Exploring and Visualising Patterns in 300'000 Consultations Collected with the MSF Clinical Decision Support Algorithm, eCARE

Abdulkadir Gokce Gunes Basak Ozgun Edin Guso

Abstract

Infectious and parasitic diseases are difficult to distinguish when the medical equipment is limited. Late detection of these diseases may cause an outbreak and endanger many lives. In collaboration with iGH lab and MSF(Médecins Sans Frontières), we built an unsupervised machine learning model to cluster the patient data to help detect potential outbreaks and their characteristics and a Power BI dashboard for visualisation of data and the model.

1. Introduction

In WHO(World Health Organization) Africa region, 2019, 34% of recorded deaths among children between 0-59 months old were caused by infectious and parasitic diseases(WHO, 2020). Limited medical resources make infectious and parasitic diseases hard to diagnose. This may prevent early detection of these diseases, thus lead to potential public health risks and increased morbidity rates.

MSF uses Power BI tool to analyse and visualise clinical data collected in Sub-Saharan Africa projects. The data collected is from consultations of children with infection suspicions. The dashboard plays a crucial role as a visualisation layer that makes early detection possible. With early detection, it is possible to predict and act upon potential outbreaks. We will use unsupervised machine learning to create spatio-temporal clusters from the consultation data, which will allow us to recognise patterns in patients, see important features that hint the diseases.

Additionally, we created a Power BI dashboard to visualise these clusters and other useful insights into the dataset.

2. Dataset

The dataset consists of 310.000 consultations between the dates 21 November 2016 and 25 September 2021, and 271 features that collected by MSF clinical decision support algorithm, eCare. The patients were 0 to 60 months old. The features represent demographic data, diagnosis, lab test results, patient's past medical history,data related to patients outcome, signs, vital signs, symptoms, recommended treatments and prescriptions.

3. Methodology

3.1. Data Preprocessing & Feature Engineering

We developed two data preprocessing pipelines according to two ML tasks. For clustering, we only keep demographic information (numerical values) and a new list of features derived by collecting related signs and symptoms in groups. A value of such a feature is 1 if any of the member features of the group is 1. In this operation, we treat `NaN` entries as 0's in order to get binary labels for each new feature. Although this might generate bias, it is usually the case that a value is `NaN` when a patient does not that show that sign or symptom. Moreover, with binary categorical features we don't need to use one-hot encoding for clustering, which enables faster computation and more meaningful cluster centroids. Grouping of features also boosts the interpretability of the cluster representatives as most of the categorical features have the value 0 more than the 99% of the time. We use median imputation for numerical features as it is more robust to outliers than mean imputation and normalize them to 0 mean and unit variance. The final dataset for clustering has 32 features.

For predictive analysis, we don't perform the grouping operation and use one-hot encoding for categorical features hence each sign or symptom is mapped to 3 dimensional vector representing 0, 1, or `NaN`. We also include date information of each data sample with the sine and the cosine of week and month numbers to encode cyclic nature of time. Furthermore, we map the health facility numbers to cities they are located in and add longitude & latitude of information as features using `geopy` (Kostya, 2014–2021) Python package. As all data samples have the spatiotemporal information we don't use imputation on them. Finally, we normalize all numerical (demographic and spatiotemporal) features to 0 mean and unit variance. The final dataset for predictive analysis has 200 features.

3.2. Clustering

Clustering is an unsupervised machine learning method which partitions unlabeled data into groups of data samples that are similar to each other with respect to a predefined distance metric. Most of the clustering techniques are tailored for a single-type data modality, i.e. K-Means for numerical data and K-Modes for categorical data. However, many real-world datasets comprised of both kinds of data, which

is also the case for our dataset. Although one can utilize algorithms for single data type on mixed-type datasets by discretizing continuous values or one-hot encoding of categorical features, other issues arises with this approach (Foss et al., 2019). On the other hand, mixed-type data clustering is an ongoing research topic (Ahmad & Khan, 2019) and most of such techniques have not been yet implemented in Python's libraries.

For clustering, we used `kmodes`(de Vos, 2015–2021) Python library to implement K-Prototypes algorithm and `scikit-learn`(Pedregosa et al., 2011) package to test Gaussian Mixture Models.

3.2.1. K-Prototypes

K-Prototypes clustering algorithm is an extension of K-Means and K-Modes techniques for mixed-type datasets (Huang, 1997)(Huang, 1998). For a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ of N samples where each $\mathbf{x}_i$ is a D-dimensional vector with both D_1-many continuous and D_2-many categorical values ($D = D_1 + D_2$), K-Prototypes clustering amounts to minimize the loss

$$\min_{\mathbf{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} d(\mathbf{x}_n, \boldsymbol{\mu}_k) \tag{1}$$

$$\text{s.t. } \boldsymbol{\mu}_k \in \mathbb{R}^{D_1} \times \{0,1\}^{D_2},\ z_{nk} =\in \{0,1\},\ \sum_{k=1}^{K} z_{nk} = 1$$

where $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]^\top$ is the cluster assignments matrix ($\mathbf{z}_n = [z_{n1}, \dots, z_{nK}]^\top$) and $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K]^\top$ is the matrix of representative (prototype) points.

One crucial distinction between K-Prototypes and K-Means & K-Modes is that the distance metric $d(\cdot, \cdot)$ should accommodate both continuous and discrete variables. As such, it can be decomposed into two parts $d(\mathbf{x}_n, \boldsymbol{\mu}_k) = d_c(\mathbf{x}_n, \boldsymbol{\mu}_k) + \lambda_k d_d(\mathbf{x}_n, \boldsymbol{\mu}_k)$ where d_c and d_d are distances metrics defined on the continuous and the discrete portions of the features while λ_k is the parameter determining the proportional weight of these two distances. If we select d_c as the Euclidean distance and d_d as the Hamming distance, the composite distance becomes

$$d(\mathbf{x}_n, \boldsymbol{\mu}_k) = \sum_{d=1}^{D_1} (x_n d - \mu_k d)^2 + \lambda_k \sum_{d=D_1+1}^{D} |x_n d - \mu_k d|$$

Minimization of K-Prototypes follows alternating optimization iterations similar to K-Means: fix $\boldsymbol{\mu}$, compute $\mathbf{z}$; fix $\mathbf{z}$, compute $\boldsymbol{\mu}$. However, the computation of $\boldsymbol{\mu}_k$ is a bit tricky since some the entries of $\boldsymbol{\mu}_k$ are discrete variable. In this case, those entries are selected as the mode of the feature in a given cluster. In another words, if for cluster k and the discrete feature d the most frequent label is 1, then assign 1 to the d-th entry of the k-th representative. The continuous variables of $\boldsymbol{\mu}_k$ are still calculated as the average of the data samples in a given cluster. The algorithm iterates until assignments does not change and converges to a local minimum. Hence, we initialize the algorithm 5 times and keep the instance with the minimum loss.

3.2.2. Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a probabilistic method describing a set of points as samples from a weighted combination of a finite number of Gaussian distributions. Compared to K-Means, a GMM facilitates elliptical clusters and allows data points to be sampled from a multinomial distribution, which leads to soft-clustering. Nevertheless, GMMs are designed to model discrete variables as the distribution itself is continuous. Furthermore, the means of Gaussian distributions are not enforced to be discrete so the reprentatives of a given mixed-type dataset will not be of the same form as the dataset. Regardless, we can interpret the discrete labels as continuous variables and apply a GMM model on a mixed-type dataset. Let $\boldsymbol{\pi} = [\pi_1, \dots, \pi_k]^\top$ be the cluster weights $p(z_n = k) = \pi_k$, μ_k and Σ_k be the mean vector and the covariance matrix of kth cluster, respectively.

Then, the model aims to maximize the log-marginal likelihood

$$\max_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

Eq. 2 is unbounded and not concave. A non-unique solution for Eq. 2 can be found using Expectation-Maximization (EM) algorithm(Wikipedia, 2021), which iteratively compute assignments and optimizes parameters in an alternating fashion.

3.3. Selecting the Number of Clusters and Clustering Method

As there is no optimal solution for choosing the number of cluster points, we compute the cost of K-Prototypes 10 times as a function of number of clusters, as can be seen in Figure 1 . With Elbow method, we decided on using 8 clusters for K-Prototypes algorithm. We also calculated Bayesian Information Criterion for GMM instances, which penalizes high number of parameters in models. Regardless, since GMM treats categorical labels as continuous variables and the cluster representatives does not necessarily have discrete values in categorical entries, we opted for using K-Prototypes as our clustering method.

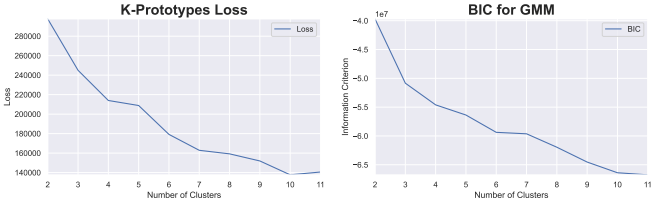

Figure 1. left K-Prototypes clustering loss a function number of clusters, *right* Bayesian Information Criterion for GMM as a number of clusters

3.4. Predictive Analysis

To understand which features impact diagnoses the most, we trained binary linear classifiers (regularized logistic regression and linear SVM) for each diagnosis to predict diagnoses of patients given demographic and spatiotemporal features as well as signs and symptoms. We utilized recursive feature elimination (RFE) method to identify top k most discriminant features to predict a given diagnosis. Given a linear classifier, the RFE method recursively selects a subset of features by pruning the weights of the classifier with lowest norms. To accomplish that, the model is repeated trained and pruned until the desired of number of features is left. We then compared F1-scores of classifiers before and after the pruning to determine how much information is captured by the remaining features for a given diagnosis.

4. Results and Visualisation

4.1. Predictive Analysis

With recursive feature elimination method, we chose the most important 4 features for each class and compared the F1-score of classifiers utilizing all features and with top 4 features for a given diagnosis. A subset of these results are shown in Table 1. We observed that if a diagnosis is highly correlated with one or two of the signs or symptoms, only using 4 features is sufficient to predict the illness of a patient. For example, `hisdx_ot_mouth` can be predicted with only `mouth_trush_0`, `mouth_trush_1`, `s_mouthpb_0`, and `s_mouthpb_1` features without any drop in F1-score. As such, this diagnosis can inferred solely based on `mouth_trush` and `s_mouthpb` symptoms of a patient. On the other hand, if a diagnosis has a lower base rate of occurance and is not highly correlated with any signs or symptoms, using only top 4 features completely breaks the classifiers. As example, F1-score of the classifier for `hisdx_malaria_simple` undergoes a subtle decline whereas `hisdx_malaria_sev` cannot be predicted correctly at all (in terms of F1-score), which has a base rate of 0.7% (`hisdx_malaria_simple` has base rate 19.7%).

Table 1. F1-scores of binary classifiers with all and top 4 features.

DIAGNOSIS	FULL MODEL	PRUNED MODEL
HISDX_OT_MOUTH	0.975	0.975
HISDX_LRTI	0.869	0.834
HISDX_COUGH_PERSIST	0.571	0.0
HISDX_DIARRHWATERY	0.995	0.992
HISDX_MALARIA_SIMPLE	0.987	0.985
HISDX_MALARIA_SEV	0.673	0.0

4.2. PowerBI

Power BI is a data analysis tool developed by Microsoft to create interactive, immersive dashboards and reports that provide actionable insights and drive impactful results.

It is being used by administrators in MSF to track health care workers and consultation insights such as the duration, patient information, symptoms, tests, and diagnosis.

There are 271 features in the dataset with more than 300.000 entries and counting, thus it is a growing challenge to understand the critical relationship between each feature and their predictive power. One of our main goals in this project was to create a simple yet insightful dashboard for administrators to monitor overall performance and go into detail by filtering if need be.

For the dashboard, we created a filter page, overview page and 6 other pages focusing on specific features. See Appendix for captures of each page.

4.2.1. OVERVIEW

This page is designed to be the home page of the application. Users can navigate to 6 other data related pages and a filter page. There is a Q&A tool where a simple question can be queried, and a plot or numerical answer shows up. On the page, a line chart with number of consultations in y axis and date information in x axis is present. For quick summary total number of consultations and the last update date of the table is shown. (Figure 2)

4.2.2. FILTERING

Filter page can be accessed by pressing the "i" button. There are 6 main filter parameters, project name, health facility code, health worker code, date, gender, and age of patients. The available values of each parameter change as the data and the filtering changes. As the user chooses filters in this page, all changes are propagated to other pages simultaneously. Users also have the option to clear all filters with a single button on every page. This change also propagates to all pages. (Figure 3)

4.2.3. DEMOGRAPHIC INFORMATION

This page focuses on the patient information. For the age feature, statistical information (mean, median, standard deviation, max and min values) is displayed. Weight for Age feature is very important to monitor nutritional information of children. We used a bar graph for this feature. Also, to display the gender distribution, we used a pie chart. (Figure 4)

4.2.4. PROJECT INFORMATION

In this page, the data is organised by the project name attribute which contains the location information semantically. Thus, we used the map tool in Power BI to show each project and their impact in terms of consultations. It is possible to zoom in and out of the map to explore data deeper.

We also wanted to look deeper into the size of each project, thus three bar charts created for number of health facilities, number of health workers and median of consultation duration per project. Median of consultation duration is especially important as MSF is interested in analysing which

regions use the tool properly. (Figure 5)

4.2.5. Health Facility Information

For each facility, number of consultations were shown. However, one may conclude that having a lot of consultations mean that the health facility is working well. Thus, we also show the number of consultations divided by health worker in each facility. With this information in hand, the previous statement does not always hold true.

A tool developed by MSF proposes treatments to the health workers when they enter the symptoms which is called "Proposed Treatment". The health worker has the liberty to follow the proposed treatment or create his/her own. It was important for MSF to observe the ratio of agreeing and disagreeing with the proposed treatment. To show the contrast between agreements and disagreements we used 100% Stacked Column graph. However, having a high disagreement percentage does not mean that the health worker is bad. There might be an unusual disease, or the worker might suspect from something else, and MSF is interested in following up with such cases. (Figure 6)

4.2.6. Antibiotic

Antibiotic consumption is a very sensitive issue, over consumption may lead to antibiotic immunity which makes the medicine less effective. Thus, we wanted to create a separate page for antibiotic prescription data to track the usage. The main relationship we wanted to observe was the number of times antibiotic was suggested in the proposed treatment and the actual number of antibiotics prescribed by the health worker.

In a large bar chart, the antibiotic prescription ratio (prescribed / recommended) by each health facility is shown. We also provided special colouring for bars with certain ratio numbers, to make the graph more readable. If the ratio is larger than 1.5, the bar is red; if it is larger than 1.25, it is orange; if it is larger than 1.10, it is yellow and else it is green. We also provide recommended and prescribed antibiotic by year to see whether there is a large mismatch between them.

Additionally, a bar chart that displays which diagnosis are noted by the health worker when they disagreed with the recommended diagnosis and prescribed antibiotics. On the y axis, the number of such occurrences are shown. (Figure 7)

4.2.7. Data Analysis and Key Features

This page displays four questions that ask about main features that affect a certain topic. A new page that displays analytics about the given question is opened. (Figure 8)

First, the key influencers tab is displayed. This visualises the most important features that affect the chosen topic. For example, the most effective feature in diagnosis disagreement is ear pain (whether the patient has ear pain symptoms). Our analysis displays that if a patient does not have ear pain, the health worker is 1.64 times more likely to disagree with the proposed treatment, which is significant to say the least. It is possible to click on each feature and further analyse how this feature and its values affect the outcome. (Figures 9 11 13 15)

The second tab, top segments, displays the largest segments (clusters) which deviate from the mean. These are the cases that have certain features and common and their cluster mean is significantly higher than the total mean. For example, for the prescribing antibiotics when none are recommended analysis, there is a segment of 6664 cases (more than 2% of the dataset) which has 4,5 times higher rate of prescribing antibiotics when none are recommended. (Figures 10 12 14 16). It is possible to click on each segment and further analyse which features are common in these clusters. (Figure 17)

4.2.8. Unsupervised Learning: Clustering

In this page, we are displaying the results of our Unsupervised ML model we built. Previously defined filters in the filter page are not applied here since we are using a separate dataset containing the clusters. Only the date filtering is available in this page.

We display each cluster and their sizes with circles on the map. The colours of the circles are based on the cluster they're assigned to. When filtering different time intervals, the colours of the clusters change automatically depending on the time interval selected. There is a table that shows each clusters' features. It is possible to refer to this table to make further conclusions about these clusters and gather important insights on what were the common trends in which regions and time intervals. (Figures 18 19)

5. Discussion and Future Work

We created an easy to use and at same time powerful dashboard using Power BI. It is possible to gain useful insights about the dataset from healthcare facility information, to antibiotic usage. We also applied data analysis tools to understand the relationships between some key features and the rest of the features. Finally, we also applied unsupervised learning techniques to create spatio-temporal clustering. This clustering, when visualised inside our Power BI presentation, allows users to gather information about the dominating trends by region and time (Figure 18).

Our work has showed that it is possible to extract valuable insights from the vast medical record database that MSF has provided and these insights can be used in pattern recognition for early detection of outbreaks.

MSF aims to create three dashboards: for project administrators, for health facility managers and for health workers. In our work, we have fully developed the dashboard that will be used by project administrators. In the future, two more dashboards need to be developed.

References

Ahmad, A. and Khan, S. S. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902, 2019. doi: 10.1109/ACCESS.2019.2903568.

de Vos, N. J. kmodes categorical clustering library. https://github.com/nicodv/kmodes, 2015–2021.

Foss, A. H., Markatou, M., and Ray, B. Distance metrics and clustering methods for mixed-type data. *International Statistical Review*, 87(1): 80–109, 2019. doi: https://doi.org/10.1111/insr.12274. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12274.

Huang, Z. Clustering large data sets with mixed numeric and categorical values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 21–34, 1997.

Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values, 1998.

Kostya, E. geopy is a python client for several popular geocoding web services. https://github.com/geopy/geopy, 2014–2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

WHO. Global health estimates: Leading causes of death, 2020. URL https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death.

Wikipedia. Mixture model, Nov 2021. URL https://en.wikipedia.org/wiki/Mixture_model#Expectation_maximization_(EM).

Appendices

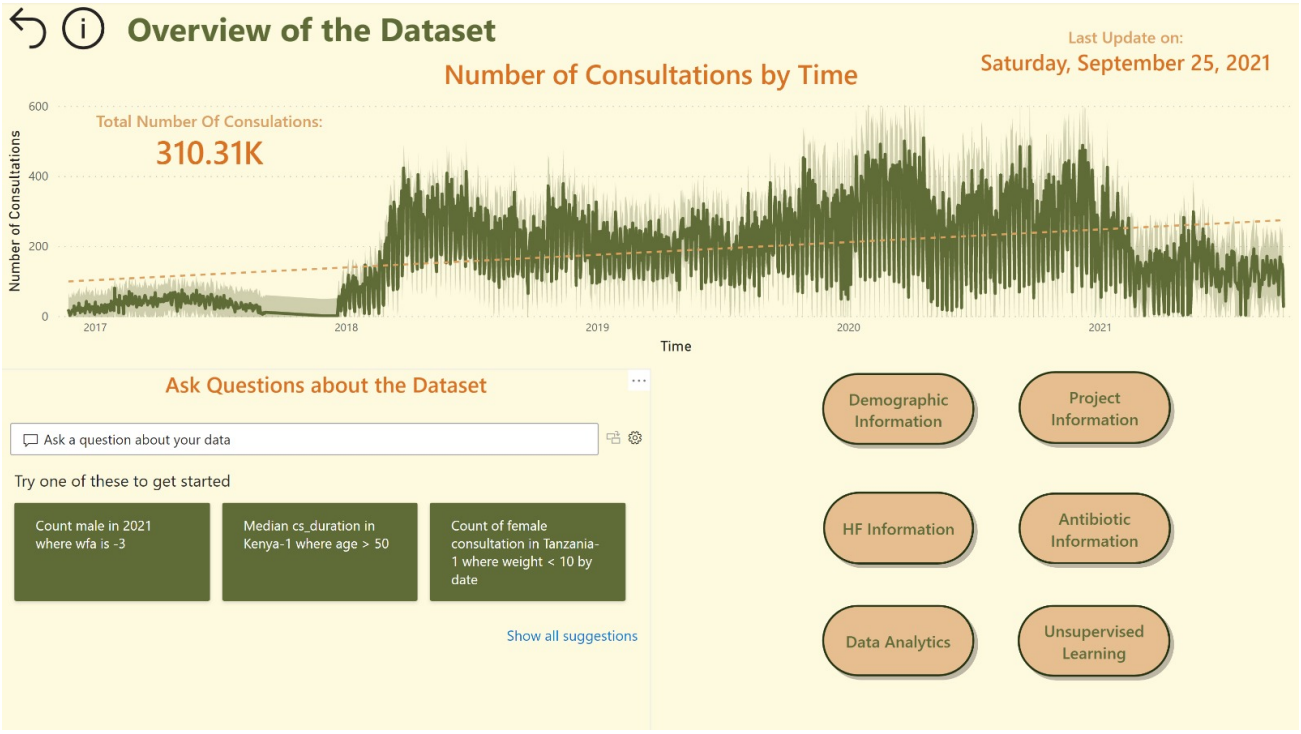

Figure 2. Overview Page: Displays the consultation over time alongside total number of consultations. Provides a powerful Q&A tool. Has buttons for easy access to the rest of the report.

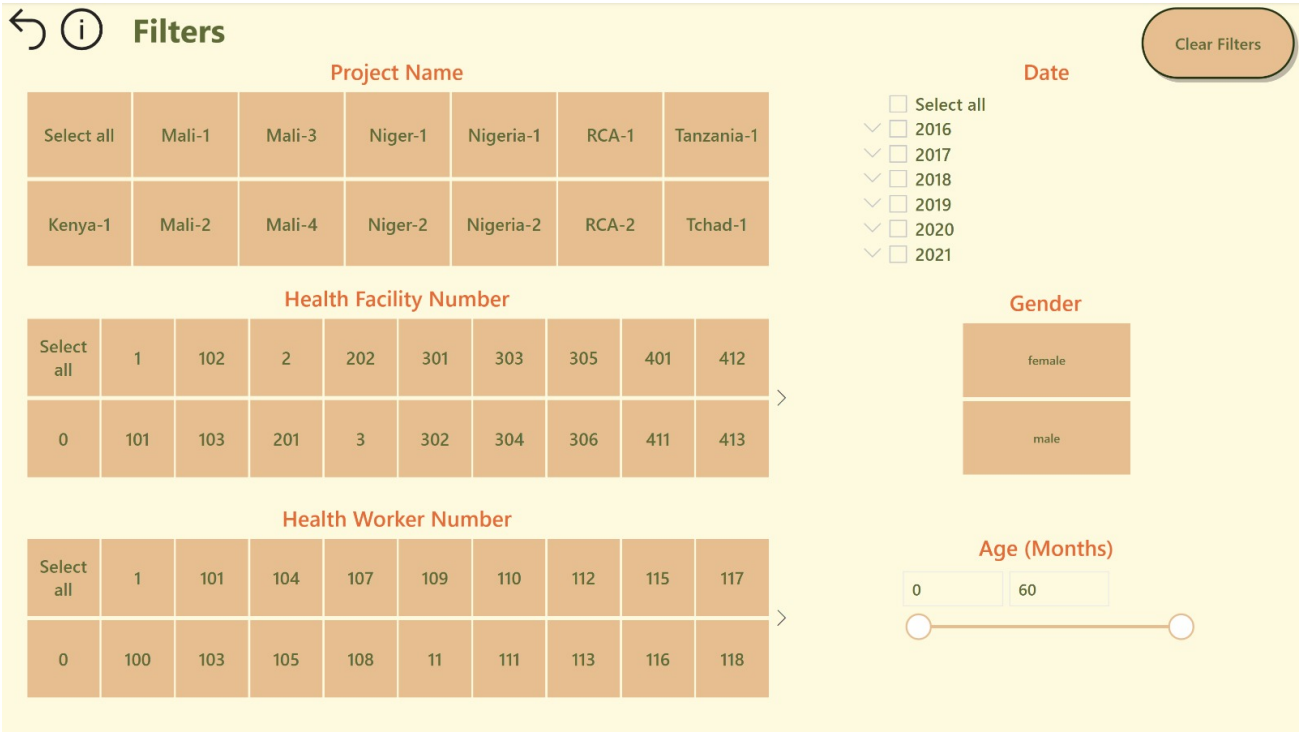

Figure 3. Filters Page: Allows filtering over six different features. Filtering one of the features automatically updates the selection options for rest of the features.

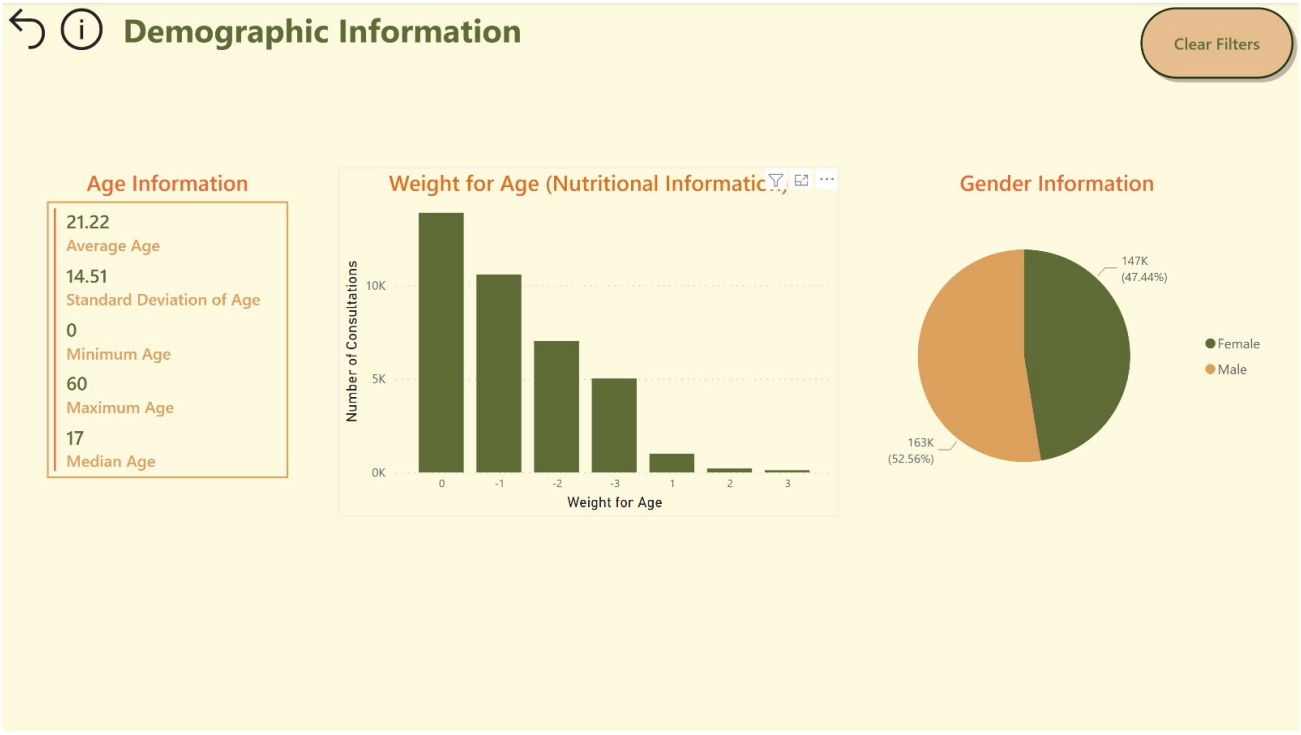

Figure 4. Demographic Information Page: Summarizes the demographic information by providing statistical information on age, bar chart on nutritional information and a pie chart on gender distribution.

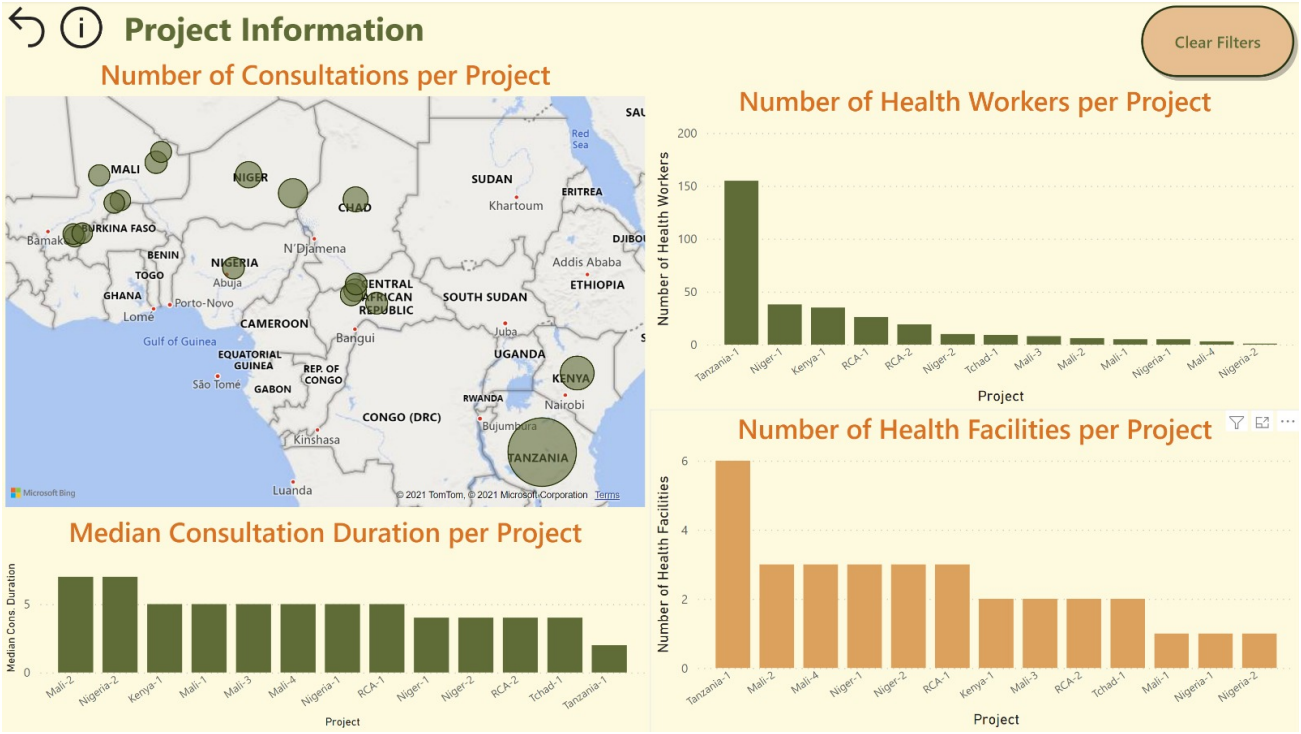

Figure 5. Project Information Page: Displays the distribution of consultations over the regions. Additionally compares the projects by number of health workers, number of health facilities and median consultation duration.

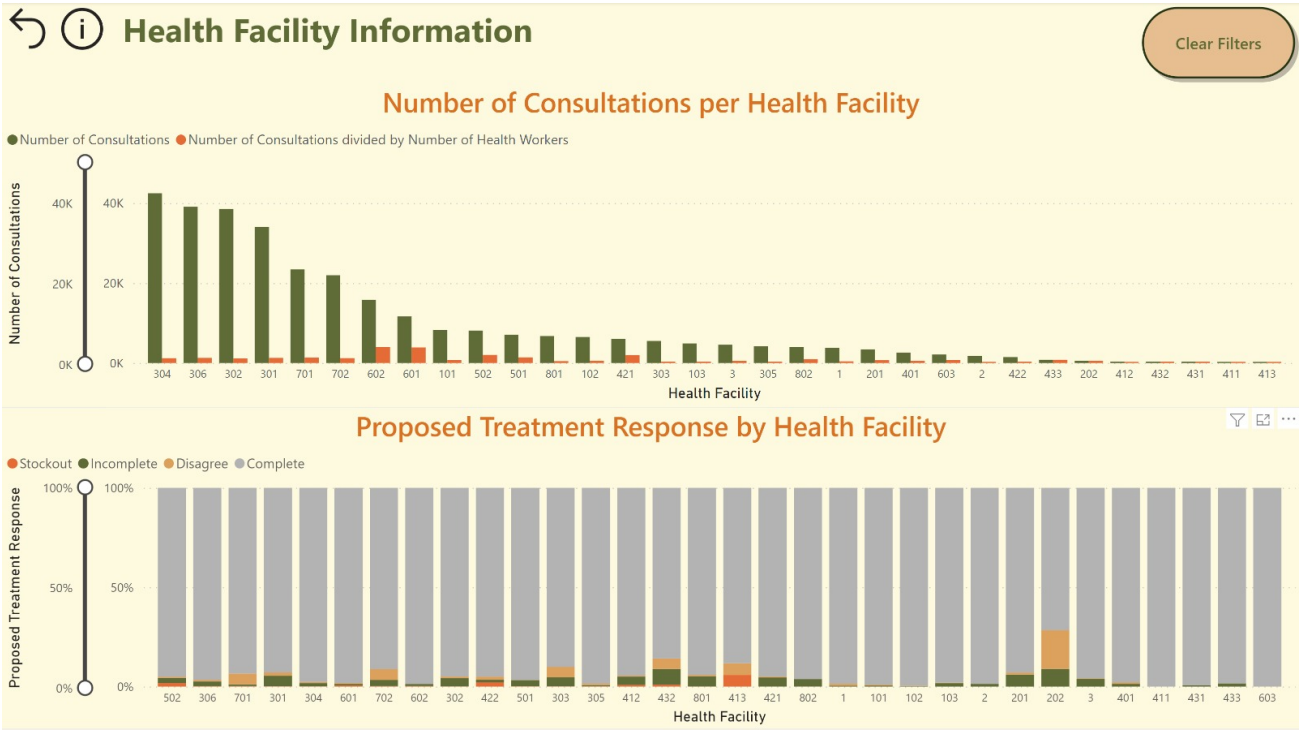

Figure 6. Health Facility Information Page: Displays the number of consultations as well as number of consultations done per health worker. Additionally, summarizes the proposed treatment response of each health facility.

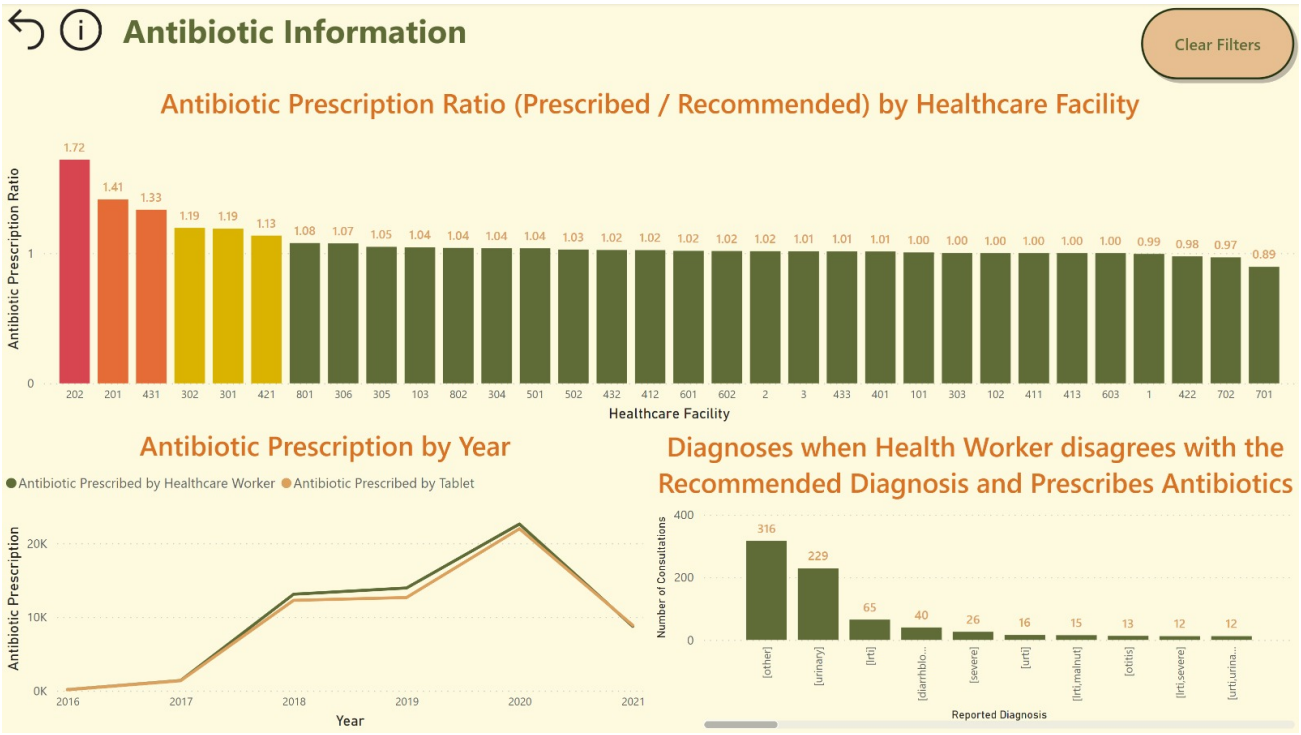

Figure 7. Antibiotics Information Page: Indicates which healthcare facilities prescribe too many antibiotics and the change of antibiotic prescription rate per year. Also, shows the top reported symptoms when antibiotics are recommended

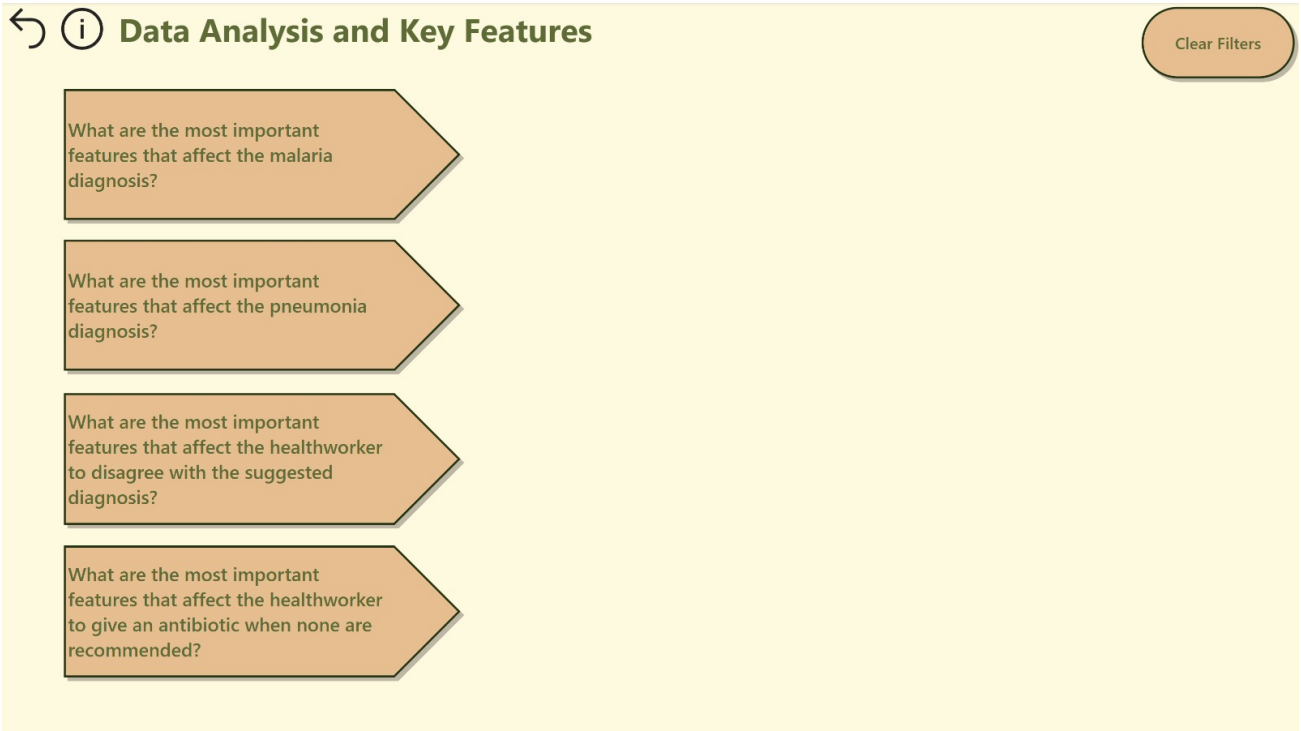

Figure 8. Data Analysis and Key Features Page: Menu page for moving to data analysis segments.

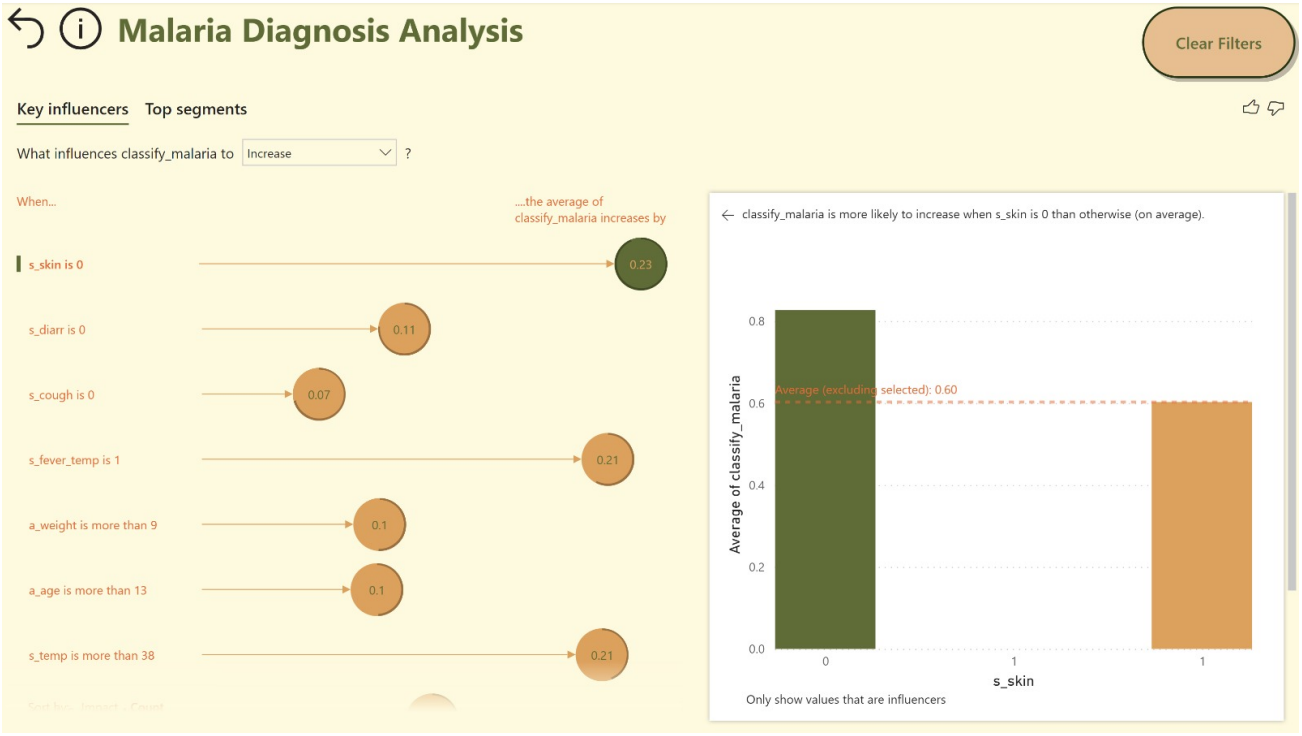

Figure 9. Malaria Diagnosis Analysis Page: Analyzes what are the most influential features that increase the malaria diagnosis rate.

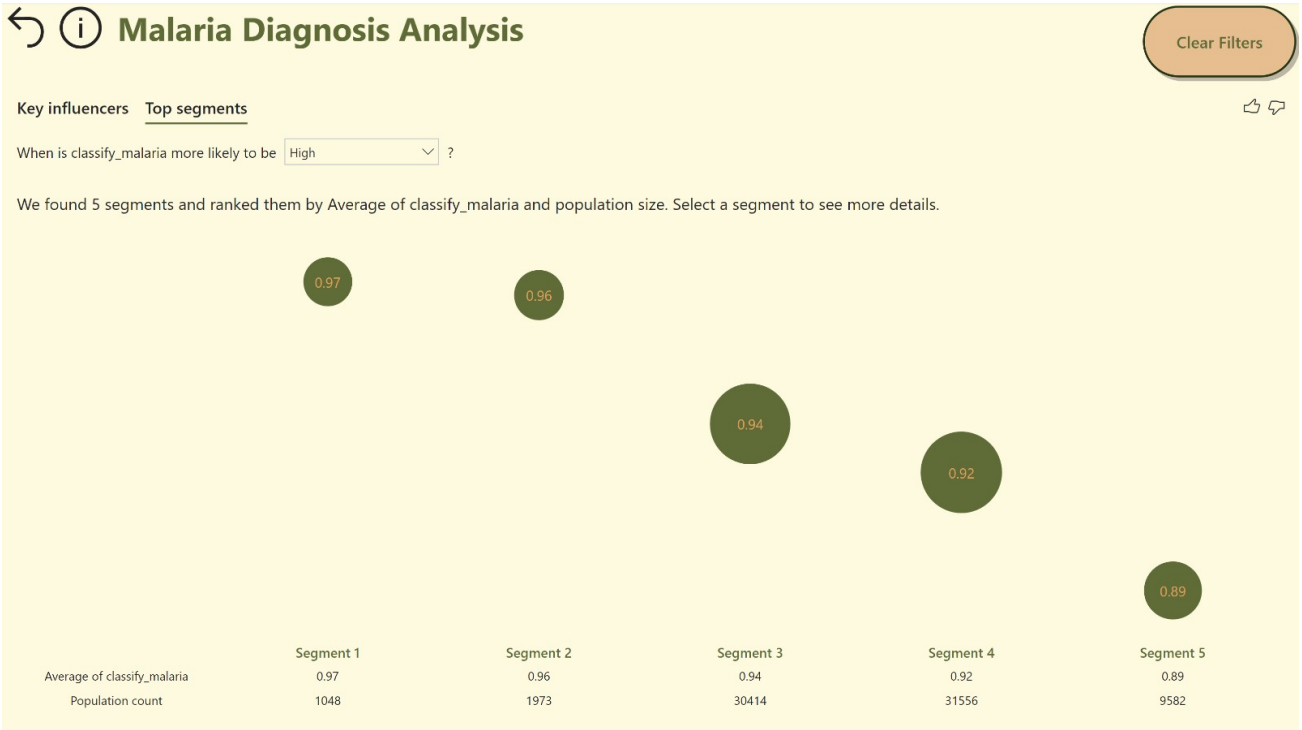

Figure 10. Malaria Diagnosis Analysis Segments Page: Displays the largest clusters that have above average malaria diagnosis rates.

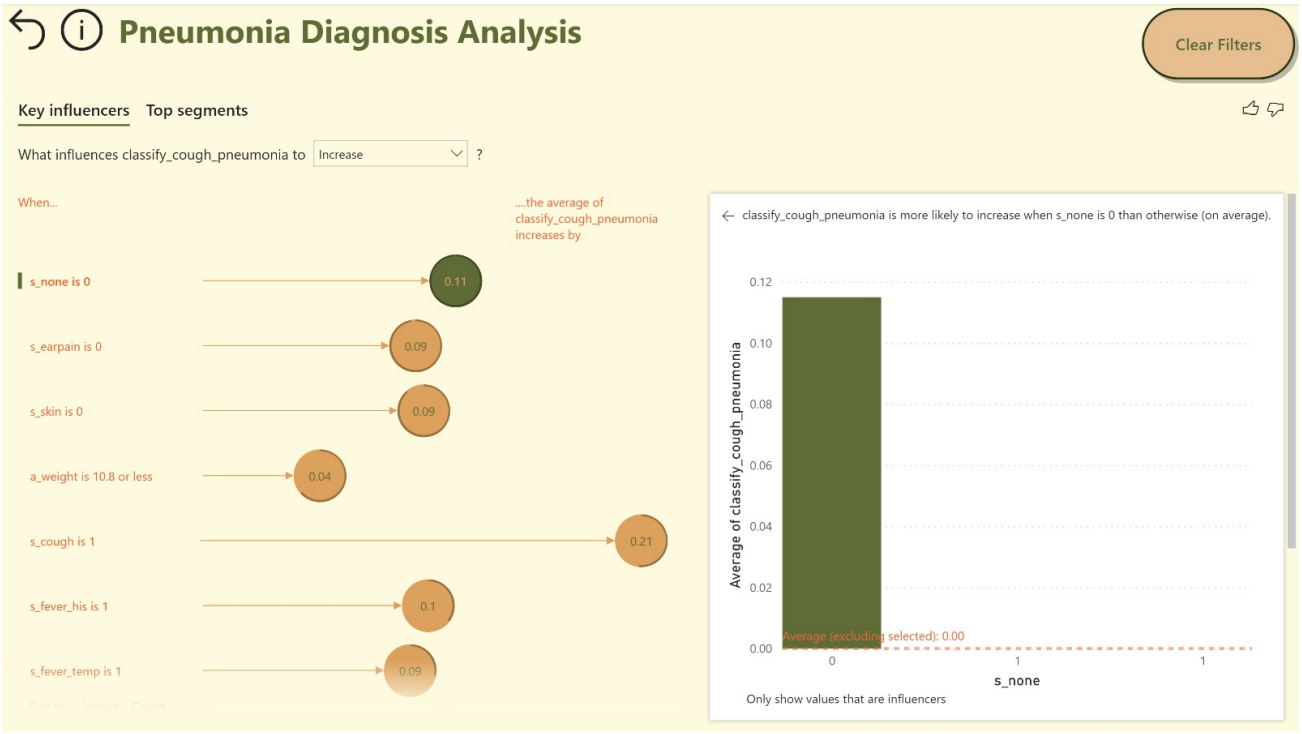

Figure 11. Pneumonia Diagnosis Analysis Page: Analyzes what are the most influential features that increase the pneumonia diagnosis rate.

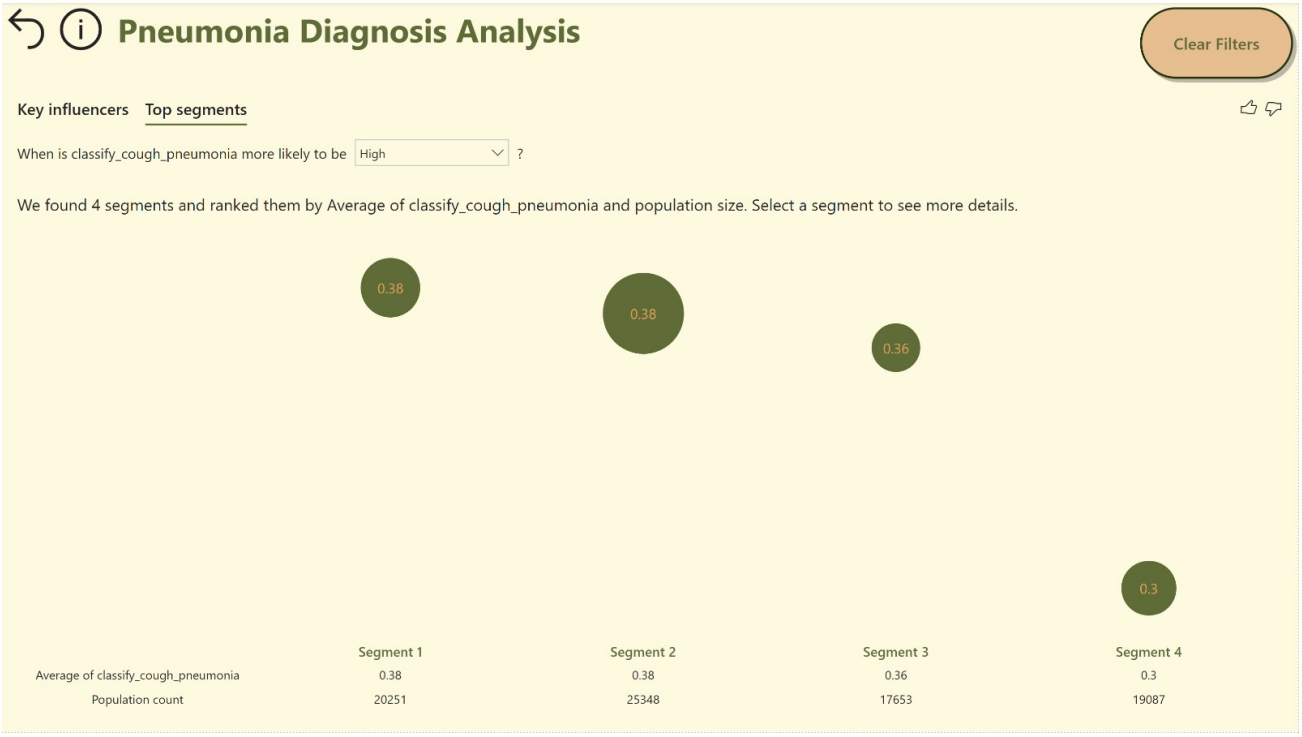

Figure 12. Pneumonia Diagnosis Analysis Segments Page: Displays the largest clusters that have above average pneumonia diagnosis rates.

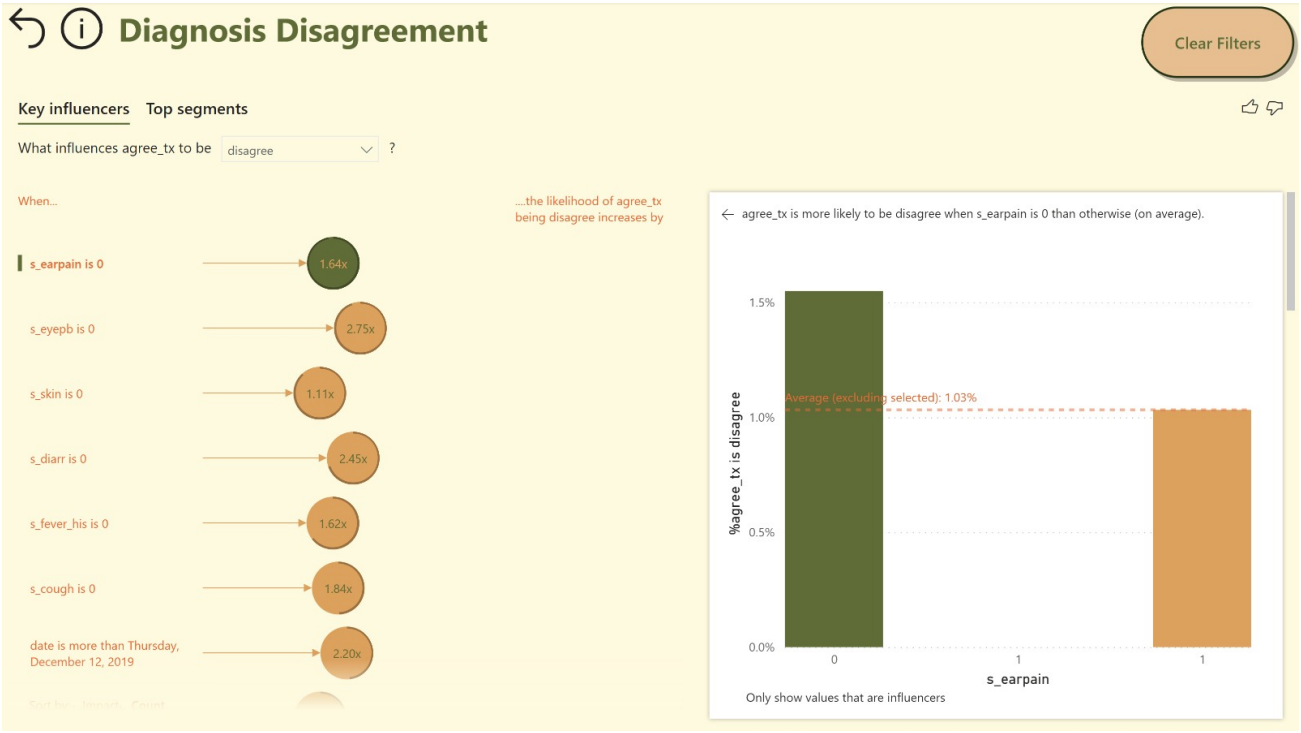

Figure 13. Diagnosis Disagreement Page: Analyzes what are the most influential features that increase the diagnosis disagreement rate.

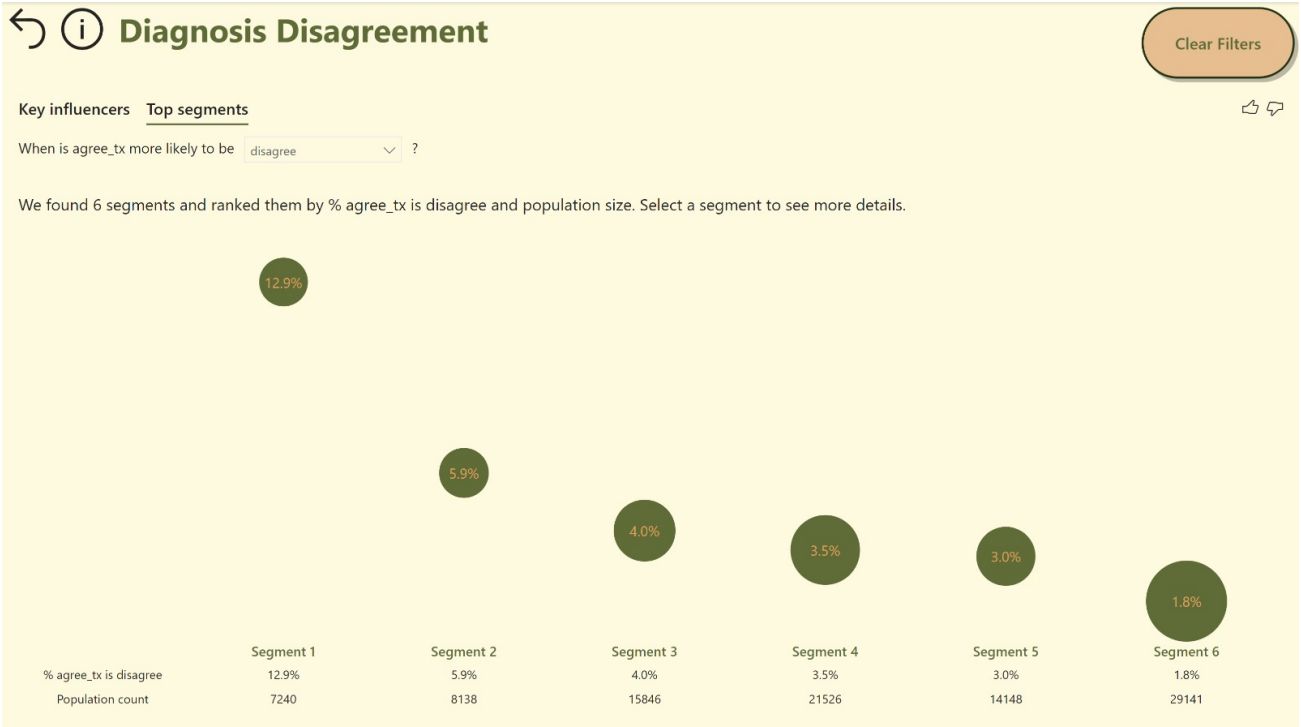

Figure 14. Diagnosis Disagreement Segments Page: Displays the largest clusters that have above average diagnosis disagreement rates.

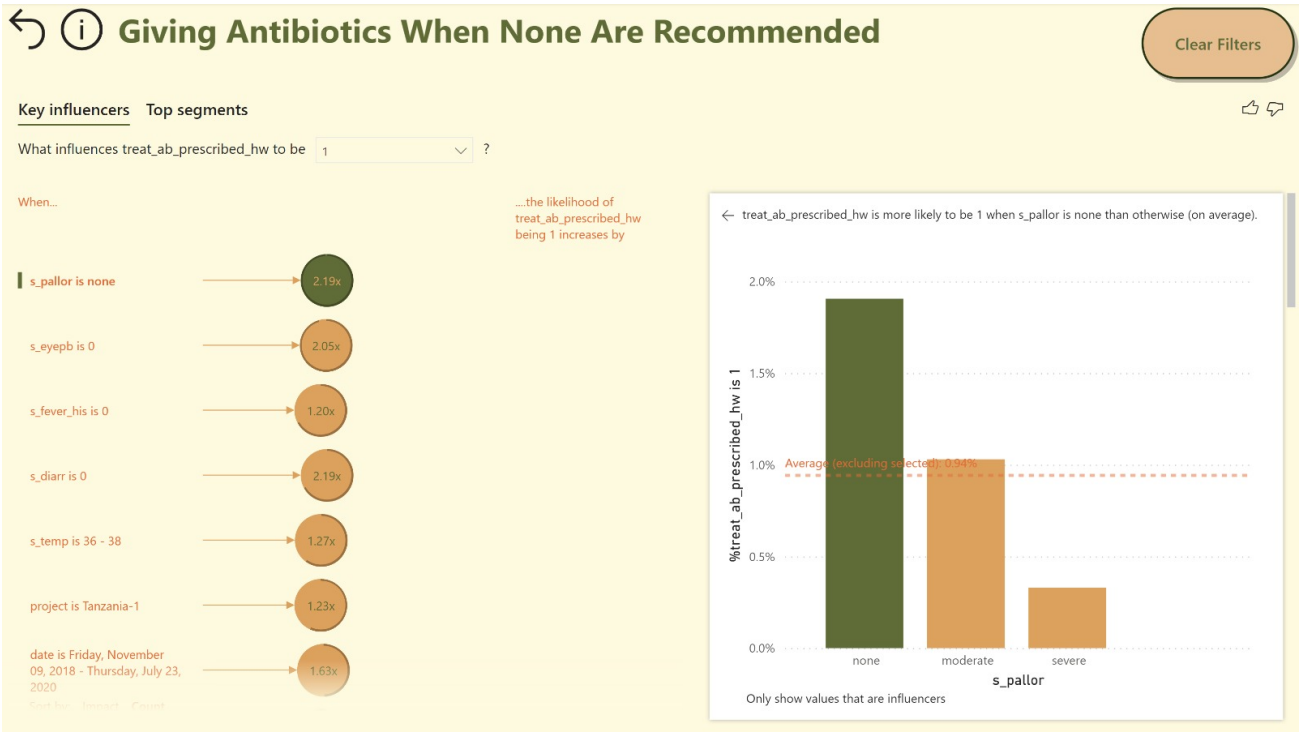

Figure 15. Giving Antibiotics When None Are Recommended Page: Analyzes what are the most influential features that increase the giving antibiotics when none are recommended rate.

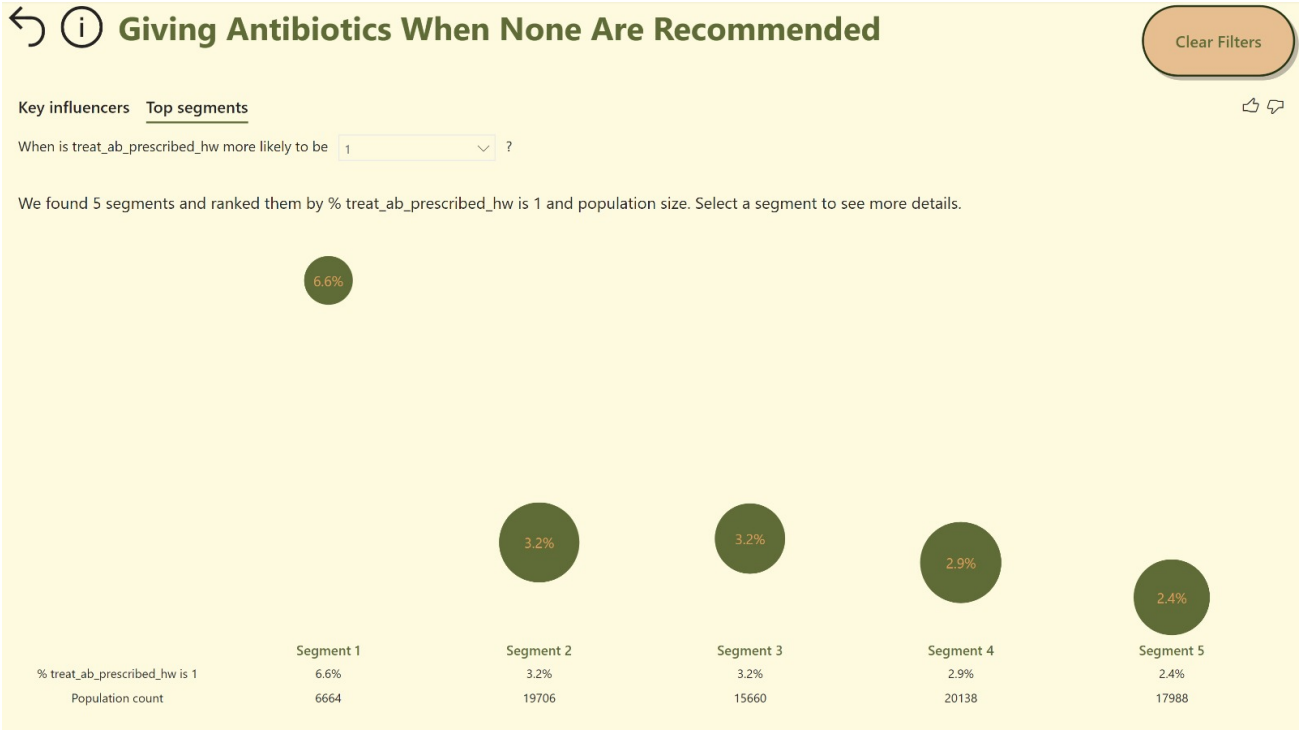

Figure 16. Giving Antibiotics When None Are Recommended Segments Page: Displays the largest clusters that have above average giving antibiotics when none are recommended rates.

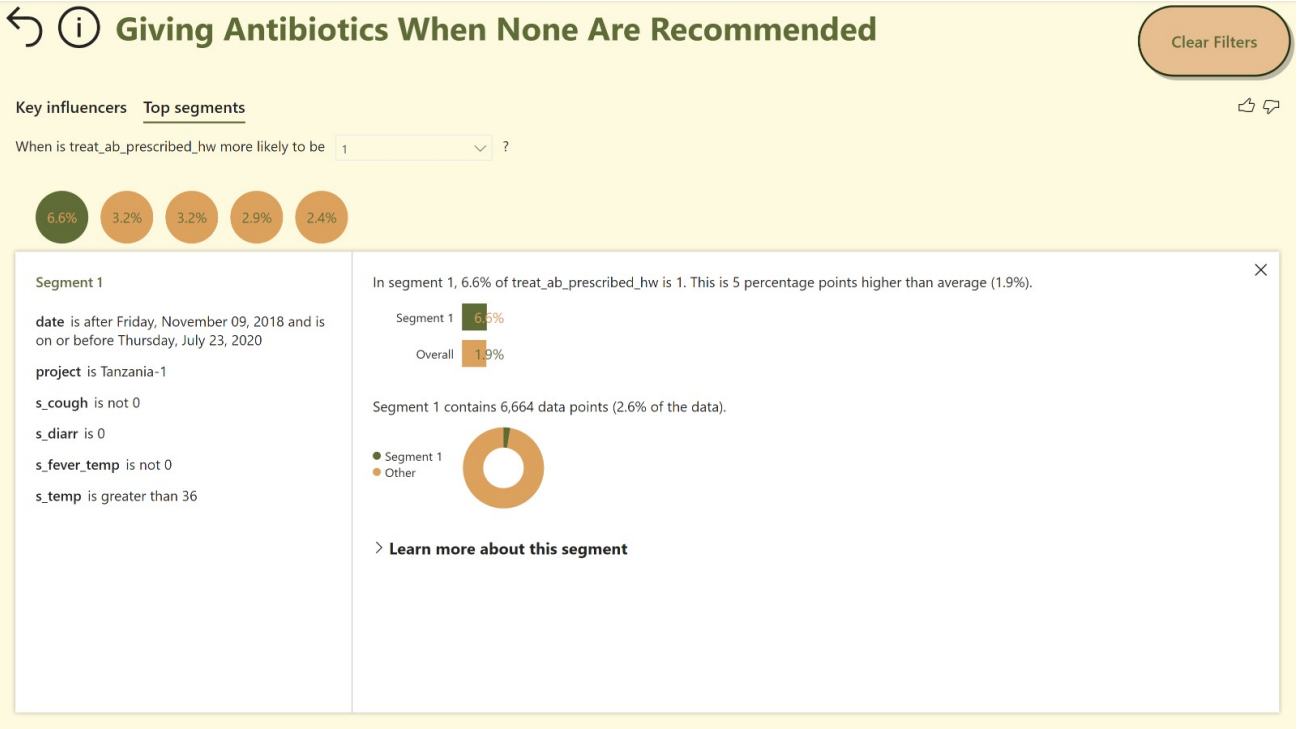

Figure 17. Giving Antibiotics When None Are Recommended Segment Details Page: An example of what happens when any of the clusters are clicked. It displays detailed information about the cluster.

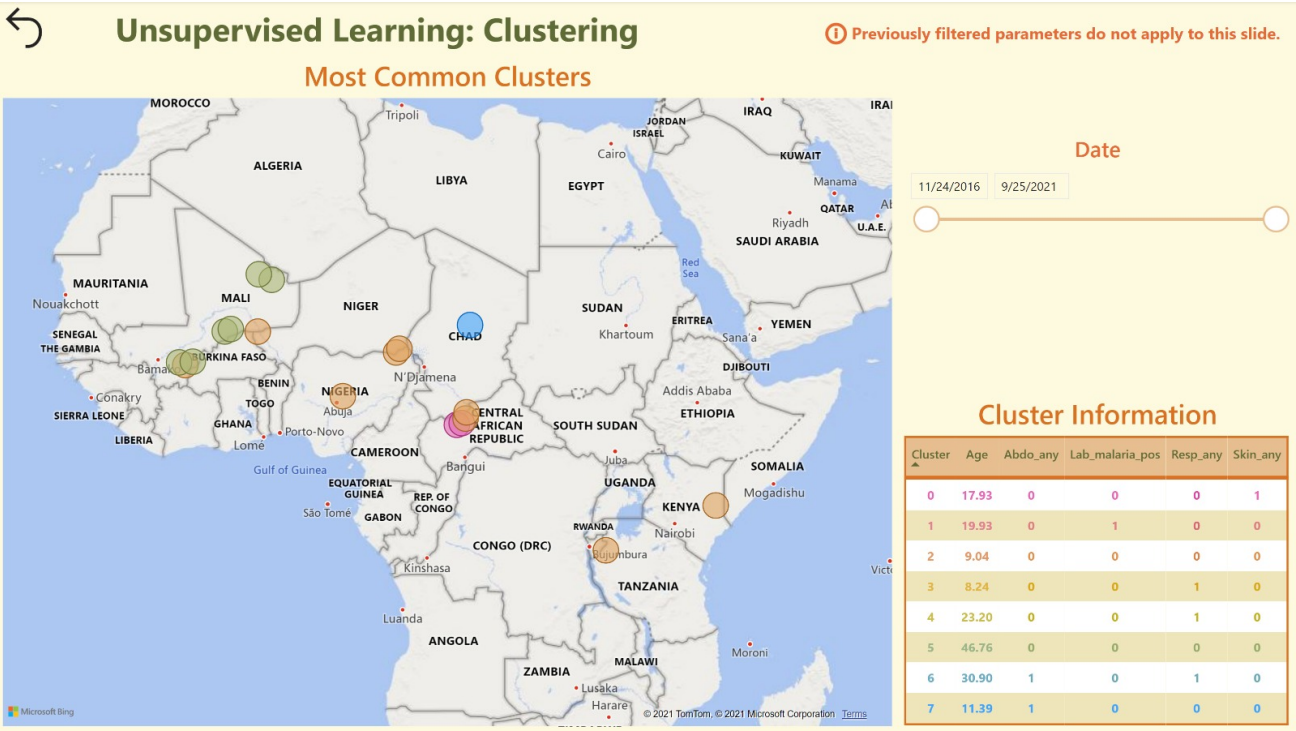

Figure 18. Unsupervised Learning: Clustering Page: Displays the clusters we have created. Each color represents a differrent cluster. Date can be filtered to see how clusters change over time. There is a cluster legend on the bottom right which describes the specific feature of clusters.

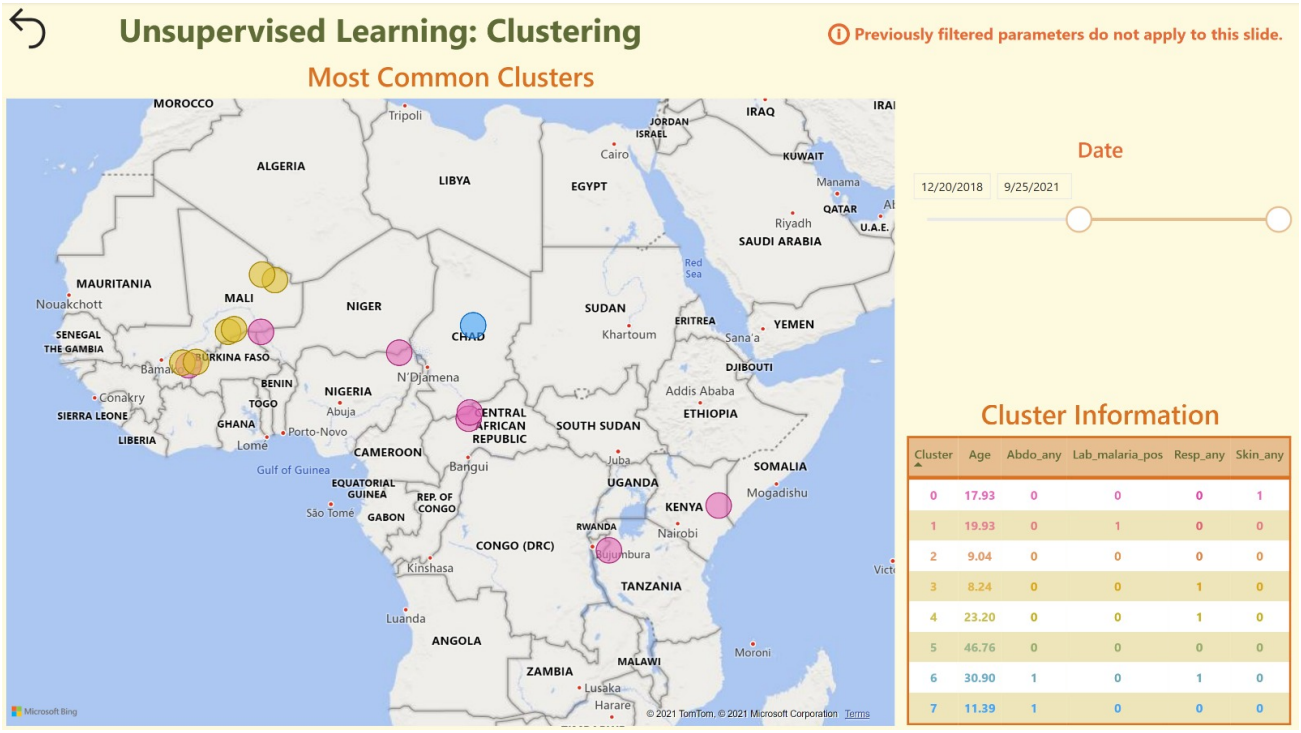

Figure 19. Unsupervised Learning: Clustering Page (filtered by date): Previous visual after the date has been filtered.