TYPHOID DISEASE ANALYSIS USING REGRESSION TECHNIQUES.

MSc. IN ARTIFICIAL INTELLIGENCE

2024/2025

LEELA PARYHAR AND SONI NIDHI NARESH

ABSTRACT

Typhoid fever, also known simply as typhoid, is a disease caused by Salmonella enterica serotype Typhi bacteria, also called Salmonella Typhi. Typhoid continues to be a public health issue in Uganda and therefore it is essential for us to have proper analysis and prediction of the patterns and occurrences of this disease. This project aims to analyze and predict the number of Typhoid cases in the different districts in Uganda using various regression models, including Decision Tree, Random Forest, Support Vector Regression (SVR). This is with the use of environmental and population related data.

TABLE OF CONTENTS

SECTION 1

1.1 Motivation of the Project

Typhoid fever is a systemic febrile illness caused by *Salmonella enterica* serovar Typhi (Typhi) responsible for an estimated 11–21 million illnesses and 65,000–188,000 deaths worldwide each year. Typhoid fever remains an important public health problem in low- and middle-income countries, with large outbreaks reported from Africa and Asia. (Grace D Appiah, 2020) [1].

In Uganda, Typhoid continues to be a serious public health issue, with over 50,000 citizens suffering from the disease annually (Oguntimilehin, A., Adetunmbi, A. O., Olatunji, K. A. 2014) [2]. Early results indicate that population density has a great correlation with the number of typhoid incidences, and Kampala shows up with the highest number of typhoid incidences. It is often associated with poor sanitation, contaminated water and poor hygiene. This project aims to analyze and predict the number of Typhoid cases in the different districts in Uganda using various regression models, including sDecision Tree, Random Forest, Support Vector Regression (SVR). This is with the use of environmental and population related data, collected from the years 2011 to 2017 by the Uganda National Meteorological Authority and the Ministry of Health.

Our project aligns with the research field of Epidemiology and Public Health. We hope to better understand disease trends and identify risk factors by analyzing the patterns and occurrences of Typhoid as they relate with environmental and population data within Uganda's different districts.

The motivation for this project is to carry out proper analysis and prediction of the patterns and occurrences of Typhoid and to address the primary challenge of lack of accurate and timely predictions of typhoid outbreaks in Uganda, which hinders effective interventions.

1.2 Significance of the Project

Typhoid still poses a significant burden on the health care system of Uganda, especially in urban areas like Kampala where there is a high population density and poor living conditions may contribute to the fast spread of the disease. The prediction of the occurrences of Typhoid per unit population in each district can assist in implementing targeted measures that can improve the standard of living for people. This analysis can aid in implementing targeted measures in places with the possibility of greatest outbreak. These measures include water treatment and improved hygiene and cleanliness. Due to limited healthcare resources, analyzing the patterns of Typhoid diseases can help to identify high-risk regions and periods, thus leading to the efficient allocation of resources.

SECTION 2

OBJECTIVES OF THE PROJECT

2.1 Main Objective

To develop an accurate Machine Learning model that analyzes and predicts the patterns and occurrences of the number of Typhoid cases in the different districts in Uganda using regression techniques.

2.2 Specific Objectives:

1. To preprocess and clean the dataset that contains historical typhoid outbreak data (environment and population data).
2. To implement and compare regression techniques mainly Decision Trees, Random Forests, and Support Vector Regression.
3. To evaluate model performance using accuracy metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R2) Score.

SECTION 3

METHODOLOGY

3.1 Dataset

This is with the use of environmental and population related data, collected from the years 2011 to 2017 by the Uganda National Meteorological Authority and the Ministry of Health.

The dataset has dimensions of 112 rows and 17 columns and contains numerical data for most of the columns, and string data under the column for district names.

The target variable is Typh_Rate, which is the cases of typhoid per unit population of people in each district, and eleven features were used.

The features used are:

- **Typh_Inc:** typhoid incidences
- **X_coord:** x coordinate for geo spatial information
- **Y_coord:** y coordinate for geo spatial information
- **HH_Wash:** % of population exercising hand washing practice after using toilet
- **PH_Lands:** % of population that stays in highlands
- **P_Density:** population per unit area??
- **Urban_leve:** Urbanization level, proportion of the district that is urban
- **ARainfall:** Average rainfall amount in mm
- **Temp_Max:** maximum temperature
- **Temp_Min:** minimum temperature
- **Typh_Rate:** cases of typhoid per unit population
- **Pn_Floods:** proportion of people affected by floods per district
- **P_male P_Female:** gender proportions
- **Typh_Per:**
- **OBJECTID**

3.2 Methodology and Architecture

Regression techniques were selected due to the continuous nature of data in the project.

The following algorithms were implemented in the project are as follows:

Decision Tree Model

The Tree starts with a root node that represents input data. The algorithm recursively splits the data into subsets based on the most informative features, using a splitting criterion (e.g., information gain and entropy). Each internal node in the tree represents a decision made based on a feature and a threshold. The terminal nodes (leaves) represent the predicted values recursively partitioning the data into smaller subsets based on new input data flows down the tree, making decisions at each node, until it reaches a leaf node, which outputs the predicted value. The tree grows by rec the most informative features.

Random Forest (RF)

Random Forest is a supervised machine learning algorithm that creates multiple decision trees and merges them together to get a more accurate and stable prediction. The Random Forest algorithm creates an ensemble of decision trees. Each tree is constructed using a random subset of the dataset to measure a random subset of features in each partition. In prediction, the algorithm aggregates the results of all trees, either by voting in the case of classification or by taking the average in the case of regression.

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning algorithm used for regression tasks. It's goal is to predict the cases of typhoid per unit population based on various independent variables. The input layer takes in the feature vector, which is a set of input variables that are used to predict the target variable. A radial basis function kernel is used to transform the input data into a higher-dimensional space, where the data can be linearly separated. A subset of the training data is selected as support vectors, which are used to define the hyper plane that separates the data. The hyper plane is the decision boundary that predicts the cases of typhoid per unit population based on the input variables (X). The output layer generates the predicted target variable based on the hyper plane and the support vectors.

3.3 Computational and analytical tools used for the implementation of the project:

The key libraries that were supported the implementation of this project are as follows:

Pandas – Used for **handling and cleaning typhoid dataset**.

NumPy - Providing support for **numerical operations** on the dataset.

Matplotlib – For creating **visualizations of the typhoid data.**

Seaborn – Used for enhancing **visualization of statistical data**.

Sklearn – This is the major library that provides the major modules for model selection and evaluation, as well as for importing the modules that support the different regression techniques, and finally the module that supports data processing.

SECTION 4

EXPERIMENTS AND RESULTS

This project is implemented on the dataset that contains typhoid occurrence data, collected from the years 2011 to 2017 by the Uganda National Meteorological Authority and the Ministry of Health. The data set used has a dimension of 112 rows and 17 columns.

The Conditions required for this project are as follows:

To evaluate and compare the performance of different regression models on our dataset, 5-fold Cross-Validation (CV) was implemented to ensure that the evaluation is not biased by a particular train test split.

Cross-Validation Setup - Use of KFold cross-validation with k=5 folds, shuffling the data with a fixed random seed (random_state=42) for reproducibility. In each fold, the dataset is split into a training set (80%) and a validation set (20%).

Since Linear Regression and SVR are sensitive to the scale of features, applied Standard Scaling (zero mean, unit variance) on the training data and used the same transformation on validation data.

Tree-based models (Decision Tree, Random Forest) do not require scaling, used the raw features for them.

Evaluation Metrics

For each fold and model, we computed four metrics:

- MSE (Mean Squared Error): Measures average squared prediction error.
- RMSE (Root Mean Squared Error): Square root of MSE, interpretable in the same units as the target variable.
- MAE (Mean Absolute Error): Average absolute difference between predicted and actual values.
- R^2 (Coefficient of Determination): Proportion of variance in the target explained by the model (higher is better, can be negative if model is worse than baseline).

Aggregated Results

After running all 5 folds, we calculated the average of each metric across folds for every model. This smooths out variability due to different splits and provides a fair comparison.

GridSearchCV is used for optimizing the hyperparameters to achieve better accuracy.

Recursive Feature Elimination (RFE) is applied to identify the most influential factors contributing

to typhoid occurrences.

The Evaluation Metrics for the Project are as follows:

As seen from the table, the Random Forest model has the most promising results.

Table 1

Model Evaluation Scores

Model	MSE	**RMSE**	MAE	R2
Linear Regression	0.000269	0.011304	0.005010	-2.1102
SVR	0.000349	0.018662	0.018002	-6.091
Decision Tree	0.000073	0.008169	0.003919	-0.438
Random Forest	0.000032	0.005462	0.002713	0.4567
Random Forest (Tuned)	0.0000298	0.005456	0.002618	0.5105
XGBoost	0.0000441	0.006642	0.003240	0.2745

SECTION 5

CONCLUSION

The importance of this work is due to the fact that Typhoid attacks over 50,000 Ugandans annually despite the inadequate medical facilities.

With the use of machine learning, we can not only analyze but also predict trends in Typhoid cases in Uganda and thus act accordingly in order to control the levels of Typhoid attacks.

From the comparison above of the four regression models, Decision Tree, Random Forests, Support Vector models, the Random forests model has the best metric scores and therefore fits to be the best model for this regression problem.

The main limitation of the work is based on the data set. The meta data that was used to describe the data set is quite ambiguous due to the poor elaboration of column names. Additionally, there is uncertainty about the data's level of integrity due to the lack of specification about the methods of data collection. The data was also collected from 2011 to 2017, thus the relevancy of the data will change due to the dynamic nature of the target variable.

Some of the future work to improve on this project include:

- Address the previously mentioned limitations with the use of up-to-date data which has a reliable degree of integrity.
- Find data sources popular for having a standard quality of data sets such as the World Health Organization.
- Use other regression models to further investigate which model best fits this regression problem.

REFERENCES

[1] :Grace D Appiah, A. C.-E. (2020). Typhoid Outbreaks, 1989–2018: Implications for Prevention and Control. *The American Journal of Tropical Medicine and Hygiene*.

[2] : Oguntimilehin, A. A. (2014). A machine learning based clinical decision support system for diagnosis and treatment of typhoid fever. . *A Machine Learning Based Clinical Decision Support System for Diagnosis and Treatment of Typhoid Fever*.