Statistical Modelling for Determinants of Anaemia among Women at Reproductive Age

Pooja S. Zanjurne[1], Vaishali V. Patil[2]
[1]*Research Scholar, Department of Statistics, Shri JJT University, Jhunjhunu, Rajasthan, India*
[2]*Research Guide, Department of Statistics, Shri JJT University, Jhunjhunu, Rajasthan, India*
Corresponding Author: Pooja S. Zanjurne, Email: poojaszanjurne21@gmail.com

Abstract:

Anaemia is a condition where the amount of red blood cells or the amount of haemoglobin they contain is unusually low. The blood's capacity to carry oxygen to the body's tissues will be diminished. The most common causes of anaemia are dietary deficiencies, particularly iron deficiency, as well as viral diseases like malaria, TB, and HIV and nutritional inadequacies. Anaemia primarily affects women between the ages of 15 and 49 who are of reproductive potential. WHO estimates that 40% of pregnant women have anaemia. According to a WHO report, anaemia affects three out of every ten non-pregnant women. Anaemia stunts children's mental development and doubles the risk of pregnancy death. According to recent studies, Indian women who are of childbearing age have an anaemia prevalence that is typically 20% higher than the global average. The purpose of this study is to predict anaemia among women at reproductive age and also find the significant factors associated with anaemia. To predict anaemia decision tree and ordinal logistic regression model were developed. It was found that the individual, household and community level factors like husband's occupation, women's occupation, women's age at first marriage, number of children, husbands age, women's age at first birth, marital status, number of household members, age of women, toilet facility, women's educational level, wealth index, urban rural status, age at period, number of children under 5 year are associated with status of anaemia.

Keywords: Anaemia, WRA, Decision tree, ordinal logistic regression, confusion matrix.

1. Introduction

Anaemia is a condition when your body unable to produce sufficient healthy red blood cells in order to deliver oxygen to your tissues. Anaemia is characterized by weakness and exhaustion. There were several varieties of anaemia, and each has its own particular set of causes. Anaemia can ranged from mild to severe, and it can also be short-term or long-term. There are numerous causes for the various types of anaemia. I studied iron deficient anaemia in this research. This is the most typical type of anaemia, and it results from a deficiency of iron in your body. As the name suggests a lack of iron leads to iron deficiency anaemia. Our body needs iron to produce sufficient amounts of a substance in red blood cells that allows for them to carry oxygen (hemoglobin). Thus, iron deficiency anaemia may make us feel tired and exhausted.

Iron deficiency anaemia can initially be so mild that it is not observed. However, the signs and symptoms grow as the body loses more iron and the anaemia increases. The group of peoples such as women, infants, children, frequent blood donors, vegetarians have increased risk of anaemia. Women at reproductive age have greater risk of anaemia because they lose blood during menstruation. The most prevalent condition is the Anaemia in women at reproductive age. Anaemia is a widespread public health issue that affects roughly one third of women of reproductive age worldwide. The WHO goal is to reduce anaemia in women of reproductive age by 50% in 2025, acknowledging it as a global public health issue. From the previous research it was observed that individual, household, and community levels factors affects the anaemia. Anaemia had a strong association with HIV infection, pregnancy, more births, and married status on an individual level. However, anaemia had a negative association with education in the secondary and above levels as well as the usage of injectable, implantable, or contraceptive tablets. At household factors living with household with large family members and with poorest, poorer and middle wealth index was positively associated with anaemia.

In this study we are interested to identify individual, household and community level factors of anaemia in women of reproductive age. Our aim is to develop model using machine learning algorithm to predict anaemia among the women of reproductive age. In this research two supervised learning models such as decision tree and ordinal logistic regression model were developed to identify the factors associated with status of anaemia and to predict anaemia.

2. Literature Review

Priyanka Anand et. al. have predicted the anaemia in children by using four Machine learning algorithms. The performance of machine learning algorithm was measured by various measures like accuracy, precision, specificity, sensitivity. Author discovered that the random forest algorithm shows great performance than other three algorithms to predict anaemia in children.

Moloud Abdar et. al. (2015) compares various data mining algorithms for prediction of heart diseases. After fitting the data mining models, model evaluation and comparison was done by using ROC, AUC and confusion matrix. It was observed that Support Vector Machine (SVM) gave best results rather than other models.

3. Research Methodology

In this research the secondary data was taken from IDHS 2015. Demographic and health surveys, which have been conducted in low- and middle-income nations since the 1980s, are easier to analyze because to IPUMS-DHS. Thousands of reliably coded variables on the health and happiness of women, children, new mothers, men, and all members of randomly chosen households are present in the IPUMS-DHS. Initially there were 6,99,686 samples. For Maharashtra state 28,648 samples were extracted. Data pre-processing was done after the data pre-processing 179 samples are taken for analysis purpose. In the final data there were total 46 variables such as status of anaemia, individual level variables like age, weight, education, occupation, husband's age husband's occupation, menstrual information etc. Community level variables like region, community education etc. Household level variables like number of household members, number of children, cooking fuel, toilet facility etc. Decision tree and ordinal logistic regression model was fitted for the prediction of anaemia.

3.1 Decision Trees

A non-parametric supervised learning technique for classification and regression is the decision tree. The goal is to understand simple decision rules derived from the data attributes in order to build a model that predicts the class label. Decision tree classifier design is appropriate for exploratory knowledge discovery because it doesn't require any parameter configuration or domain understanding. In the community of machine learning, CART is one of the often used techniques for creating decision trees. By dividing records at each node in accordance with a function of a single attribute, CART creates a binary decision tree. The best split is chosen by CART using a GINI Index. We attempt to split a root node from each of the two nodes produced by the initial split in the same way. Once more, we look over every input field to identify potential splitters. We designated a node as a leaf node if there was no split that considerably reduced the node's variety. Eventually, just the leaf node remained, and we had completed our decision tree. A new set of records cannot be properly classified using the complete tree because of over fitting.

Each training set record was assigned to a leaf of the whole decision tree at the ending of the tree-growing procedure. Now a class can be given to each leaf. The percentage of inaccurate classifications at a node is known as the error rate of a leaf node. The weighted average of the error rates at each leaf makes up the decision tree's overall error rate. The error rate of where the record will ultimately end up is the sum of the contributions of each leaf. High dimensional data can be handled via decision trees. Decision tree classifiers are often accurate. A popular inductive method for learning classification information is decision tree induction.

3.1 Ordinal Logistic regression

A statistical analysis technique called ordinal logistic regression can be used to simulate find the association between an ordinal response variable and one or more explanatory variables. A categorical variable with a distinct ordering of the category levels is called an ordinal variable. The explanatory variables could be categorical or continuous. Although it is not difficult to estimate ordinal logistic regression models using statistical software, doing so can be challenging. The extension of logistic regression known as ordinal logistic regression considers the linear relationship between the independent variables and the logit (or log chances) of a binary answer. There are k-1 logits if the response variable has k levels instead. The proportional odds assumption, which states that an independent variable's effect is constant with each rise in the level of the response, is a key tenet of ordinal logistic regression. As a result, the output of an ordinal logistic regression will include a single slope for each explanatory variable and an intercept for all response levels other than one. An ordinal regression model can be parameterized in a variety of ways, and many statistical software programs employ various parameterizations. So, while evaluating the results from ordinal regression models, extra caution should be exercised.

4. Statistical Analysis

Decision tree model was developed to identify the important factors which are associated with anaemia. Following table shows the variable importance for anaemia.

Variable	Importance
HUSJOB	19
CURRWORK	16
AGEFRSTMAR	10
RESIDENTYR	9
CHEB	8
HUSAGE	8
AGEAT1STBIRTH	8
MARSTAT	7
HHMEMTOTAL	5
AGE	2

BIOFHHAGE	2
TOILETTYPE	2
EDUCLVL	1
WEALTHS	1
URBAN	1
AGEMINARCHE	1
HHKIDLT5	1

Table 4.1

From the variable importance table, we can say that, husband's occupation, women's occupation, women's age at first marriage, number of children, husbands age, women's age at first birth, marital status, number of household members, age of women, toilet facility, women's educational level, wealth index, urban rural status, age at period, number of children under 5 year in household are found to be significant factors for status of anaemia.

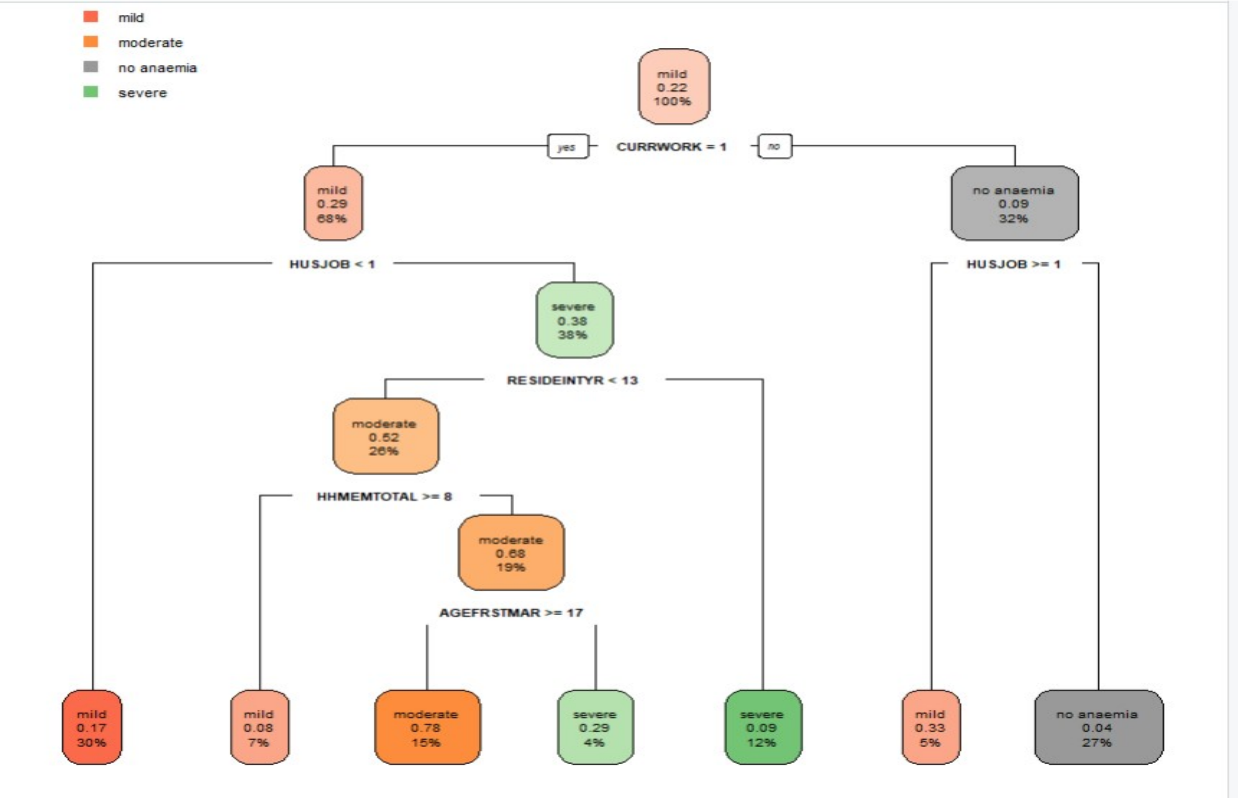

Fig. 4.1

Fig. shows that the if the women is house wife then there is 32 % sample found to be no anaemia otherwise 68% sample found to be mild anaemic. If the women currently working and her husband did not work then there is 30% sample found to be mild anaemic. Women currently working and her husband also has job then there is 38% sample found to be severe anaemic, also number of years in lives in that area is greater than 13 years then there is 12% sample found to be severe anaemic. If the women and her husband both are currently working have household members less than 8 then there is 7% sample found to be mild anaemic otherwise 19% sample found to be moderate anaemic. If the age of women at first marriage is less than 17 years then 15% of the sample shows moderate anaemia else it has severe anaemia. Ultimately it was discovered that the women's occupation is significantly associated with the status of anaemia.

Confusion matrix

	Predicted class				
		Mild	Moderate	No anaemia	Severe
Actual class	Mild	**46**	1	3	0
	Moderate	13	**21**	2	4
	No anaemia	6	1	**41**	1
	Severe	10	4	2	**24**

Table 4.2

From the above confusion matrix the accuracy is 73.74%. That means we can say that the decision tree gives 73.74% accurate results which is considerable. So the decision tree can be used to predict the anaemia among women at reproductive age (WRA).

Ordinal logistic regression

The main objective of this study is to identify the factors which are mostly affects the anaemia. From decision tree this objective was fulfilled but for more accuracy here ordinal logistic regression was developed. The results are as follows

Variable	**Coefficient**	**Std. error**	**t-value**	**p-value**
RESIDENT	107.702276	1.158125	92.9971079	0.00E+00
ELECTRC	-19.706343	1.16061584	-16.9792121	1.17E-64
WKCURRJOB	-2.05562	0.57343391	-3.58475482	3.37E-04
HUSJOB	1.24176601	0.23307649	5.32771885	9.95E-08

Table 4.3

Table show the significant factors associated with anaemia. RESIDENT (resident or visitor), ELECTRC (using electricity as a cooking fuel), WKCURRJOB (Women's occupation), HUSJOB (husband's occupation) are most significant factors associated with anaemia..It was observed that the women’s occupation and husband’s occupation and resident show significant importance with anaemia in both the models.

5. Conclusion

From both the models it was observed that In both models, it was found that factors such as husband's occupation, wife's occupation, wife's age at first marriage, number of children, husband's age, woman's age at first birth, marital status, number of household members, age of the women, availability of toilets, educational level of the women, wealth index, urban-rural status, age at period, and number of children under the age of five living in the household are associated with the status of anaemia. So not only individual level factors but also household and community level factors are associated with anaemia. Therefore, to reduce anaemia in reproductive age women we have to focus on these factors or women from these regions.

6. References

1. Abdar, M., NiakanKalhori, S. R., Sutikno, T., Much, I., Subroto, I., &Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, *5*(6), 1569–1576.
2. Alquaiz, J. M., Abdulghani, H. M., Khawaja, R. A., & Shaffi-Ahamed, S. (2012). Accuracy of Various Iron Parameters in the Prediction of Iron Deficiency Anemia among Healthy Women of Child Bearing Age, Saudi Arabia. In *Iranian Red Crescent Medical Journal Iran Red Crescent Med J* (*Vol. 14*, Issue 7).
3. Alrifaie, M. F., Ahmed, Z. H., Hameed, A. S., & Mutar, M. L. (2021). Using Machine Learning Technologies to Classify and Predict Heart Disease. *International Journal of Advanced Computer Science and Applications*, *12*(3), 123–127. https://doi.org/10.14569/IJACSA.2021.0120315
4. Anand, P., & Sharma, A. (n.d.). Prediction of Anaemia among children using Machine Learning Algorithms. https://www.researchgate.net/publication/341853966
5. Bari Antor, M., Jamil, A. H. M. S., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer’s disease. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/9917919.
6. P. S. Zanjurne, V. V. Patil (2021) Comparing the performance of data mining algorithms in the prediction of teacher’s performance. Vidyabharati *International Interdisciplinary Research Journal (Special Issue), ISSN 2319-4979.*