STA363 Project 1

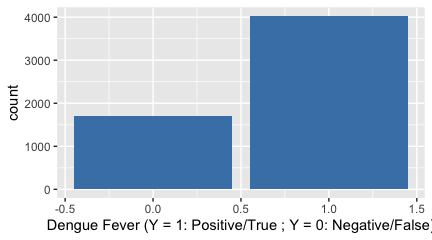
Zishan Shao

2022-09-05

# Section 1: Introduction and Data

In this report, we are working with patients data from the local hospitals of Vietnam. The data was composed by 5726 children who has symptom of dengue fever when they were brought to clinics. The dengue was . While most of them only develop mild illness, some develops into a potentially lethal complication, called severe dengue (citation: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>). The study of the dengue has long been conducted worldwide, but detection method without accession of advanced medical care is yet to be invented. This caused difficulty for diagnosis of dengue in underdeveloped regions. Therefore, the main goal of this analysis is to construct a relatively applicable statistical model that efficiently, yet accurate, predicts the dengue with common symptoms of the disease. The model will be built with features that has highest potential influences to the outcome, and trained with the data from real-life data. The model is expected to be applicable in most places in the world even without accession of advanced medication, so the simplicity of the model will also be considered. Finally, we expected to select the best model from multiple trained models with different tuning parameters, thresholds, and sampling techniques. We will evaluate our model based on precision metrics of statistical models (e.g. accuracy, geometric mean, false positive rate)

### Label Analysis



In this data, there are totally 15 variables comprised by 14 features (categorical or numerical) and 1 labels. Our goal is to predict the label Y according to features with different weights. Meanwhile, the Y is a binary variable: Y = 1 means the patient has dengue fever and Y = 0 means the patient does not have dengue fever. In other words, this report is dealing with a prediction task with classification problem. From the Figure 1.1, we could see that the distribution of the outcomes (0 & 1) is about 1 : 2.5 in ratio, which is valid for model building because there is a fair chance to meet each scenario.

### Data Cleaning

|  | Num Observations |
| --- | --- |
| Before Cleaning | 5726 |
| After Cleaning | 5726 |

Along with the label analysis, it is also necessary to clean the data by removing or refilling the rows with missing information. However, the cleaning results indicate that the total observation number remain unchanged after the cleaning (both equals to 5726), indicating that there is no missing data existed in the explanatory variables. This also indicates there is no missed information in any features (indicated by columns), so the dimension of the entire data set remain same. Therefore, we proceed to construct the model.

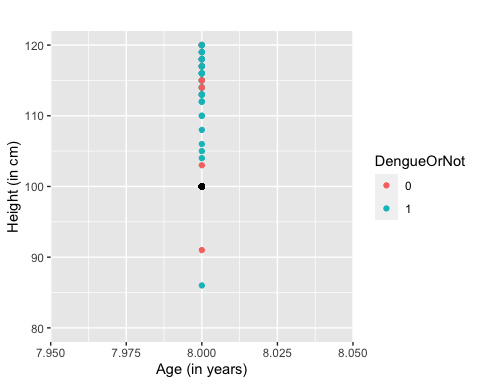
# Section 2: K-Nearest Neighbor (KNN)

Because the label Y is a binary variable, there are two kinds of ubiquitous methods to apply: multivariable logistic regression model (MLR) and K-Nearest Neighbor (KNN). The KNN is a clustering technique/algorithm that is effective in classifying labels based on the neighboring outcomes. The KNN assumes the similarity exists among outcomes of nearby observations, and is effective if there is clustering of outcomes in the data. Therefore, KNN is expected to be suitable because the diagnosis of dengue is difficult without accession of standard test.

## 2.1: Brief Example of K-Nearest Neighbor (KNN)

KNN algorithm is surprisingly simple for interpretation. In the following example, a mock sample was provided with age = 8 and height = 100 cm and a KNN based model was constructed with feature Age and Height. We set the tuning parameter k, the number of neighbors for consideration, equals to 5 and make predictions with model and a scatterplot.

### Prediction with Scatterplot & Model



The figure 2.1 provides with all observations with age = 8. The orange dots indicates that the patient is negative and cyan dots indicates the patient is positive. The black point, which is the mock observation, lays at 100 cm height. The closest 5 observations to the mock observation has the outcome of [0,1,1,1,1], which the positive result was the mode of the neighbors. Therefore, it is likely that, by assumption, the mock observation also has a positive outcome.

The model prediction also support our prediction with the scatterplot. The predicted label of the mock observation indicates that the result should be Y = 1, which means the kid get the dengue fever.

However, the prediction result does not necessarily reflects the accuracy of the model. To determine if this model is effective, we need to compare the performance metrics such as geometric means of KNN models adapting different sampling methods or tuning parameters. This part will not be specified in the example but will be included and explained later.

## 2.2: KNN models (with all features) & train-test split method

In section 2.2, we will construct our first model with all 14 features via train-test split method. The train-test split method, by its name, splits the data into training and testing datasets. It is necessary when the testing dataset was not provided and was effective in minimizing the overfitting problem. In this case, I will compare the traditional train-test split by 20% to 80% split and k-fold cross validation to construct the train-test dataset.

The tuning parameter is also considered. The k is the tuning parameter of every KNN-based models, which defines the number of neighbors to be considered. This parameter enables the KNN algorithm to adapt with different scenairos. For instance, if the sample size was very small, the neighbor number should be comparably smaller than the neighbor numbers in relatively larger datasets. In this case, I will construct models with k = 3, 5, 7, 9 and compare their geometric means (computed by square root of sensitivity x specificity). The model with highest geometric mean will be selected.

### Train-Test Spliting

Observations in Train/Test Sets & True/False Counts

|  | Count |
| --- | --- |
| Num observations in test set | 1145 |
| Num observations in training set | 4581 |
| True Dengue Number in test set | 803 |
| True Not Dengue Number in test set | 342 |

The training set has 4581 observations and the test set has 1145 observations, which in total has 5726 observations. The test set includes 803 patients who diagnosed dengue and 342 patients who does not diagnosed dengue. The later two numbers are essential for computation of geometric means.

### Model Creation (via loop)

In order to create and comparing models with different tuning parameters, a loop is adapted to iteratively train and test the model. The predicted result was recorded in a 4x2 data frame with value of k and corresponding geometric mean (GMean) in table 3.

Models with different k values with train-test split method

| K | Geometric Mean |
| --- | --- |
| 3 | 0.526 |
| 5 | 0.554 |
| 7 | 0.540 |
| 9 | 0.548 |

The table 3 indicates that the k = 5 has the highest geometric mean, indicating that, at k = 5, the model classifies the observations at highest precision because this means we have highest proportion of correctly classified positive and negative cases. Therefore, the from KNN model built with train-test split method.

### Estimation Variation Test

It is worth mentioning that the train-test split method has two drawbacks. Firstly, it could cause reduction in training data size by splitting 20% of it to testing data. Such large proportion of data lose could potentially cause inaccuracy in prediction. Secondly, the randomly chosen rows could cause high estimation variation because the training results are highly dependent on the chosen observations, especially for data set with limited number of observations. Therefore, it is necessary to test if alternative sample could lead to a different conclusion.

Models with Resampled Data

| K | Geometric Mean |
| --- | --- |
| 3 | 0.539 |
| 5 | 0.555 |
| 7 | 0.543 |
| 9 | 0.556 |

Under seed = 1919810 (previously 114514), the k = 9 obtains a geometric mean equals 0.556, while the k = 5 is 0.555 with 0.001 lower than k = 9. Nonetheless, the closeness of two geometric means failed to indicate a statistically significant difference in prediction capacity. Meanwhile, the value of the geometric mean of k = 5 only varies by 0.001 when seed = 114514. Therefore, it is proven that the model with k = 5 is not significantly sensitive to the observations chosen.

### Confusion Matrix of KNN (train-test split method) & Metrics Computation

Confusion Matrix of KNN

|  | True Not Dengue | True Dengue |
| --- | --- | --- |
| 0 | 120 | 109 |
| 1 | 220 | 696 |

### Metric Computation

## 2.3: KNN Models (with All Features) & 10-Fold Cross Validation

In section 2.3, we applied the 10-fold cross validation method as the sampling method to obtain the training and testing data. The k-fold cross validation, like the train-test split, applied when testing data was not provided. K-fold cross validation splits the data to k folds and use each folds’ data iteratively as the testing data (the rest data as training data). The results of prediction will be computed each time with the size of a fold, and finally the size equals to 5726.

Two most common fold numbers are 5 and 10. Considering the size of the data, the 10-fold cross validation should be a more appropriate choice. The dataset was not exactly divisible by 10, so we will create a folder with larger storage space and balance the number of observations in each fold with difference no larger than 2. This could be achieved by taking the ceiling of the data assigned to folds so that all observations will be sampled to folds.

Models with different k values with 10-Fold Cross Validation

| K | Geometric Mean |
| --- | --- |
| 3 | 0.533 |
| 5 | 0.529 |
| 7 | 0.523 |
| 9 | 0.538 |

The testing result in table 6 indicates that the k = 9 with 10-fold cross validation has the highest geometric mean (0.538), which means that the KNN-based model with k = 9 achieve the highest precision under the 10-fold cross validation method.

### Confusion Matrix of KNN (10-fold cross validation method)

According to the previous result, we construct our model with k = 9 and fold number = 10. We then construct the confusion matrix of the model and compute the metrics to evaluate the prediction capacity of the model.

Confusion matrix (fold=10) of DengueClean

| True Not Dengue | True Dengue |
| --- | --- |
| 525 | 508 |
| 1173 | 3520 |

### Metrics Computation

| General Table of Metrics |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GMean | Sensitivity | Specificity | Accuracy |
| ———— | ——- | ————- | ————- | ———- |
| train-test | 0.554 | 0.865 | 0.353 | 0.713 |
| 10-fold | 0.538 | 0.874 | 0.309 | 0.706 |

The general table of metrics indicates that the train-test split model has higher geometric mean, specificity, and accuracy, while the 10-fold method has higher sensitivity. Therefore, the KNN-based model with train-test split method and k = 5 should be our final choice as it has better performance on most of metrics.

The reason behind this result make sense because the data set is not limited in size. With 5726 observations, it turns that losing 20% of the data does not harm the prediction capacity of the model or cause very large estimation variation. What’s more, the estimation variation test also indicates that the prediction capacity of the model with k = 5 and train-test split method is not overly sensitive to the observation selections. Considering that the 10-fold cross validation method is computationally expensive, the choice of model with k = 5 and train-test split method make sense.

# Section 3: Multivariable Logistic Regression

Noticing that the label Y (Dengue Fever) of the dataset is binary, we could also apply the multivariable logistic regression model in this case. The multivariable logistic regression model is a statistical model that conduct binary classification. It assumes that the binary outcome of a sample follows a Bernoulli distribution and has a population model as following:

In logistic regression model, we firstly conduct the exploratory data analysis to select features that has visually statistical significant influences on the outcome. Then we construct the model with the features and get rid of variables with p-values larger than 0.05, which means we failed to prove there is a statistical significant relationship between the feature and the label.

## 3.1 Exploratory Data Analysis (EDA)

The exploratory data analysis provides researchers a chance to explore the relationship between individual features with the label. The logistic regression model requires the feature to be linearly related with the label, which can be explored by the empirical logt plot. If there is a visually non-lineaer relationship between the feature and the label, we then need to conduct the linear transformation (usually convert the feature to log of the feature) to fix the non-linearity. However, if the non-linearity exist after the transforamtion, we could decide to choose another transformation function or choose not to use the feature.

### EDA (Part 1): Variable Type Analysis & Standardization

It is worth mentioning that the linear relationship is only required for numerical features, while the categorical features need mosaic plot to determine their relationship with the model. Therefore, we need to determine the type of the features and convert the defined-numerical features to categorical if they perform more similar to a categorical feature and vice versa.

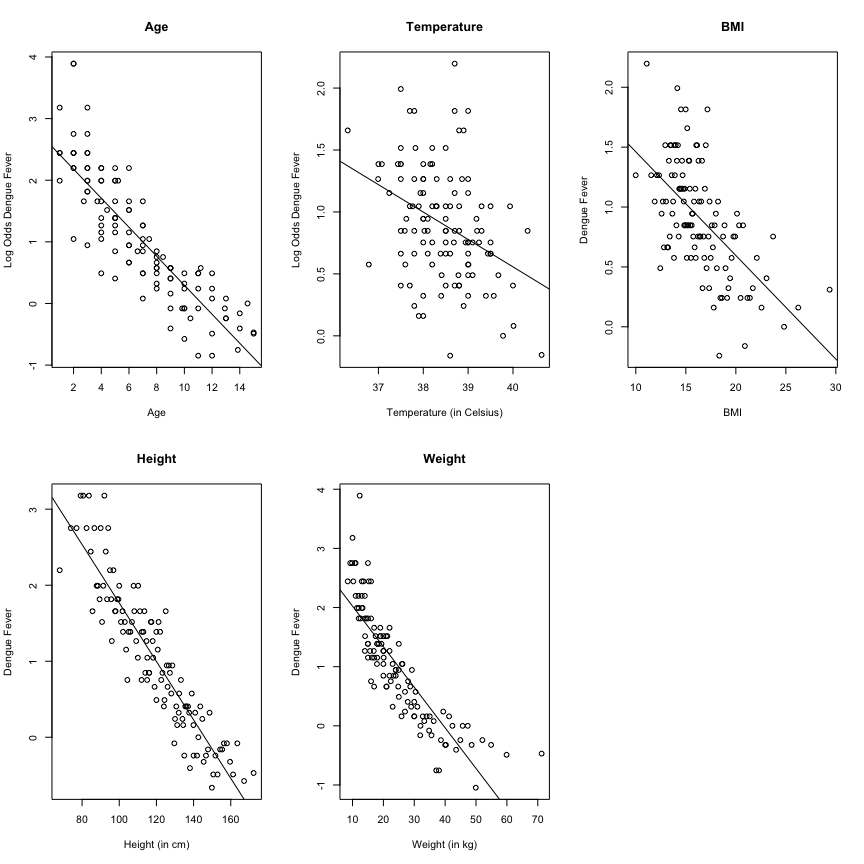
Feature Range

| feature | range begin | range end |
| --- | --- | --- |
| Sex | 1.000000 | 2.00000 |
| Age | 1.000000 | 15.00000 |
| DayDisease | 1.000000 | 3.00000 |
| Vomiting | 1.000000 | 2.00000 |
| Abdo | 1.000000 | 2.00000 |
| Muco | 1.000000 | 2.00000 |
| Skin | 1.000000 | 2.00000 |
| Temp | 35.800000 | 41.00000 |
| BMI | 8.503401 | 34.70986 |
| Height | 58.000000 | 176.00000 |
| Weight | 7.200000 | 91.00000 |
| Flush | 0.000000 | 1.00000 |
| Hepatomegaly | 0.000000 | 1.00000 |
| Rash | 0.000000 | 1.00000 |

From the feature range table, we noticed that, in the description of features, DayDisease is defined as a numerical rather than categorical variable. Therefore, we need to convert it to a categorical variable.

### EDA (Part 2): Empirical Logit Plots for Numerical Features

From the feature range of the data, we could see that the sex, DayDisease, vomitting, Abdo, Muco, Skin, Flush, Hepatomegaly, and Rash all have only 2 or 3 categories, indicating that these are categorical variables. Therefore, there are 5 remaining features that are numerical: Age, Temp, BMI, Height, Weight. We need to construct empirical logit plots for all of these features and explore if there is non-linearity existed.



The Figure 3.1.1 shows that the age, temperature, BMI, and height has a strong linear relationship with the log odds of dengue fever, indicating that these four features should be added to the model with no need for transformation.

However, the empirical logit plot between the log odds of dengue fever and the weight reflects a curved relationship, which is not linear, so it is necessary to perform the linear transformation to modify the relationship.

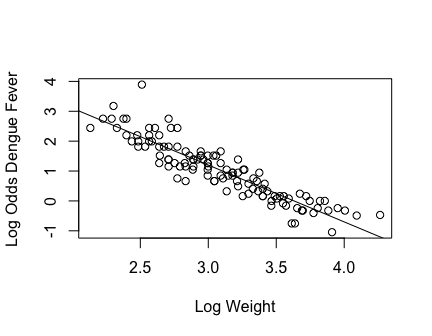
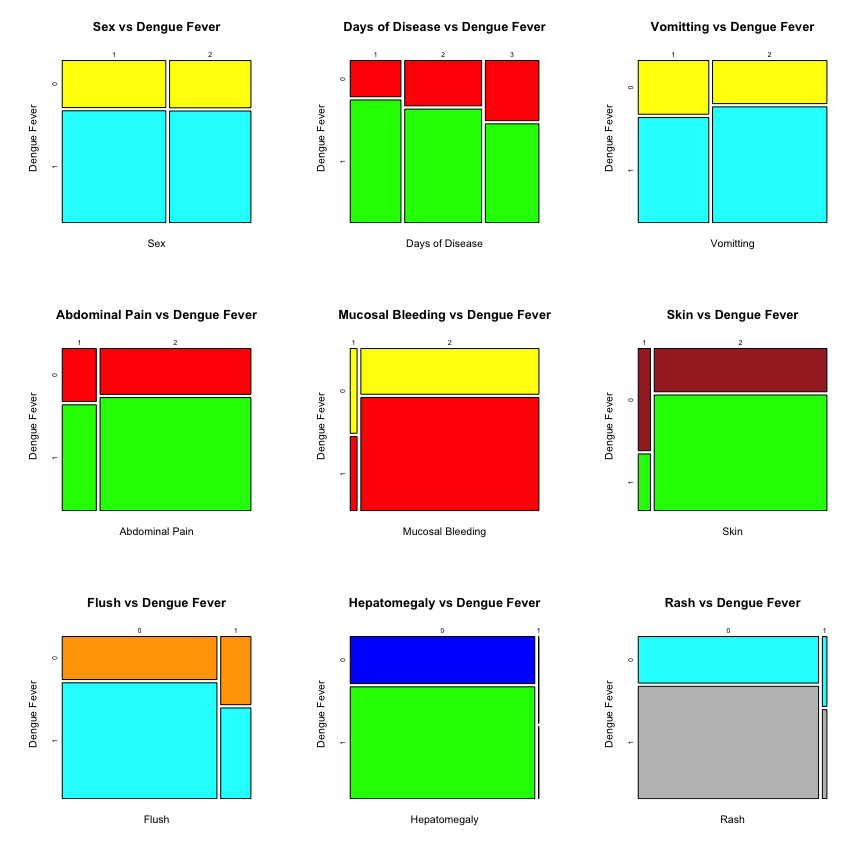


Figure 3.1.2 indicates that, the log weight is strong linearly related with the log odds of dengue fever. Therefore, we add the log weight into our model.

### EDA (Part 3): Categorical variables analysis

Along with the numerical variables, we use mosaic plot to select categorical features with statistically significant influences of the outcome.



From the Figure 3.1.3, we see that there is no relationship between the sex and dengue fever. Although there is an evident change in proportion of outcomes with different categories for hepatomegaly, one category dominates the entire feature, and therefore unlikely to make statistical significant influence on the outcome. Therefore, both the sex and hepatomegaly are excluded from the model.

## 3.2: Model building and training

In section 3.2, we will construct model with the variables selected from the EDA: DayDisease, Vomiting, Abdo, Muco, Skin, Flush, Rash, Age, Temp, BMI, Height, and log(Weight). We applied the train-test split method in this case to explore the prediction capacity of the model.

### Model 1 Construction & Coefficients Evaluation

Coefficients Table of Logistic Model 1

|  | Estimate | Std. Error | z value | Pr(>|z|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 10.912 | 1.696 | 6.434 | 0.000 |
| DayDisease2 | -0.343 | 0.082 | -4.205 | 0.000 |
| DayDisease3 | -0.610 | 0.086 | -7.064 | 0.000 |
| Vomiting | 0.287 | 0.067 | 4.268 | 0.000 |
| Abdo | 0.051 | 0.083 | 0.613 | 0.540 |
| Muco | 0.115 | 0.179 | 0.644 | 0.520 |
| Skin | 1.305 | 0.135 | 9.638 | 0.000 |
| FlushTRUE | -0.620 | 0.086 | -7.230 | 0.000 |
| RashTRUE | -0.390 | 0.206 | -1.888 | 0.059 |
| Age | -0.090 | 0.026 | -3.443 | 0.001 |
| Temp | -0.218 | 0.042 | -5.147 | 0.000 |
| BMI | 0.142 | 0.044 | 3.252 | 0.001 |
| Height | 0.031 | 0.014 | 2.255 | 0.024 |
| log(Weight) | -3.162 | 0.769 | -4.112 | 0.000 |

From the critical test, we could see that the P-value for Abdo, Muco, and RashTRUE are all over 0.05. This means we failed to testify that there is a statistically significant relationship existed between these features and the labels. Therefore, I recommend to remove these variables from the model. If the remaining variable still perform relatively same level of prediction capacity, then we should adapt the simpler model.

### Model 2 Construction & Coefficients Evaluation

Coefficients Table of Logistic Model 2

|  | Estimate | Std. Error | z value | Pr(>|z|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 10.870 | 1.681 | 6.465 | 0.000 |
| DayDisease2 | -0.348 | 0.081 | -4.268 | 0.000 |
| DayDisease3 | -0.618 | 0.086 | -7.181 | 0.000 |
| Vomiting | 0.293 | 0.066 | 4.455 | 0.000 |
| Skin | 1.367 | 0.125 | 10.955 | 0.000 |
| FlushTRUE | -0.647 | 0.085 | -7.639 | 0.000 |
| Age | -0.089 | 0.026 | -3.420 | 0.001 |
| Temp | -0.212 | 0.042 | -5.025 | 0.000 |
| BMI | 0.141 | 0.043 | 3.262 | 0.001 |
| Height | 0.031 | 0.014 | 2.267 | 0.023 |
| log(Weight) | -3.161 | 0.764 | -4.137 | 0.000 |

From the table, all features in model 2 has p-value less than 0.05, indicating that there is a statistical significant relationship between each feature and the label dengue fever.

### AIC Table of Model 1 & Model 2

AIC Table of M1 & M2

|  | AIC values |
| --- | --- |
| Model 1 | 5897.657 |
| Model 2 | 5896.046 |

From the table, the Model 2 has smaller AIC than the model 1. This indicates that, after removing all these variables,

## 3.3: Prediction & Result Evaluation

In KNN section, our result proves that the data size is large enough that the influence from the selection of observations does not influence the prediction capacity of the model. Therefore, we also apply the train-test split method in logistic model and make predictions with threshold of 0.5.

The model 2 predicts 192 patients who is not dengue and 953 person who is dengue. The total prediction numbers equals the observations in the test dataset. To determine the accuracy of the prediction, we construct the confusion matrix for analysis.

### Confusion Matrix of Logistic Regression Model (Model 2)

Confusion Matrix of Logistic Regression

|  | True Not Dengue | True Dengue |
| --- | --- | --- |
| 0 | 119 | 73 |
| 1 | 223 | 730 |

### Metrics Computation

# Section 4: Compare and Evaluate: Logistic/KNN

### Summary Table

| Summary Table of Metrics (Logisitc vs KNN) |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GMean | Sensitivity | Specificity | Accuracy |
| ———— | ——- | ————- | ————- | ———- |
| KNN | 0.554 | 0.865 | 0.353 | 0.713 |
| Logistic | 0.562 | 0.909 | 0.347 | 0.741 |

### Conclusion

From the summmary table, we could see that all four metrics of logistic regression model is higher than the KNN model. Therefore, we could conclude that the multivariable logistic regression model is more effective in predicting the dengue fever through the symptoms. Therefore, we should adapt the multivariable logsitic regression model (Model 2) as our final chosen model.