


chidf <− modelo_2$df . null − modelo_2$df . residual chisq . prob <− 1 − pchisq ( modelChi , chidf )

chisq . prob


# We calculate the Devianza statistic, the Chi Square and the significance of # the contrasts:


sum( residuals ( modelo_ 2 , type=" deviance" ) ^2 )

1 − pchisq (sum( residuals ( modelo_ 2 , type=" deviance" ) ^ 2 ) , 1 ) # deviation


# Pearson's chi-square is an alternative method for testing the same hypothesis.

# It is only the application of the Pearson family formula to compare the number of observed events with the expected one (and no events).

sum( residuals ( modelo_ 2 , type=" pearson" ) ^2 )

1 − pchisq (sum( residuals ( modelo_ 2 , type=" pearson" ) ^ 2 ) , 1 ) # square chi


l ibrary ( ResourceSelection )

hoslem . test ( training $ Nivel _anemia , f i t t e d ( modelo_ 2 ) , 14 ) # the null hypothesis holds that the model fits the data

# Hosmer − Lemeshow Goodness-of-Fit (GOF) Test


# Selection of training sample (70%) and Validation (30%)
```

```
l ibrary ( caret )

set . seed( 123 )

index <− create DataPartition ( tabla _ f i nal $ Nivel _anemia , p= 0 . 7 , l i s t =FALSE) training <−
tabla _ f i nal [ index , ]

testing <−          tabla _ f i nal [ − index , ]


# Checking the structure of partitioned data

100 *prop . table ( table ( tabla _ f i nal $ Nivel _anemia ) )

100 *prop . table ( table ( training $ Nivel _anemia ) )

100 *prop . table ( table ( testing $ Nivel _anemia ) )


#SMOTE

l ibrary (DMwR) colnames ( training ) head( training )

training $ Nivel _anemia <− as . factor ( training $ Nivel _anemia) smote_ train <− SMOTE( Nivel
_anemia~ . ,

                              training ,

                              perc . over = 50 ,

                              perc . under = 300) %> % as . data . frame ( )

table ( training $ Nivel _anemia)

table ( smote_ train $ Nivel _anemia)


summary( tabla _ f i nal )

modelo_4<− glm ( Nivel _anemia ~        Altura _ centi    + Departamento_Amazonas +
                    Departamento_Ancash + Departamento_Apurimac
```

```
                + Departamento_Arequipa + Departamento_Ayacucho +
                Departamento_Cajamarca + Departamento_ Callao
                + Departamento_Huanuco + Departamento_ Ica + ‘ Departamento_La Libertad ‘
                + Departamento_Lambayeque + Departamento_Lima
                + Departamento_Moquegua      +
                Departamento_ Piura + Departamento_Puno
                + ‘ Departamento_San Martin ‘ +
                Departamento_Tacna + Departamento_Tumbes + Sexo_Hombre +
                ‘ nivel _edu_madre_ Sin educacion ‘ + nivel _edu_madre_ Primaria + nivel
                _edu_madre_Secundaria ,
                family = binomial ,
                data=smote_ train )
summary( modelo_ 4)


# Advantage Ratio (Odd Ratio )
l ibrary (MASS)
OR <− exp ( coef ( modelo_ 4 ) ) # ODDS_RATIO calculation
Probabilidad <− 100 * (OR / (1 + OR) ) # Probability calculation


z <− summary( modelo_ 4) $coefficients [ , 1 ] /summary( modelo_ 4) $coefficients [ , 2 ] z
p <− (1 − pnorm( abs ( z ) , 0 , 1 ) ) * 2     # We use pnorm () to estimate the probability,
since it is a cumulative probability function we use 1−pnorm( ) p
```

```
# Advantage Ratio and Confidence Interval at 95%

a= cbind ( Coef=round ( modelo_4$coef , 2 ) , wald=round ( z , 2 ) , pvalue=round ( p, 2 )            ,
OR=round (OR, 2 ) , round ( confint . default ( modelo_ 4 ) , 2 ) )


# Difference of residues

dif _ residuos <− modelo_4$null . deviance − modelo_2$deviance


# Degrees of freedom

df <− modelo_4$df . null − modelo_4$df . residual

# p−value

p_ value <− pchisq ( q = dif _ residuos , df = df , lower . t a i l = FALSE)

test _smo = predict ( modelo_ 4 , newdata = smote_ train , type = " response" ) test _roc _smo =
pROC: : roc ( smote_ train $ Nivel _anemia ~ test _smo,

plot = TRUE, print . auc = TRUE)

predicciones _1 <− i f e l s e ( test _smo> 0 . 49 , 1 , 0 )

matriz _confusion_1 <− table ( smote_ train $ Nivel _anemia , predicciones _ 1 ,

dnn=c ( ’ observaciones ’ , ’ predicciones ’ ) )

matriz _confusion_1


h <− sum( diag ( matriz _confusion_ 1 ) ) h/sum( matriz _confusion_ 1)


test _smo = predict ( modelo_ 4 , newdata = testing , type = " response" ) test _roc _smo = pROC: :
roc ( testing $ Nivel _anemia ~ test _smo,

plot = TRUE, print . auc = TRUE)
```

```
predicciones _1 <− i f e l s e ( test _smo> 0 . 49 , 1 , 0 )
matriz _confusion_1 <− table ( testing $ Nivel _anemia , predicciones _ 1 ,
                              dnn=c ( ’ observaciones ’ , ’ predicciones ’ ) )
matriz _confusion_1

h <− sum( diag ( matriz _confusion_ 1 ) )
 h/sum( matriz _confusion_ 1)

test _smo = predict ( modelo_ 4 , newdata = training , type = " response" )
test _roc _smo = pROC: : roc ( training $ Nivel _anemia ~ test _smo,
plot = TRUE, print . auc = TRUE)
predicciones _1 <− i f e l s e ( test _smo> 0 . 49 , 1 , 0 )
matriz _confusion_1 <− table ( training $ Nivel _anemia , predicciones _ 1 ,
                                                                  dnn=c ( ’ observaciones ’ , ’ predicciones ’ ) )
matriz _confusion_1

h <− sum( diag ( matriz _confusion_ 1 ) )
 h/sum( matriz _confusion_ 1)

par ( mfrow=c ( 1 , 2 ) )
train _prob = predict ( modelo_ 4 , newdata = smote_ train , type = " response" ) train _roc =
pROC: : roc ( smote_ train $ Nivel _anemia ~ train _prob ,
plot = TRUE, print . auc = TRUE)
```

```
train _prob = predict ( modelo_ 4 , newdata = training , type = " response" ) train _roc = pROC: :
roc ( training $ Nivel _anemia ~ train _prob ,

plot = TRUE, print . auc = TRUE)

test _prob = predict ( modelo_ 4 , newdata = testing , type = " response" ) test _roc = pROC: : roc (
testing $ Nivel _anemia ~ test _prob ,

plot = TRUE, print . auc = TRUE,

levels = c ( 0 , 1 ) , direction = "<" )
```