MODELING OF VITAMIN A SUPPLEMENTATION IN KILIFI, KWALE AND SIAYA COUNTIES IN KENYA

By

EUGENE ODUMA

Table of Contents

Introduction

Vitamin A is an important nutrient required for the growth and development of children and combat infections. It helps in the development and repair of bones, teeth, muscles and tissues, development of proper eyesight as well as healthy growth of soft membranes and skin cells. Vitamin A deficiency is thus detrimental to the health of children as it can result in visual impairment and increase the risks of illness. However, research by WHO shows that about 190 million preschool-age children mainly from Africa and South-East Asia suffer from Vitamin A deficiency. In Kenya, it is estimated that 9.4% of pre-school children suffer from Vitamin A deficiency. Most Countries from the affected regions including Kenya curb this deficiency by offering vitamin A supplements to affected areas as public health intervention programs. In Kenya, vitamin A supplementation is administered from six months into infancy. The results of the main study have been published here http://www.panafrican-med-journal.com/content/article/32/96/full/

The aim of this study is to identify and examine the factors influencing Vitamin A supplementation campaigns in Kenya using descriptive and predictive data analysis methods.

The code is available here

The remaining section includes data section which explains how the survey was conducted, the methodology and data analysis sections which present and explain the results of the analysis.

Data

The data for this study was provided by Dr. Shadrack Oiye on request. Dr. Shadrack Oiye is a public health researcher and nutrition consultant by profession with a focus on research, programs design, programs implementation and monitoring and evaluation in food and nutrition security and public health-related issues. He has worked and consulted for the UN, NGOs, regional organizations and the private sector. He has been extensively involved in programs as well as in researches in Kenya, Uganda, Democratic Republic of Congo, Rwanda, Zimbabwe, Ethiopia, Malawi, and Somalia. You can reach him on oiyeshad@gmail.com

It was an aggregate of data collected from caregivers of infants and young children aged 6-59 months between June to July 2016 from three counties (Siaya, Kilifi, and Kwale) in Kenya. More about data collection methods can be found here. The data was presented in three categories namely: Household survey data, ECD survey data, and Health Facility survey data from each county under the study. Although the analysis utilizes the Household survey dataset, data from other categories will be used for exploration and for further analysis. The survey was conducted in 1184 households but 7 questionnaires were dropped due to data inconsistencies thus bringing the total samples to 1177.

Methodology

The study followed a common data science methodology which includes data cleaning, data visualization, data preprocessing, model training and validation

Data cleaning

Real-world data is often messy and chaotic and that is why data munging or data cleaning forms an important part of machine learning and data analysis project. Garbage-in-garbage-out applies in data science and machine learning tasks and proper data cleaning can make or break the whole project. Different types of data require different methods of cleaning but generally, it includes tasks such as naming or renaming a variable, sorting variables, changing variable types, joining tables, conditional processing of variables, summarizing columns, appending tables, imputing missing variables, standardizing or normalizing columns, and binning continuous variables among others. Fortunately, the survey responses had already been coded hence making the cleansing task easier.

The three datasets for three counties were merged to form a single dataset for analysis. There were some variables where "I don't Know" responses were coded as 98, these were converted to

missing values and imputed with median and mode for numerical and categorical data respectively. Multiple answer questions were reduced to one answer by prioritizing the first answer provided by the respondent.

Missing value

There are different methods of handling missing values including deleting the missing observations either listwise or pairwise. However, listwise deletion rarely supports the assumption of MCAR (Missing completely at Random) while pairwise assumes that the missing data is MCAR which is not the case with our dataset. The second method involves dropping the entire variable containing missing values but this was not an option in this study due to the small dataset size. The third imputation method is by filling the missing values with mean, mode and median. This was best suited for the study because it does not take advantage of the variable relationships and characteristics.

The Household survey data was in .xlsx format but was converted to .csv. The three datasets representing the three counties had different numbers of missing data. Kilifi dataset had the highest percentage of missing values at 38.39% in the place of supplementation (Place where the child was given Vitamin A supplementation) followed by 16.96% in Polygamy (Whether the caregiver/mother is in a polygamous marriage) and the others as shown in figure 1. Kwale had relatively lower percentages of missing values with the highest being polygamy (Whether the caregiver/mother is in a polygamous marriage) and Together (Whether both parents live together) both at 12.18% followed by the rest as shown in figure 2. Similarly, Siaya had relatively lower percentage of missing values except for Together (Whether both parents live together) and Polygamy (Whether the caregiver/mother is in a polygamous marriage) at 22.27% and 20.07% respectively followed by the rest as shown in figure 3. All the missing values in the categorical data were imputed using mode while numerical data were imputed using mean.

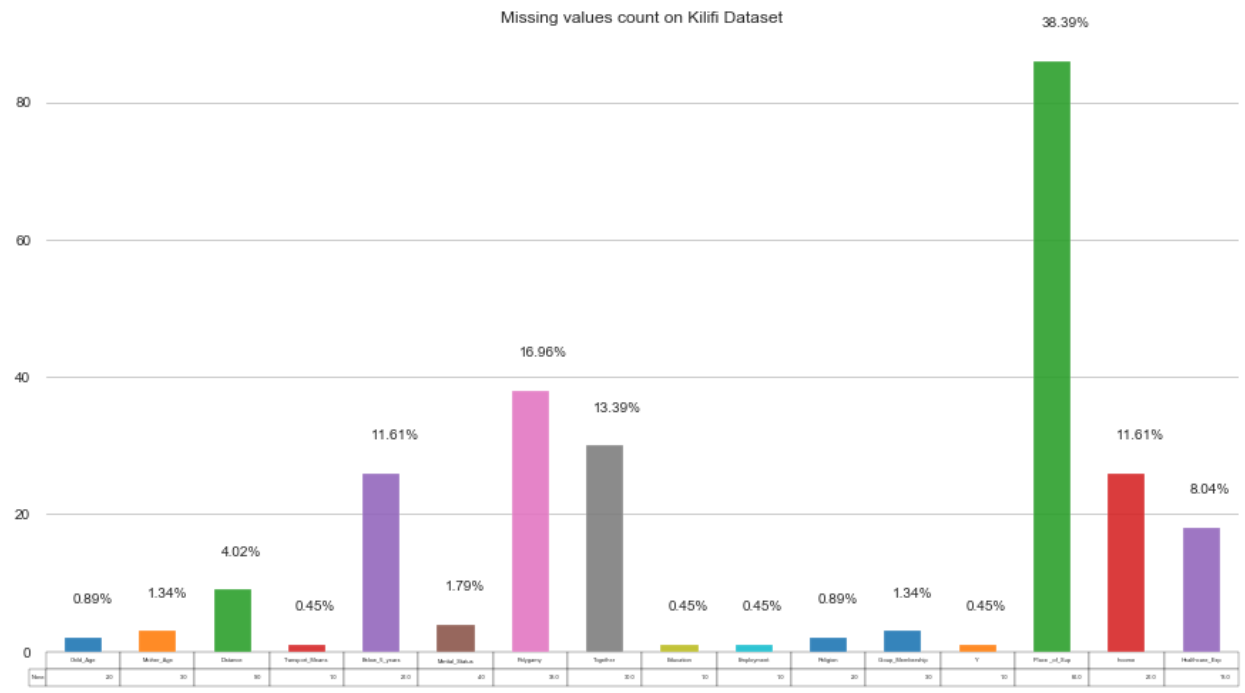

Figure 1: missing values in Kilifi dataset

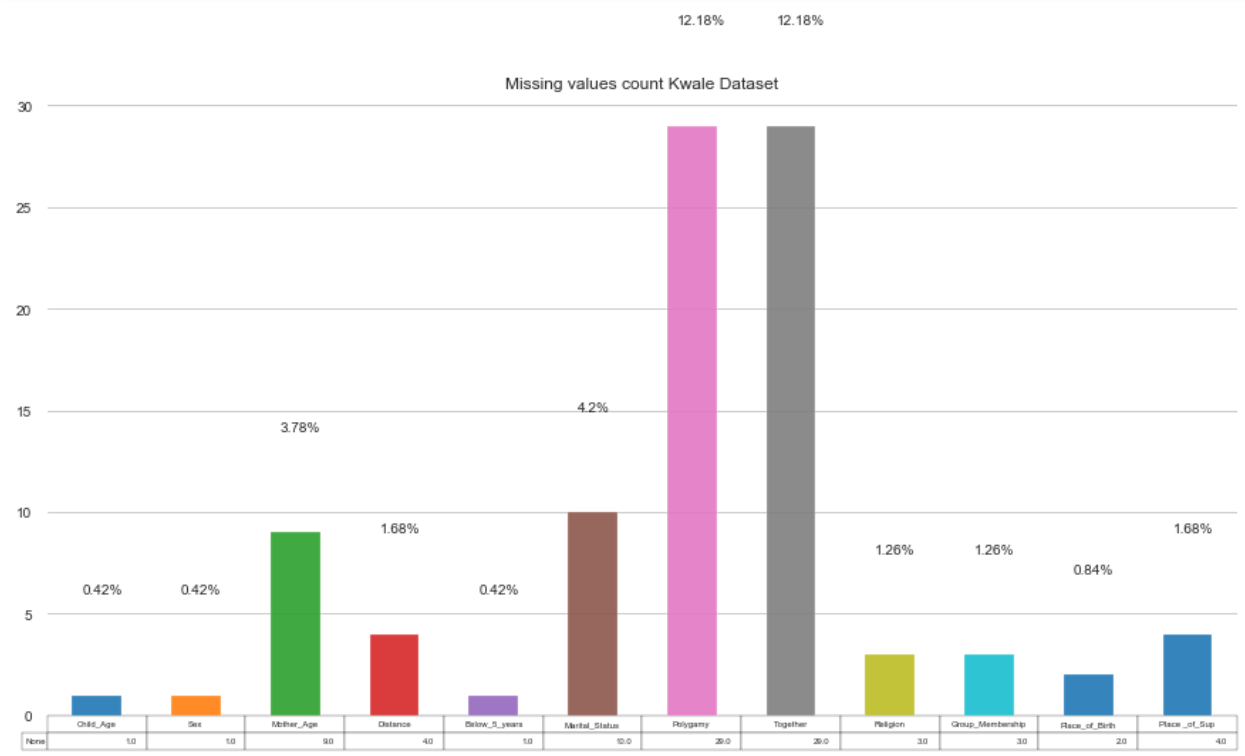

Figure 2: Missing values in Kwale dataset

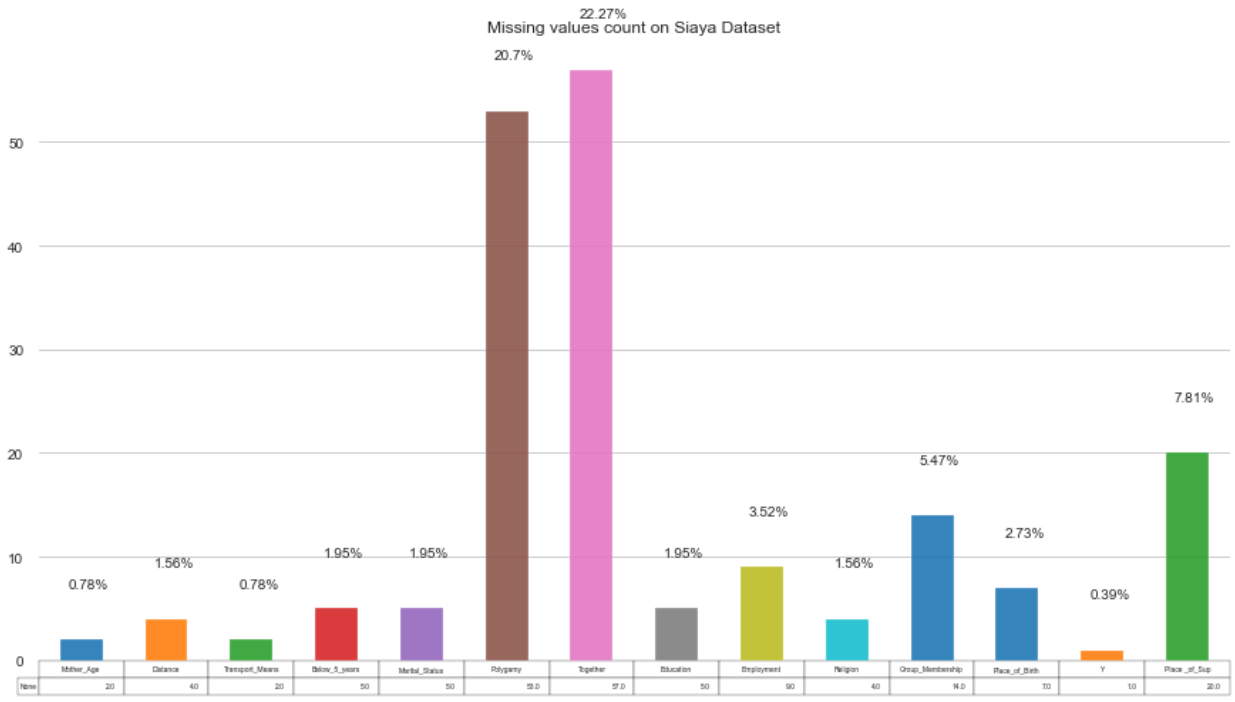

Figure 3: Missing values in Siaya dataset

Observation

Clearly, there exists a pattern in the missing values among the three countries which cannot be treated as missing at random. There was a higher percentage of missing values for two questions in the three counties: 1) IF MARRIED. Are you in a polygamous marriage 2) IF MARRIED, do you live with your spouse? Reason for the pattern in missing values was because the respondents were not married (were either single, divorced or widowed) and therefore couldn't respond to the question.

Data exploration

This section is primarily aimed at understanding patterns and relationships among different variables. The variables follow a similar trend in all counties except for Vitamin A Supplementation (independent variable), Distance to the health facility, place of supplementation, Caregiver's income, caregiver's level of education and distance to the hospital.

Supplementation

As shown in the figure below, Kilifi records the lowest number of supplemented children at 63.9% followed by Siaya and Kwale at 92.9% and 96.3% respectively. This can be used to explain any discrimination in some of the variables.

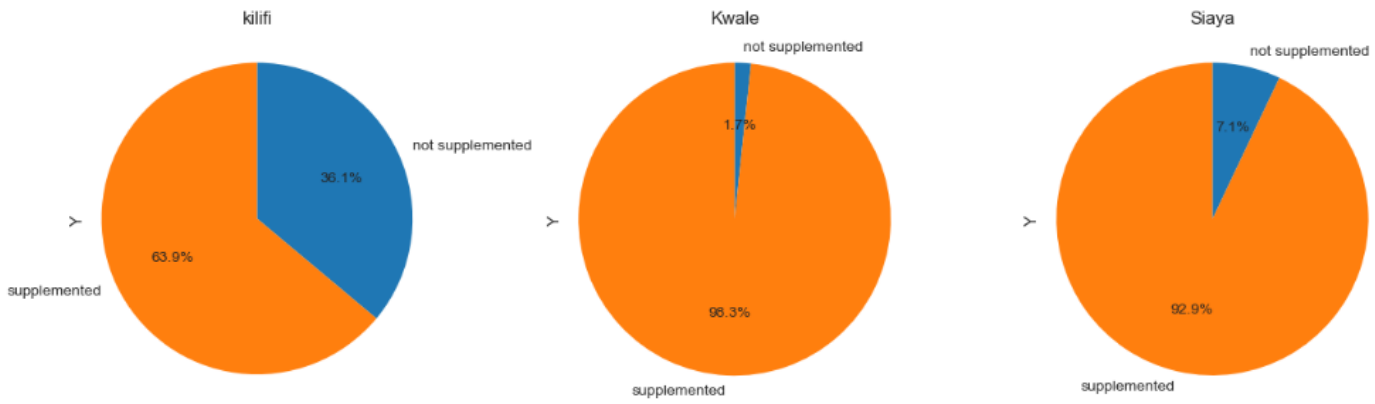

Figure 4: Vitamin A supplementation for Kilifi, Kwale, and Siaya

Distance to the health facility

As shown in the figure below Kwale county records the lowest distance to the health facility with about 31% traveling for 0.5-2 kilometers to the nearest health facility followed by Siaya county where more than 44% travel for 0.5-2 kilometers to the health facility and Kilifi county where more than 63% travel for 0.5-2 kilometers to the health facility. The 0.5-2 kilometers range was used for this analysis because it was the median range among the three counties.

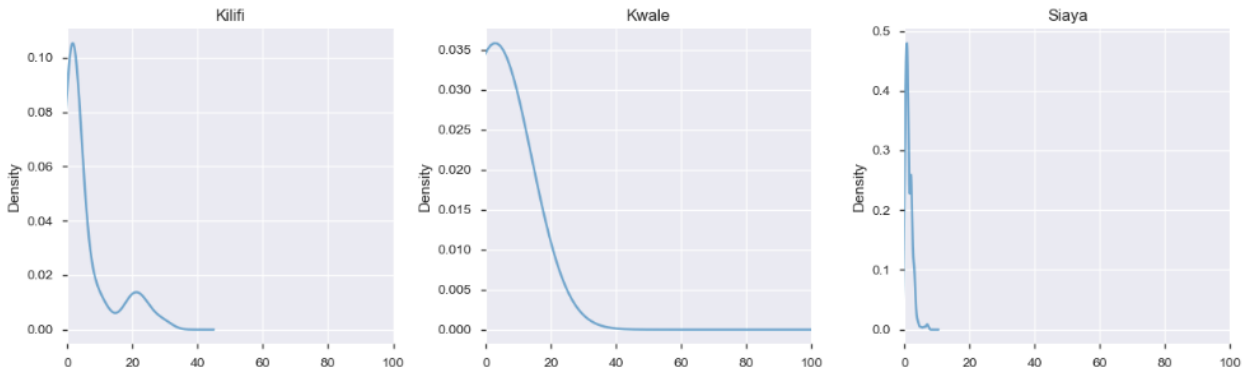

Figure 5: Distribution for distance to the health facility for Kilifi, Kwale, and Siaya

Place of Supplementation

As shown in the figure below all the counties have a varying number of supplementations taking place in different places. Vitamin A supplementations were conducted in different venues including Schools or ECDs targeting school-going children, at home and at the health facilities. In Kilifi, supplementations were majorly done at schools or the ECDs (49.3%) followed by home and (34.8%) and health facility 15.9%. In Kwale, supplementations were majorly done at home (73.1%) followed by school or ECDs (17.5%) and health facility (9.4%). In Siaya, majority of the children were supplemented at home at 43.2% followed by school or ECDs at 38.1%, health facility at 14.4% and 4.2% in other places.

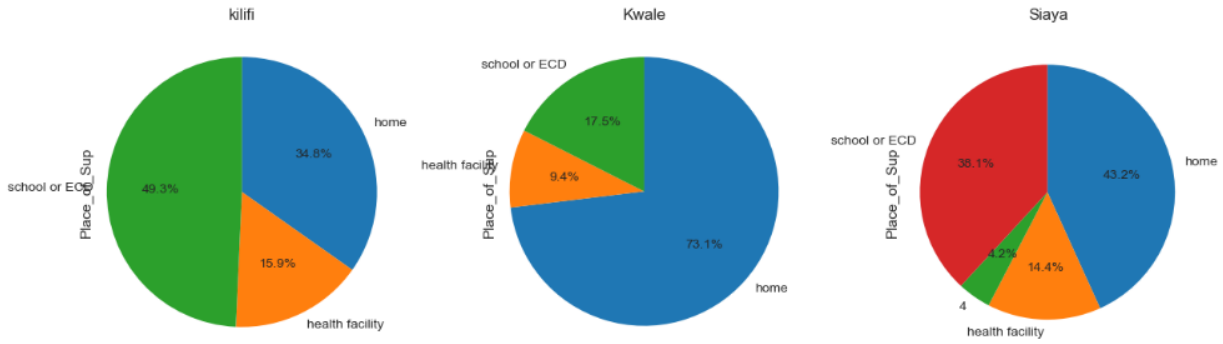

Figure 6: Place of supplementation for Kilifi, Kwale, and Siaya

Income

Kilifi records the highest rate of income with the majority of the caregivers earning between 0 and Ksh.15,000 ($150) per month followed by Siaya where the majority of the caregivers earn between 0 and Ksh. 6,000 ($60) per month and Kwale with the majority earning between 0 and Ksh.3,000 per month. This can be explained by the fact that compared to the other three counties Kilifi is an industrial and service economy with hotels and retail business as some of the significant economic activities.

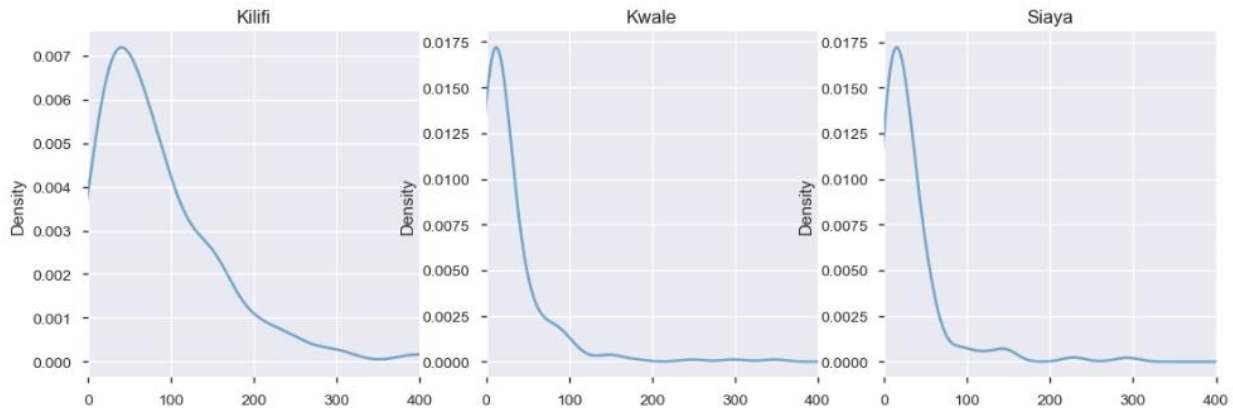

Figure 7: Income distribution for Kilifi, Kwale, and Siaya

Employment

The rate of employment is low in all the three counties with Kilifi recording the highest number of employed caregivers at 11.7% followed by Siaya at 8.9% and Kwale at 5.0%. The high rate of employment in Kilifi against other counties can be explained by its economic status described above. Notably, the statistics also show that a significant number of caregivers are engaged in small businesses (self-employment) with Siaya recording the highest at 44.9% followed by Kilifi at 36.8% and Kwale at 21.4%.

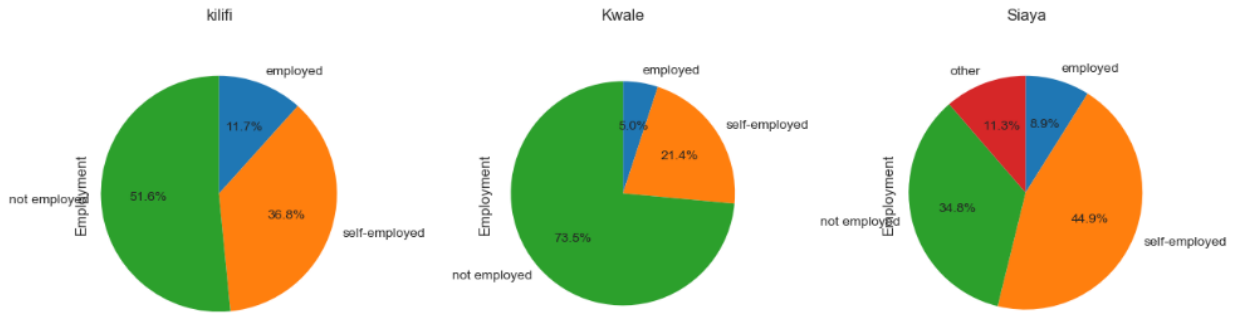

Figure 8: distribution of employment status in Kilifi, Kwale, and Siaya

Education

The highest level of education for the caregivers varies from one county to another as shown in the figure below. **Kilif**i: did not attend school=11.2%, nursery=38.1, primary=33.2% secondary=12.6%, tertiary (certificate & diploma) =4.5, university=0.4. **Kwale**: did not attend school=24.4%, nursery=0.4, primary=55.9% secondary=14.3%, tertiary (certificate & diploma)

=4.6, university=0.4. **Siaya**: did not attend school=0.8%, nursery=24.3, primary=50.6%, secondary=17.9%, tertiary (certificate & diploma) =6.4. Interestingly, Kwale has the highest rate of illiteracy in terms of the caregivers who did not attend school at all but it also records the highest number of supplemented children. This demonstrates that the level of education may not be a significant determinant for Vitamin A supplementation.

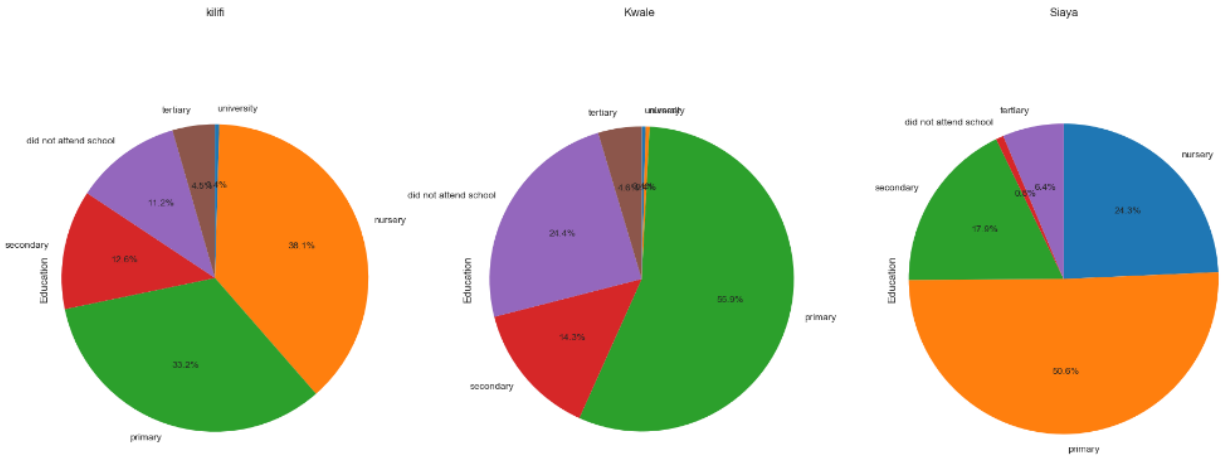

Figure 9: Level of education distribution for Kilifi, Kwale, and Siaya

ECD Survey data

ECD survey data was used as a complementary source of analysis to help understand some of the factors that influenced the Vitamin A supplementation process. The dataset consisted of a varying number of data points from one county to another representing the number of ECDs interviewed in each county. Kilifi county dataset was made up of 55 observations, Siaya 32 observations, and Kwale 16 observations. The datasets were considered import for the study because they provided insights to some of the reasons why some of the school-going children were not supplemented. The dataset contained 15 variables representing the questions but the key variables for the study included: the total number of children in the ECD, Number of children supplemented, the reasons for some children not being supplemented, and the communication instances. We also added two features: the number of communication instances and the percentage of supplemented children to help examine the relationship between communication and rate of supplementation. The number of communication instances was obtained by counting

the number of unique cases where communication took place between the community health volunteers (CHVs), the ECD and the parents. For example, a case where the CHV communicated to the ECD alone would be considered as one communication instance whereas a case where communication was made to the ECD and the area chief was considered as two communication instances.

Reasons for missing supplementation

Reason for missing supplementation was a multiple answer question but the study prioritized the first answer for analysis. The reasons included absenteeism of the child, cultural or religious beliefs, overage (the child was above 5 years), and if the child had already been supplemented. As shown in the figure below the major reasons why school-going children were not supplemented in Kilifi were Overage=47.2%, Absenteeism=47.2%, religious/Cultural belief=3.7%, and those already supplemented=1.9% respectively. In Kwale the reasons were Absenteeism=37.5%, Overage=31.2%, and those already supplemented=31.2% respectively. In Siaya Absenteeism=48.7%, Overage=28.2%, No reason=15.4, those already supplemented=5.1%, religious/Cultural belief=2.6% and respectively.

Noticeably there is a higher percentage of children who had already been supplemented in Kwale (31.2%) compared to other counties (Kilifi=1.9, Siaya=5.1%). This can be explained by the fact that the majority of children in Kwale county were supplemented at home as demonstrated earlier. “No reason” answers in Siaya were given by ECDs where all children were supplemented. Religious and cultural beliefs also play a significant role in determining the administration of Vitamin A supplements, especially in Kilifi and Siaya counties. As demonstrated in figure 10-b below, the majority of Catholics and protestant children were not supplemented. This is because there are a few churches opposed to medical treatment and thus advice their members against the supplementation campaigns. There is a considerable percentage of absenteeism of children on the day of supplementation which could be investigated further. Our hypothesis could be that some of the parents were aware of the supplementation and intentionally kept children at home.

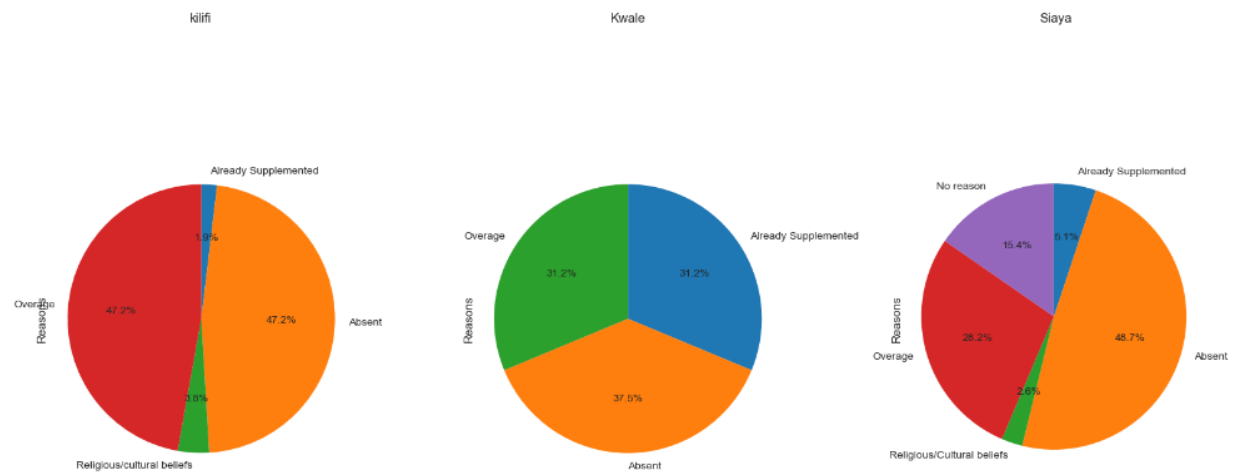

Figure 10: Reasons for missing supplementation in Kilifi, Kwale, and Siaya

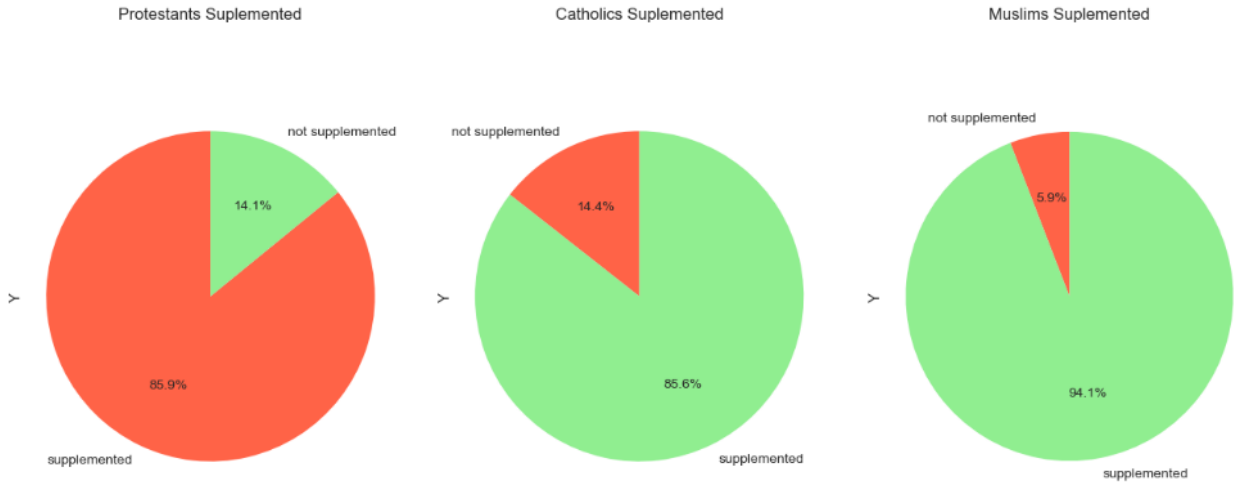

Figure 10-b Rate of supplementation based on religion

Communication instances against supplementation percentage

As shown in the scatter plot below the rate of supplementation increases with an increase in the number of communication instances. In most cases where there were two communication instances, the Community Health Volunteer (CHV) informed the ECD prior to the exercise through a phone call and then followed up by visiting the ECD to remind them. On the other hand, one communication instance cases involved the CHV calling the ECD head teacher then followed by the exercise. This implies that supplementation would have been higher had there been effective communication and coordination among the involved parties.

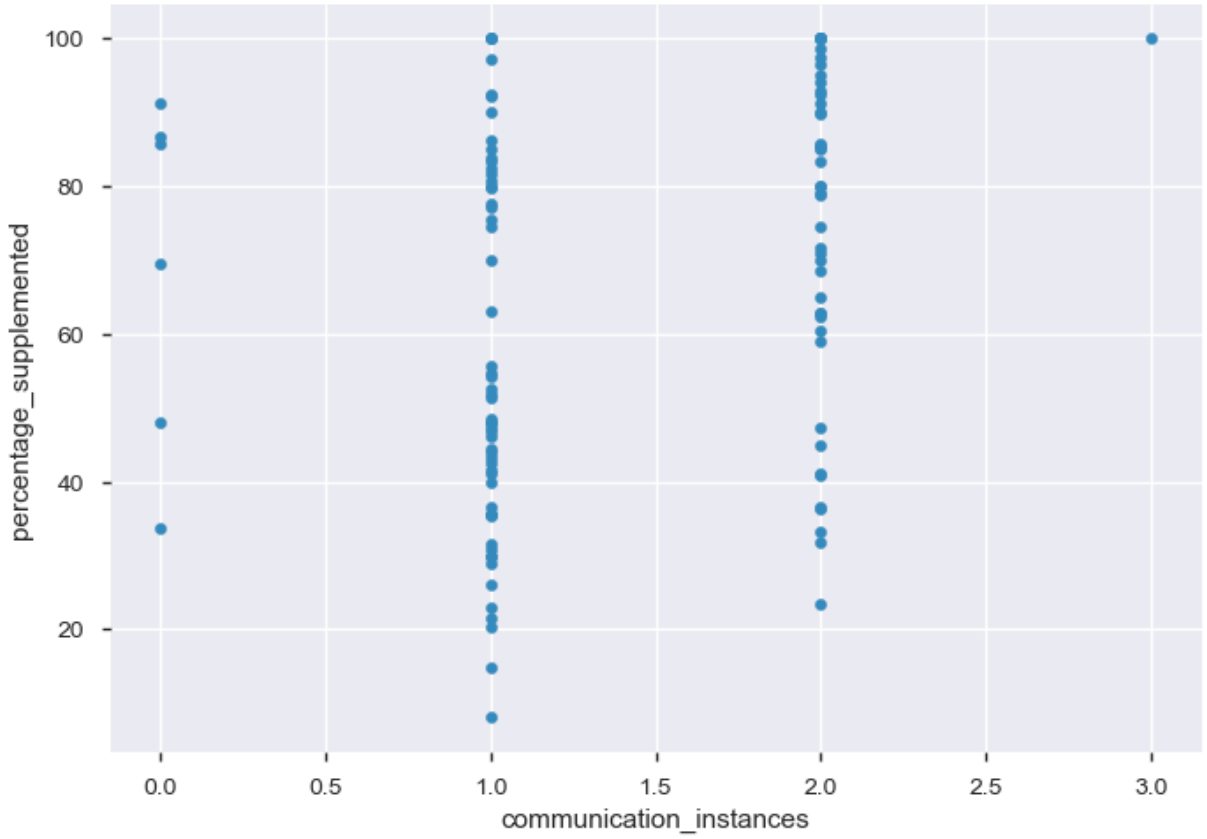

Figure 11: Relationship between communication instances and supplementation in the three counties

Analysis of the Aggregate data

Correlations

Our data is largely categorical so we used Spearman correlation to investigate any monotonic relationships but we also expect some degree of linearity so Pearson relationship will be useful to analyze linearity and constant variance. As shown in the figure below there are significant correlations among the variables. Both Spearman and Pearson correlation coefficients indicated that there is a strong positive correlation between the child age and the place of supplementation. The places of supplementation included: 1=health facility. 2=home and 3=Schools or ECD. This means that an increase in child age results in a proportional increase in the place of supplementation from 1 to 3. This is because the majority of the caregivers with young children

frequented the health facility for a clinical checkup of the infants and thus received Vitamin A supplementation in the process.

The figure below also shows a strong negative correlation between the caregiver's age and group membership. The group membership variable determines whether the caregiver belongs to a women savings and support groups popularly known as Chama: 1=belonged to a group and 2=did not belong to a group. The common groups included Merry-go-round groups, savings and loans groups, mother support groups and Farmer groups. This means that many older caregivers were more likely to join the groups compared to the younger ones. It's mainly because older caregivers or mothers are well informed and understand the benefits of the groups.

There is also a strong negative correlation between group membership and the level of education of the caregiver. This means that the higher the education level the more likely the caregiver is to join a group. Similarly, the more educated the caregiver the more informed she is about the benefits of the groups hence the higher likelihood to join.

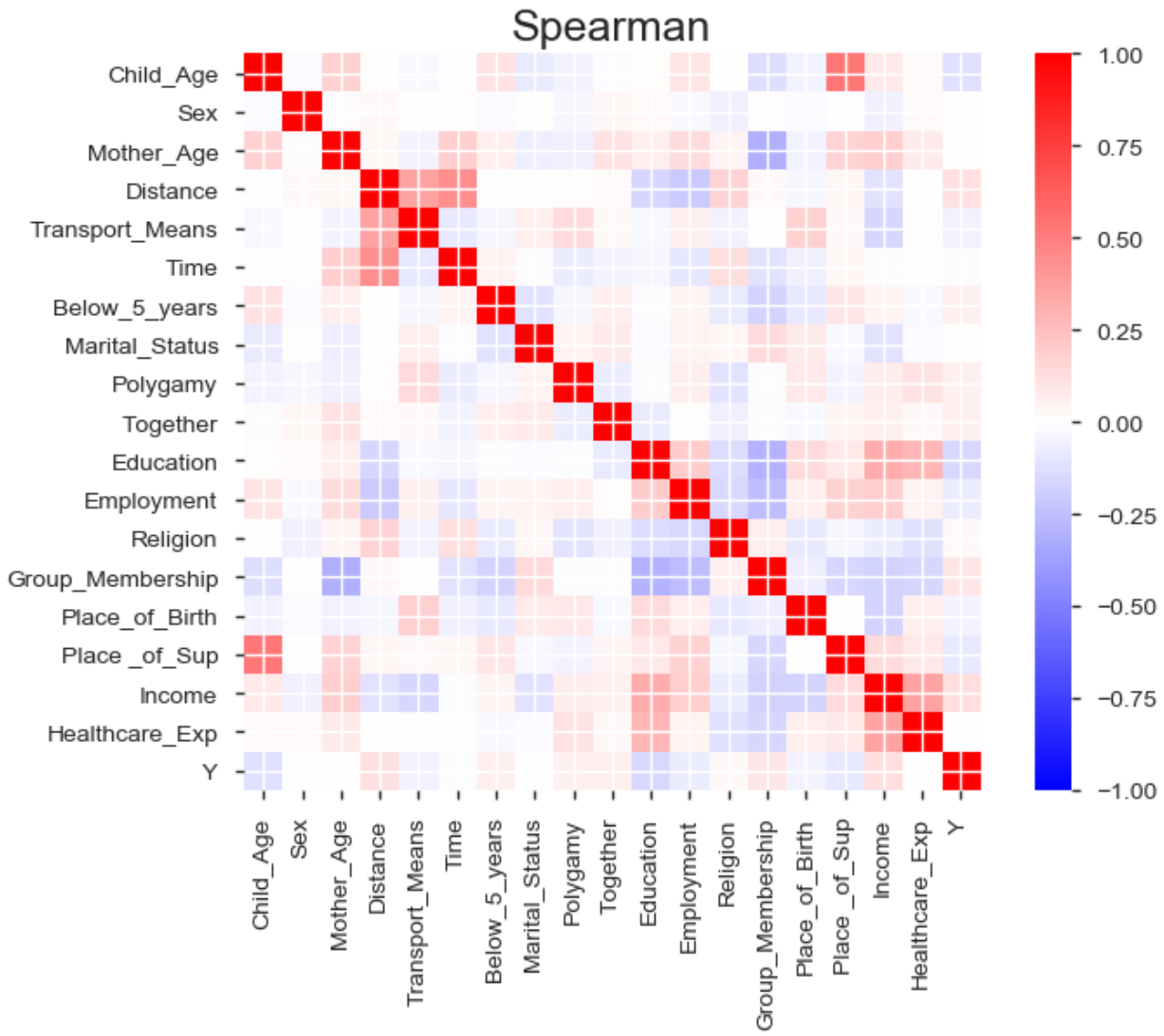

Figure 12: Spearman correlation

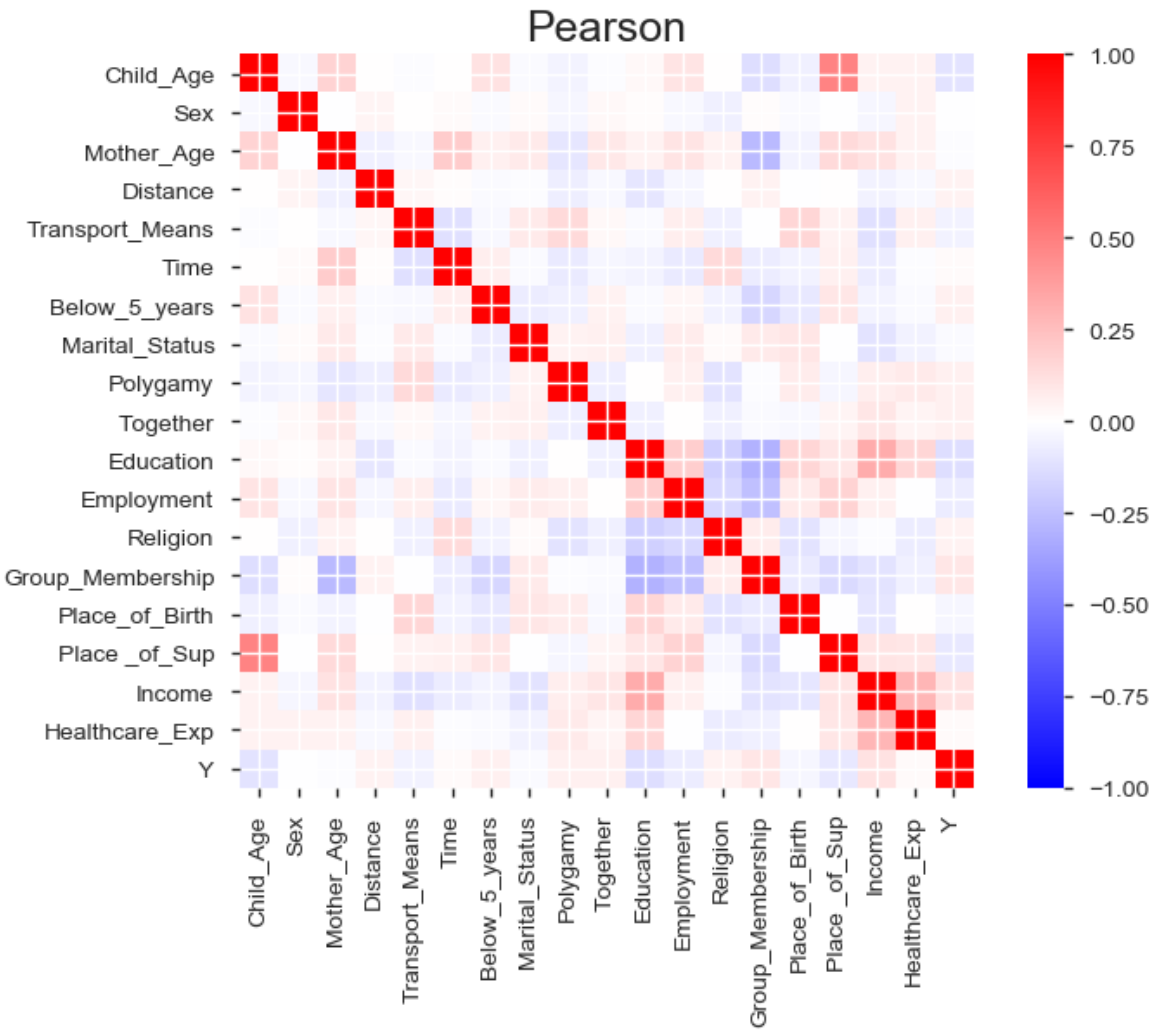

Figure 13: Pearson correlation

Data Preprocessing

Data preparation

The data was largely categorical so dummy (pandas get dummies) encoding was used to create instances of the categories to improve the performance of our model. The numerical data was not normally distributed with some noticeable outliers hence the use of Robust Scaler for data standardization. Robust Scaler utilizes the interquartile range for feature scaling which makes it effective for dealing with outliers. Moreover, this study will use the Support Vector Machine algorithm which relies on Euclidean distance hence sensitive to magnitudes. Therefore, scaling the features will ensure that they weigh in equally to improve model performance.

Feature importance

Feature importance is simply an increase in the model error after the feature values are permuted. In essence, a feature is important if a change in its values increases the model's error which shows that the feature was an important determinant of the prediction outcome. On the other hand, a feature is unimportant if the model error remains unchanged after changing its values. Important features in our study represent the key determinants for Vitamin A supplementation. Using random Forest Classifier to examine feature importance the results are as shown in the figure below. Place of Supplementation is the most important feature followed by Caregiver's income, Distance to the health facility, Child Age, Healthcare Expenditure and Time taken to reach the health facility respectively. As illustrated in the supplementation pie chart majority of supplementations took place at home. This implies that focusing the supplementation campaign efforts in homes could yield better results in terms of achieving a higher rate of supplementation.

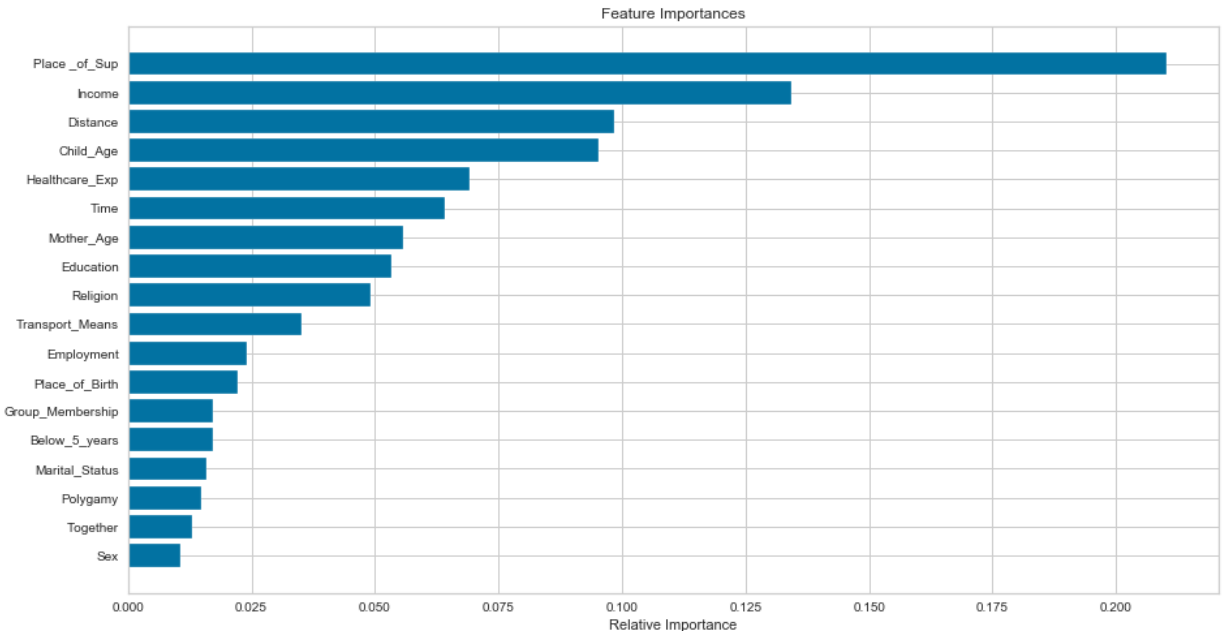

Figure 14: Feature importance with random forest

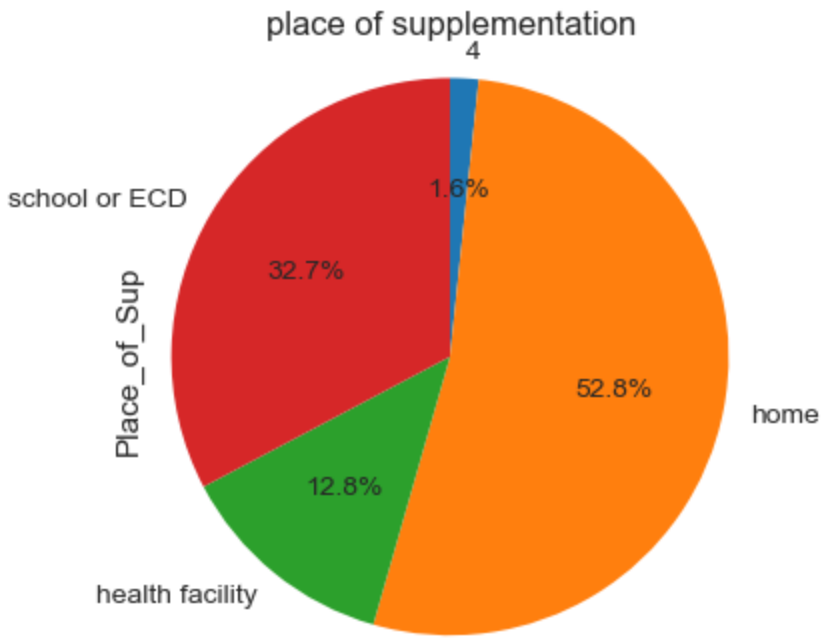

Figure 15: Place of supplementation

Model training

Model training is the process where machine learning algorithms learn to predict data by providing them with training data to learn from. The algorithm identifies patterns in the training data to enable it to map the input data to the target attribute. The trained model can then be used to make predictions on new data whose target attribute is unknown. In this case, the machine learning algorithm learns from the caregiver's data to predict whether a child received vitamin A supplement or not.

This process requires splitting the dataset into two sets; one for training the model and the other set for testing or evaluating the performance of the model or its ability to predict on unseen data. The data was split in the ratio of 8:2 that is 80% was used for training and the remaining 20% of the original data for model evaluation. Different classification algorithms were then used to perform supervised learning on the dataset.

Four supervised learning algorithms were selected for the study including Naïve Bayes, Support Vector Machine, Random Forest and XGboost. Naïve Bayes was selected due to its sheer simplicity and could be helpful for comparison purposes. Support vector machine was selected for its ability to model non-linear decision boundaries and robustness against over-fitting.

Random Forest was selected for its ability to handle data with high dimensionality. Finally, XGboost was selected for improved performance and memory efficiency.

Here is a short description of the classifiers.

Naïve Bayes: Naïve Bayes is probabilistic classifier based on conditional probability and counting. It operates under an assumption that all features are independent of each other. It simply creates a probability table using Bayes Theorem where predictions are the class probabilities of an observation.

Support Vector Machine: Support vector machine essentially uses kernels that represent data as points in space and calculate the distance between two observations. It then creates a decision boundary that separates the closest observations of separate classes as wide as possible. It builds on logistic regression by introducing non-linear kernels.

Random Forest: Random Forest is an ensemble classifier that uses votes from multiple trees to classify objects. It minimizes over-fitting by utilizing votes from all the trees to classify an object.

XGboost: XGBoost is a tree classifier from the ensemble family with improved speed and performance. It allows for parallelization of tree construction which improves the speed of modeling. It applies a boosting technique which combines weak learners to improve prediction accuracy.

Model evaluation

Receiver Operating Characteristic (ROC) curve is a performance metric primarily for evaluating the performance of binary classification models. A 1.0 represents a perfect model that made all predictions correct while a 0.5 represents a random classifier. It consists of two elements; sensitivity and specificity. Sensitivity is the true positive rate which represents the correctly predictive positive class while specificity is the true negative rate representing correctly classified instances of the negative class.

Precision: This is the ratio of correctly predicted positive observations to the total predicted positive observations. For example, in our case, if all children predicted as supplemented were actually supplemented then the model would achieve perfect precision. However, this approach

does not demonstrate how well our model is at identifying all members of the supplemented class.

Recall: Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It is a measure of how well the algorithm is at identifying all the data points or members belonging to the positive class. For example, in our case, if the model predicted that all children were supplemented then it would have a perfect recall. However, the model would not have a good precision unless all children were supplemented which is not the case in this study.

Selecting a perfect model thus results in a tradeoff between precision and recall. In this study, a high precision (lower false positive rate) is more ideal than high recall (lower false negative rate)

Accuracy: Classification accuracy is a measure of correct predictions as a ratio of all predictions. The problem with classification accuracy is that it tends to discriminate against a class with fewer observations in cases where the classes are not balanced such as in this study (Supplemented=83.5% and not supplemented=16.5%). Moreover, it assumes that the prediction errors are equally important which is not true in our case. For example, in this case, a false negative (children classified as not supplemented but were actually supplemented) would be more desirable than a false positive (children classified as supplemented but were actually not supplemented). The study combines ROC curve, precision, recall, and accuracy to evaluate performance of the algorithms.

Model evaluation results

Classifier	ROC (%)	Accuracy (%)	Precision	Recall	F1-score
Random Forest	79	83	1=0.84 2=0.50	1=0.98 2=0.08	1=0.91 2=0.14
XGBoost	87	85	1=0.90 2=0.58	1=0.93 2=0.46	1=0.91 2=0.51
Support Vector Machine	83	76	1=0.96 2=0.40	1=0.75 2=0.83	1=0.84 2=0.54
Naïve Bayes	72	51	1=0.96 2=0.24	1=0.42 2=0.92	1=0.59 2=0.38

As demonstrated in the table above XGBoost classifier generalizes well followed by Random forest, support vector machine and Naïve Bates respectively. However, all the algorithms perform well at classifying one class than the other. The algorithms are good at identifying the "supplemented" class than the "not supplemented" class. Therefore, the algorithms would classify most children as having received Vitamin A supplementation even if they have actually not been supplemented hence missing supplementation which is not ideal for this study.

Resampling

In supervised learning, resampling techniques are employed to balance classes in cases where they are imbalanced such as in this study (Supplemented=83.5% and not supplemented=16.5%). This can be achieved through undersampling where the size of majority class is reduced to that of the minority class or oversampling which involves generating synthetic instances of the minority class to match the size of the majority class. The oversampling method was used in this study to improve model performance.

Oversampled classes

Classifier	ROC (%)	Accuracy (%)	Precision	Recall	F1-score
Random Forest	98	94	1=0.93 2=0.95	1=0.95 2=0.94	1=0.94 2=0.94
XGBoost	92	87	1=0.87 2=0.87	1=0.86 2=0.88	1=0.86 2=0.87
Support Vector Machine	87	87	1=0.85 2=0.78	1=0.73 2=0.88	1=0.79 2=0.83
Naïve Bayes	84	74	1=1.00 2=0.66	1=0.46 2=1.00	1=0.63 2=0.80

As illustrated in the table, Random forest generalizes well as compared to other algorithms with ROC and accuracy of 0.98, and 0.94, respectively. XG Boost classifier also performs relatively well for oversampled classes with ROC and accuracy of 92, and 87 followed by Support Vector Machine and Naïve Bayes respectively. As mentioned earlier, a case where children are classified as not supplemented but were actually supplemented would be more desirable than when children are classified as supplemented but were actually not supplemented, ideally no child should miss the supplementation. Therefore, Random Forest becomes a perfect model for the study because it generalizes well with higher precision and recall. A precision measure of 1=0.93 means that 93% of those classified as supplemented were actually supplemented. This is a considerably better performance for ensuring that only a few children would miss the Vitamin A supplementation.

Findings and recommendations

Descriptive and predictive analysis of the data using supervised machine learning revealed the following about Vitamin A supplementation in Siaya, Kwale, and Kilifi:

- Place of supplementation is one of the major factors affecting the rate of Vitamin A supplementation. This is partly because of the long distance between the caregiver's

home and the nearest health facility. According to the analysis majority of the supplementations took place at home which involved the Community Health Workers moving from one home to another administering the supplements. Moreover, Majority of the caregivers are not employed and therefore lack enough capital to for child healthcare services including buying supplements. Therefore more efforts should be directed towards ensuring that there are enough human labor and resources to reach caregivers at their homes.

- The more informed the caregivers are about the benefits of the Vitamin A supplementation the more likely they are to participate in the exercise. Therefore, there is need for increased civic education especially for young caregivers or mothers about Vitamin and its benefits.
- Communication is a key determinant of Vitamin A supplementation. Caregivers should be well informed about the campaigns through strategic community mobilization efforts.
- Religious and cultural beliefs hinder the administration of Vitamin A supplementation hence the need for campaigns against any unfounded religious and cultural beliefs.

Conclusion

The aim of this study was to determine the factors affecting Vitamin A supplementation in three counties Kwale, Kilifi and Siaya and how they can be minimized to increase the impact of the supplementation campaigns. As demonstrated in the analysis, the government and all the stakeholders need to address factors such inadequacy of enough resources for the exercise, religious or cultural beliefs against the exercise, communication and community mobilization efforts towards the exercise.

The study had notable limitations that impacted the outcome of the model. Its performance could be improved by having a definite figure representing Vitamin A supplementation as the independent variable. What was available for the study was an assessment for supplementation within the past six months which did not accurately reflect the variable. (The independent variable question was: *Within the last six months, was [NAME] in the household] given a vitamin A dose like (This/any of these)?).*

About Me

I have worked collaboratively on various research projects around health and social development in Kenya and currently keen on finding ways to augment healthcare services using AI and

Machine Learning. I am also volunteering at Ai Kenya as a content manager and welcome anyone interested in publishing their research papers on the website or providing data for research and training. Contact me if interested in any kind of collaboration in similar or related projects either through data sharing, modeling, or analysis.

Email: eugene@kenya.ai