Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

ANAEMIA IN REPRODUCTIVE AGE WOMEN: EXPLORING KEY INFLUENCING FACTORS

Pooja S. Zanjurne ,PhD. Scholar, Shri Jagdishprasad Zabermal Tibrewala University, Rajasthan
Vaishali V. Patil, Associate Professor, T.C. College, Baramati, Maharashtra.
Farooqui M. A. Zakirhussain, Shri Jagdishprasad Zabermal Tibrewala University, Rajasthan
poojaszanjurne21@gmail.com

Abstract:
Without women, we are unable to picture how successful life would be in general. They have a major share of the blame for the continued success of life on this planet. In the past, they were only thought of as wives and mothers who had to cook, clean, and care for the entire family by themselves. While caring for responsibilities, women should take care of their own health as well, but more often than not, women tend to neglect their health. Due to both biological and gender-related distinctions, being a man or a woman has a substantial effect on one's health. Women face particular healthcare issues and are more likely than men to receive a diagnosis for some disorders. The major causes of death for women include chronic diseases and ailments such heart disease, cancer, diabetes, and anaemia. According to WHO statistics, anaemia affects 40% of pregnant women and 42% of children under the age of 5. According to DHS report 2011 anaemia affects more than 500 million women in developing countries, leaving in its wake an unacceptable burden of preventable morbidity and mortality, decreased economic productivity and lost opportunities for human, social, and economic development. According to DHS reports it observed that the anaemia affected by not only individual level factors but also household level factors and community level factors. So, the purpose of this study is to predict anaemia among the women at reproductive age (WRA). The study will also explore the key influencing factors associated with anaemia among the women at reproductive age. For the prediction decision tree and random forest algorithms was developed. It was found that random forest gives better results than that of decision tree. From the random forest algorithm key influencing factors was identified.

Keywords: WRA, Decision Tree, Random Forest, confusion matrix

INTRODUCTION:

In India, there are many factors that have an impact on women's health, including issues with maternal and reproductive health, gender-based violence, anaemia, malnutrition, limited access to healthcare, educational disparities, sanitary conditions, non-communicable diseases, stigma surrounding mental illness, and legal rights. Government and non-governmental organizations are working to solve these problems in an effort to raise the general well-being and standard of living for women in the nation.

Anaemia has diverse and significant impacts on women at various stages of life. For pregnant women, it can lead to complications such as preterm birth and low birth weight, endangering both maternal and fetal health, while causing fatigue and weakness. Women of childbearing age may experience heavy menstrual bleeding, further exacerbating their anaemia and affecting daily life. Adolescent girls with anaemia may face stunted growth and cognitive development, potentially hampering their educational and future prospects. Postmenopausal women can still develop anaemia, often due to underlying health issues, resulting in persistent fatigue and reduced quality of life.

Anaemia in older women may cause greater frailty and cognitive decline, which might affect their independence. Sports-related anaemia in female athletes can have a negative impact on their performance, and chronic health disorders in women may get worse. Anaemia is more common in low-income areas where there is less access to nourishing food and treatment, which can have an adverse effect on women's productivity and well-being since they frequently provide for their families'

nutritional needs. To lessen these profound consequences on women's health and quality of life, anaemia must be addressed through nutritional support, access to healthcare, and knowledge.

There are various factors which are affecting the status of anaemia among women at reproductive age. To identify and examine these factors is a crucial role. The aim of this research is to examine and identify the factors which are influence the anaemia in reproductive age women (WRA). For this purpose, the secondary data was taken from DHS (2015) it contains 46 variables. The anaemia was predicted by using popular machine learning techniques decision tree and Random forest. Random forest gives better accuracy than that of decision tree. From the results it was found that the Working condition, husband's job status, age at first birth, etc. found to be influential factors to status of anaemia.

RESEARCH METHODOLOGY:

In this research the secondary data was taken from DHS 2015 for India. Initially there were 6,99,686 samples of all India. 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. Dataset covers a wide range of information, including health-related factors like anaemia, pregnancy status, and various dietary habits, such as consumption of specific food items. Socio-economic indicators like household wealth, education levels, and employment status are also included. Demographic details like age, marital status, and family size are key components. Moreover, the dataset encompasses lifestyle choices like alcohol and tobacco consumption. Decision tree and Random Forest algorithm were developed for the classification of anaemia.

Decision Tree:

Decision tree is a supervised machine learning technique which is used for both classification and regression tasks. It is a simple tree-like structure where each internal node represents a predictor variable, and each leaf node corresponds to a response variable label or numerical value. The algorithm of creating a decision tree is typically begins by selecting the most influential predictor (attribute)e at each node, based on various criteria such as information gain, Gini impurity and gain ratio to partition the dataset into subsets that are as homogeneous as possible in terms of the response variable. This recursive portioning continues until a stopping condition is met, resulting in a tree that can be used to make predictions. Decision trees are easy to interpret, easy to visualize, and can handle both categorical as well as numerical data.

Random Forest:

Random Forest is a popular ensemble learning technique in machine learning which is built upon the decision tree methodology. It involves constructing a various decision trees during the training phase and combines their results to achieve more accurate and robust results. The main idea behind Random Forest is to introduce randomness in two ways: by selecting random subsets of the training data through a process known as bootstrapping, and by considering only a random subset of the features at each node of each tree. This randomness helps reduce overfitting and decorrelates the individual trees, making the ensemble model less prone to errors and more robust to various data patterns. The final prediction is typically determined by a majority vote for classification tasks or an average for regression tasks. Random Forests are popular for their high predictive accuracy, versatility, and the ability to handle large and complex datasets, making them a popular choice in various applications, including classification, regression, and feature selection.

DATA ANALYSIS

Decision Tree with 10-fold Cross validation:

The dataset contains 179 rows and 46 columns. It is divided into training and testing sets with an 80-20 split. The training set (train data) contains 80% (143) of the data, while the testing set (test_data) contains the remaining 20% (36).

After the 10 fold cross validation the results are as follows:

CART
179 samples
45 predictor
4 classes: 'mild', 'moderate', 'no anaemia', 'severe'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 161, 161, 162, 161, 161, 161, ...
Resampling results across tuning parameters:

cp	Accuracy	Kappa
0.0930233	0.625817	0.4969145
0.1937985	0.5189542	0.3426154
0.2635659	0.3633987	0.1190476

The cross-validation results suggest that a decision tree with a complexity parameter (cp) of approximately 0.093 provides the best accuracy. Accuracy was used to select the optimal model using the largest value. The final value used for the model was cp = 0.09302326.

Therefore, the CART algorithm was redeveloped by setting tuning parameter cp = 0.09302326 and the results are as follows:

The new CART (Decision Tree) algorithm was developed using 143 train sample with Complexity parameter (CP) 0.09302326.

CART algorithm:

rpart(formula = Anaemia ~ ., data = train, method = "class",
cp = 0.09302326)
n= 143

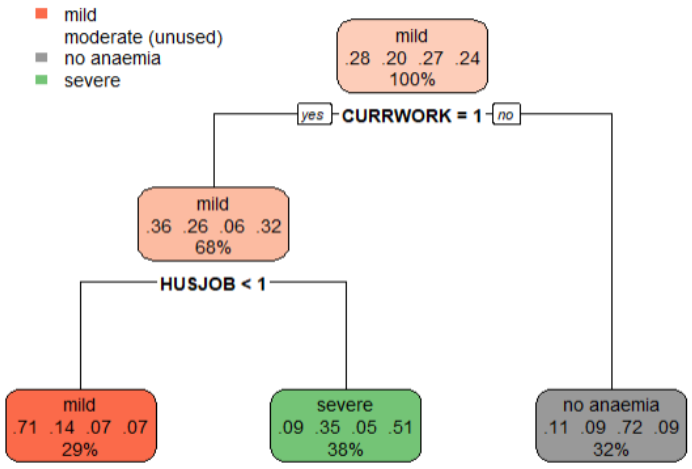
	CP	nsplit	rel error	xerror	xstd
1	0.27184466	0	1.0000000	1.0000000	0.05211267
2	0.22330097	1	0.7281553	0.7669903	0.05772957
3	0.09302326	2	0.5048544	0.5048544	0.0558492

Variable importance table:

Variable	Importance
CURRWORK	22
HUSJOB	12
AGEAT1BIRTH	11
AGEFRSTMAR	11

HUSAGE	11
MARSTAT	11
CHEB	10
AGE	1
BIOFHHAGE	1
EDYRTOTAL	1

The variable importance table displays the importance of predictor variables in the decision tree. Variables are ranked in order of their importance in making classification decisions. In this algorithm, "CURRWORK" and "HUSJOB" are the most important variables, followed by others like "AGEAT1STBIRTH," "AGEFRSTMAR," "HUSAGE," "MARSTAT" AND " CHEB".



From the above plot we can see that the Husbands job and working status of WRA is most important factors. But in the legend section we can see the moderate case is unused that says that algorithm fail to predict that class due to imbalance of data.

Confusion Matrix:

This Confusion matrix table represents the model's performance on the test data. It shows the predicted classes (Mild, Moderate, No, Severe) against the actual classes in the test dataset.

		Predicted class			
		Mild	**Moderate**	**No**	**Severe**
Actual class	**Mild**	6	0	2	2
	Moderate	3	0	1	7
	No	2	0	8	0
	Severe	1	0	0	4

The confusion matrix reveals the performance of the CART algorithm in classifying different stages of "Anaemia." The algorithm shows an overall accuracy of approximately 54.5%, indicating that it correctly predicts the Anaemia class for just over half of the cases in the test data. However, there are notable variations in its performance across different Anaemia classes. The algorithm performs well in correctly identifying cases of "No" Anaemia, achieving 100% true positives, but it struggles with the "Moderate" class, failing to predict any instances correctly. The misclassification rate, which is around 45.5%, underscores the algorithm's limitations in providing precise classifications, especially for the "Moderate" category. This suggests that while the algorithm has a moderate overall performance, there is room for improvement, particularly in accurately classifying the "Moderate" Anaemia cases. Further evaluation metrics such as precision, recall, and an understanding of the specific clinical or practical implications are essential for a more comprehensive assessment of the model's suitability for the task at hand.

To overcome this problem the ensemble algorithms may be helpful. Therefore to examine Anaemia in the WRA the ensemble algorithm Random forest was developed on same data. The results are as follows:

Random Forest:

The Random Forest algorithm is constructed using the randomForest function. The formula specifies the prediction of "Anaemia" based on all predictor variables in the training dataset. It uses a forest of 100 trees (ntree = 100), and at each split, it tries six randomly selected predictor variables. The results are as follows:

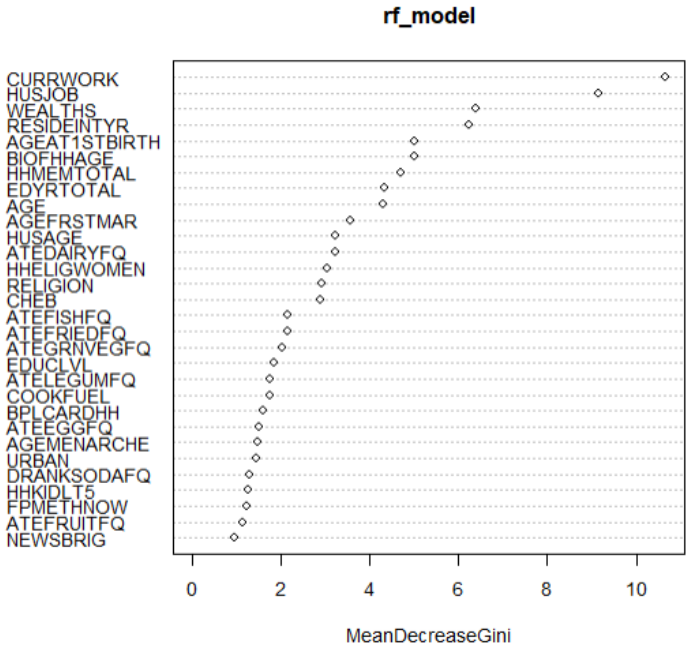
Call:
randomForest(formula = Anaemia ~ ., data = train, ntree = 100)
Type of random forest: classification
Number of trees: 100
No. of variables tried at each split: 6
OOB estimate of error rate: 42.66%
Confusion matrix:

	mild	moderate	no anaemia	severe	class.error
mild	26	5	7	2	0.3500000
moderate	11	4	2	12	0.8620690
no anaemia	5	0	31	3	0.2051282
severe	3	9	2	21	0.4000000

Variable Importance Table:

Varibale	Importance
CURRWORK	10.66916853
HUSJOB	9.155
WEALTHS	6.3937
RESIDEINTYR	6.2304
AGEAT1STBIRTH	5.01851152
BIOFHHAGE	4.9967
HHMEMTOTAL	4.69571514

EDYRTOTAL	4.3483
AGE	4.3172



The variable importance table highlights the key predictors influencing the Random Forest model's predictions for anaemia severity. Notably, "CURRWORK" (Current Work) emerges as the most influential variable, indicating that an individual's current employment status significantly impacts anaemia predictions. "HUSJOB" (Husband's Job) follows closely, suggesting that the occupation of an individual's spouse plays a pivotal role in determining anaemia outcomes, likely reflecting socioeconomic dynamics within households. "WEALTHS" (Wealth Status) emphasizes that economic well-being is a critical factor, while "RESIDEINTYR" (Residency Years) implies that the duration of residency is relevant. "AGEAT1STBIRTH" (Age at First Birth) underscores the significance of maternal age, and "BIOFHHAGE" (Age of Female Household Head) suggests the potential role of household leadership. Additionally, factors like household size ("HHMEMTOTAL"), education ("EDYRTOTAL"), and age ("AGE") are deemed important in predicting anaemia. This comprehensive view of influential variables underlines the multifaceted nature of anaemia, influenced by socioeconomic, demographic, and health-related factors. Understanding these critical predictors can guide targeted interventions and public health strategies to address anaemia more effectively.

Confusion Matrix for Random Forest:

The confusion matrix for the Random Forest model's predictions provides insights into its performance in classifying the severity of anaemia.

		Predicted class			
		Mild	**Moderate**	**No**	**Severe**
Actual class	**Mild**	7	1	0	2
	Moderate	3	3	0	5
	No	1	0	9	0
	Severe	1	0	0	4

The overall accuracy, calculated as approximately 64%, suggests that the algorithm correctly predicts the anaemia classes for about two-thirds of the cases in the test dataset. While the model performs well for some anaemia severity levels, it requires further improvement, particularly in classifying the "Moderate" cases. Since the data is only 179 so this results may change when a sample is sufficiently large. The accuracy results will be increase if we take large samples.

RESULTS & DISCUSSION

Random forest model shows relatively high accuracy than that of decision tree algorithm. Therefore, to examine the influential factors of status of anaemia the random forest algorithm was considered. From the random forest algorithm, it was observed that women's currently working situation, occupation of husband, women's socioeconomic condition, number of years a WRA has spent in a particular residence, age at which the WRA had her first child, age of female household head, total number of family members, total years of education of WRA, and her age were most influential factors on the status of anaemia. Therefore, while taking care about the anaemia among WRA we have to focus on these factors.

REFERENCES

1. Anand, P., & Sharma, A. (n.d.). Prediction of Anaemia among children using Machine Learning Algorithms. https://www.researchgate.net/publication/341853966
2. Jamnok, J., Sanchaisuriya, K., Sanchaisuriya, P., Fucharoen, G., Fucharoen, S., & Ahmed, F. (2020). Factors associated with anaemia and iron deficiency among women of reproductive age in Northeast Thailand: A cross-sectional study. BMC Public Health, 20(1). https://doi.org/10.1186/s12889-020-8248-1
3. Mog, M., & Ghosh, K. (2021). Prevalence of anaemia among women of reproductive age (15–49): A spatial-temporal comprehensive study of Maharashtra districts. Clinical Epidemiology and Global Health, 11. https://doi.org/10.1016/j.cegh.2021.100712.
4. Sunuwar, D. R., Singh, D. R., Adhikari, B., Shrestha, S., & Pradhan, P. M. S. (2021). Factors affecting anaemia among women of reproductive age in Nepal: A multilevel and spatial analysis. BMJ Open, 11(3). https://doi.org/10.1136/bmjopen-2020-041982
5. Teshale, A. B., Tesema, G. A., Worku, M. G., Yeshaw, Y., & Tessema, Z. T. (2020). Anemia and its associated factors among women of reproductive age in eastern Africa: A multilevel mixed-effects generalized linear model. PLoS ONE, 15(9 September). https://doi.org/10.1371/journal.pone.0238957