Knowledge Discovery in Medical Database Using Data Mining Techniques

Introduction

The dataset contains different healthcare centres ranging from primary health care centres to teaching hospital spread all over the country. The features in the dataset includes number of full time doctors, number of full time nurses, number of full time midwives, number of community and health workers, availability of skilled birth attendant, electricity, availability of caesarian section, measles, immunization for children, type of management of hospitals (faith based, private or public), improved water supply, improved sanitation, emergency transport services, vaccines freezer, antenatal care, family planning services and artemisin malaria drug availability. The dataset also contains longitude, latitude, ward, LGA, state, and geo-political region. The target variable is availability of maternal health delivery services which is a simple yes or no and that falls under classification. The following classification algorithm were compared Logistic Regression, Bagging, RandomForest, Boosting and were evaluated using accuracy, precision and recall.

Methodology

The dataset used contains about 33,149 hospitals, unfortunately they contain many missing values, the decision to remove missing values that have to do with features that bothers on services availability such as family planning, caesarian section, vaccination and also personnel which also include doctors, nurses, midwives, community health worker services. After removing records with missing features meeting the criteria, we are left with 26,238 hospitals. We equally dropped the following features; facility_name, community, ward, formhub_photo_id ,gps, survey_id, unique_lga, sector, state, geopolitical region and facility_id.

The next step was the encoding of the categorical features which includes the facility type and the type of management using the following methods; Find and replace, Label encoding and one hot encoding, but eventually the Find and replace method was used. The scikit learn train-test split was used at the ratio 70:30, 70% for training and 30% for testing with random seed of 101. The feature importance was also carried out using the Random Forest model. The Logistic Regression uses only default parameter and nothing was changed. Th Random Forest Classifier uses number of estimators to be 35, maximum depth is 20. The Boosting uses number of estimators to be 50 and a learning rate of 0.1. For Bagging, we used the logistic regression as the base estimator. The means of evaluation was the percentage accuracy, recall and precision on each model.

Exploratory Data Analysis

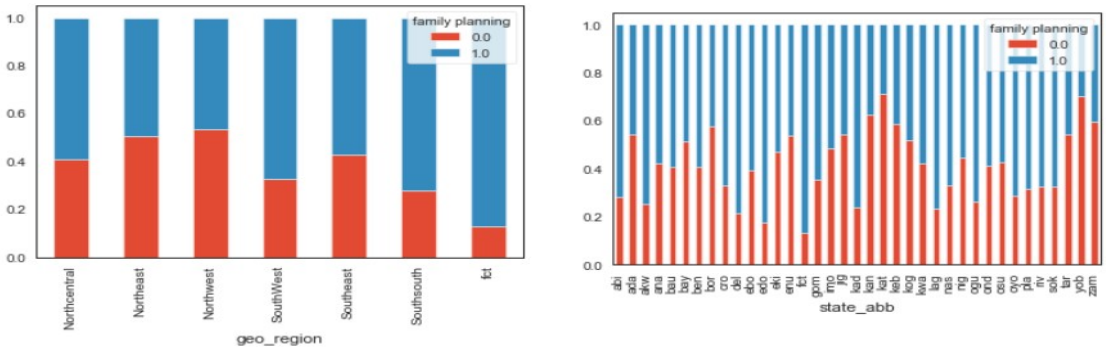

Figure 1a: Distribution of family Planning services across Regions and States

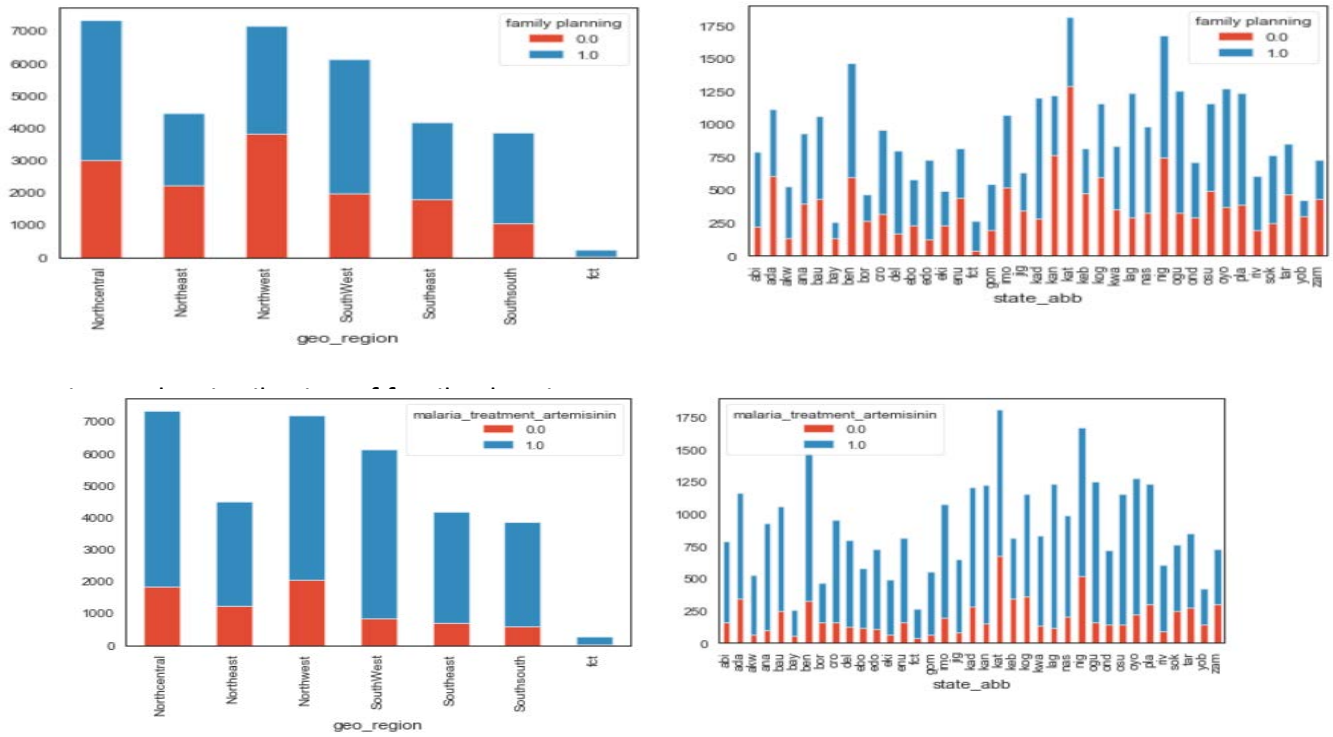

Figure 1c: Distribution of malaria treatment services across Regions and States

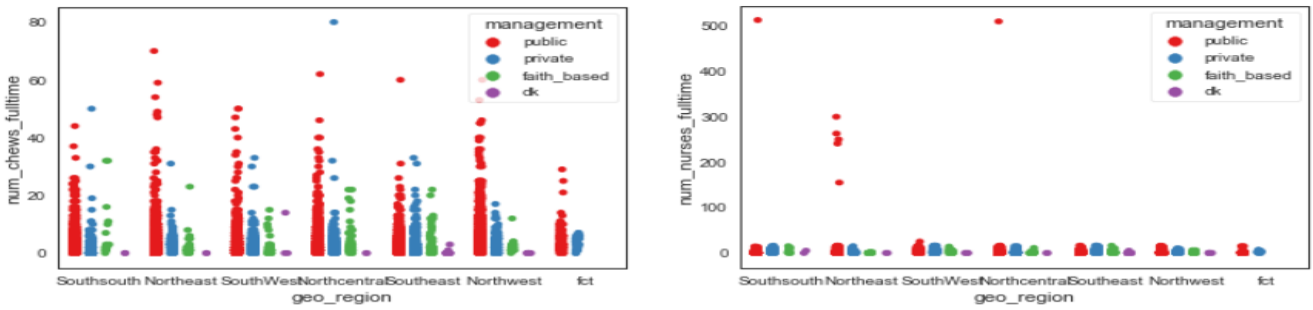

Figure 2a: Strip Plot of Community Health Workers and Nurses across Region by Management Type

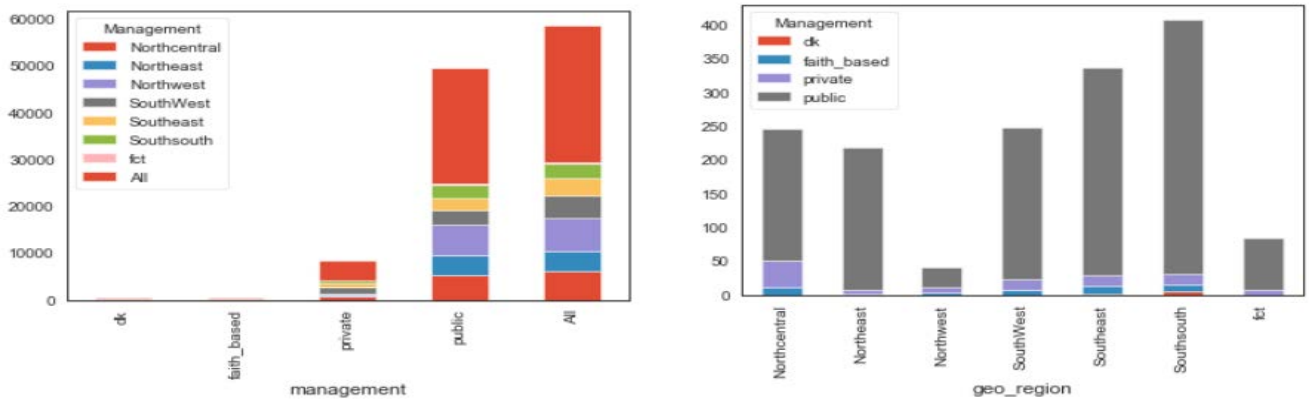

Figure 3: Bar Plot of Management Type across regions and by Doctors

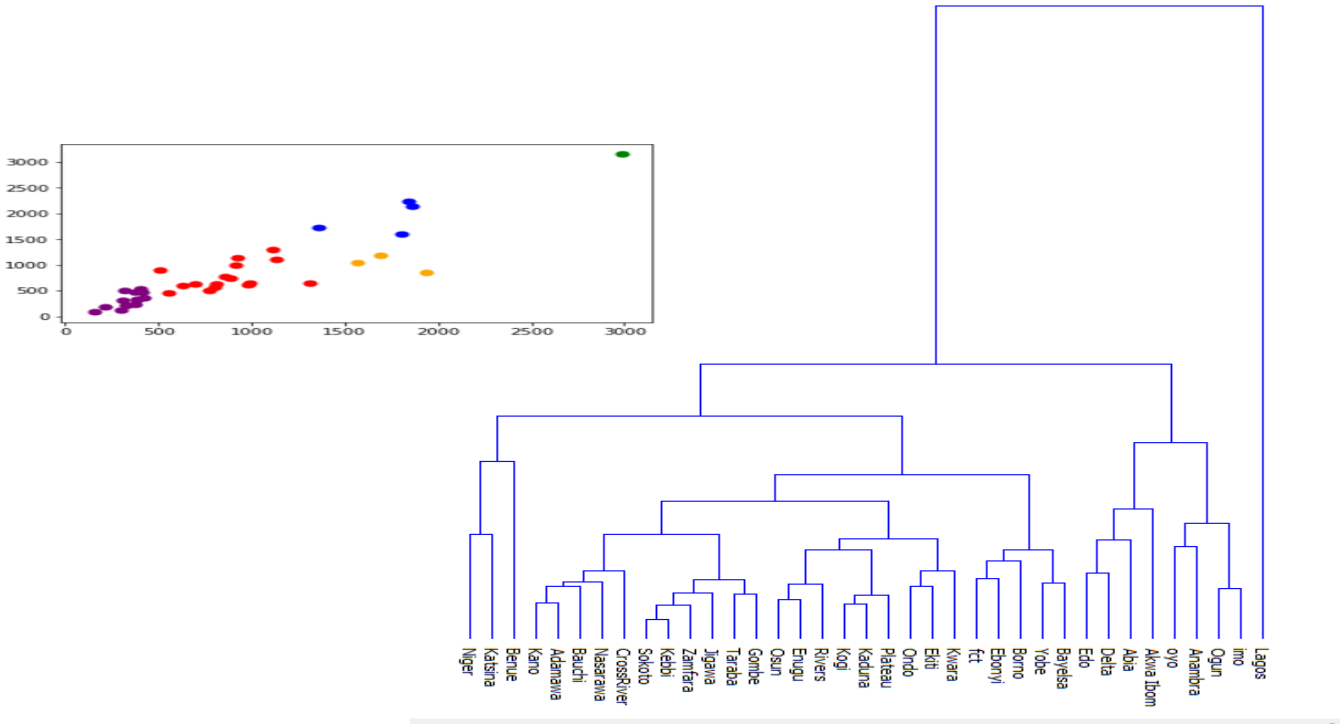

figure 4: A Dendogram of the Agglomerative Hierarchical Clustering of States by personnel using Average Linkage Criteria

Discussion

The feature importance reveals very interesting discovery which are quite interesting and also confirming the obvious, looking at the output shows that antenatal availability is the most important feature determining if a hospital will offer maternal health delivery services or not. It is also noteworthy to note that the availability of a full-time doctor is not in any way too important for having maternal health delivery services in a hospital

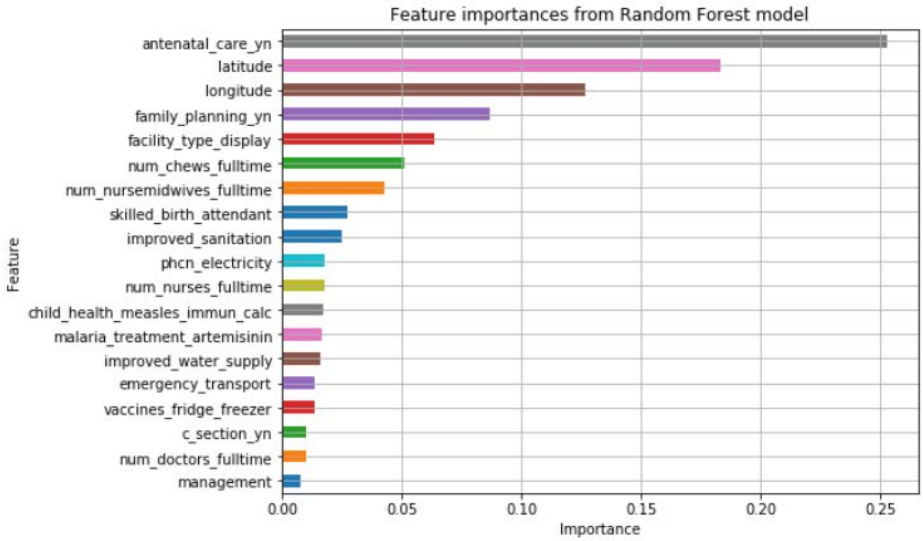

Figure 1: Feature importance using Random Forest Model

The evaluation metrics result is presented below

	LogisticReg	Bagging	RandomForest	Boosting
accuracy	84.4004	84.2988	85.9248	83.7398
precision	82.5085	82.4575	85.72	81.2846
recall	93.5332	93.4047	91.5203	94.3041

Figure 2: Evaluation metrics for Accuracy, Precision and Recall

It can be seen that each of the model closely related results, across each of the various evaluation metrics for each of them. The above table is also represented in a graphical form

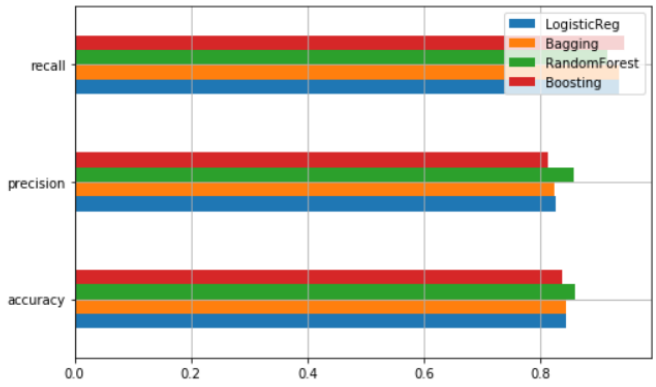

Figure 3: Evaluation metrics for The Models compared

Conclusion

From the result submitted so far, it can be concluded that Logistic Regression, RandomForest and Boosting are great classifiers, despite the fact that bagging which is an ensemble algorithm was used did not show any outstanding performance based on the evaluation metrics