PROJECT FINAL - CAPSTONE

OSCAR GONZALEZ FRUTOS

June 22, 2020

Introduction

Dengue has become a public health problem, in Paraguay every year there is an epidemic outbreak called Dengue caused by the bite of a mosquito (Aedes aegypti). Dengue is a feverish viral disease, characterized by having a wide variability in its presentation, generating from asymptomatic processes to severe symptoms that can lead to death. This project aims to determine if climatic factors such as temperature, humidity, insolation, evaporation, atmospheric pressure and rainfall influence the level of infection of this disease, as well as another factor that may be related to the outbreak, the level of variation del Río of the area in question.

The actual data on dengue fever in Paraguay was obtained from the PAHO website https://www.paho.org/data/index.php/es/temas/indicadores-dengue/dengue-nacional/9-dengue-pais -ano.html whose report is weekly and accumulated by year, the daily levels of the Paraguay River were obtained from the DINAC website of the national government https://www.meteorologia.gov.py/nivel-rio/vermascalendario.php?estacion=2000086218&fechadesde=01-01-2017&fechahasta=31-12-2018 and the weather factors provided by the Polytechnic Faculty of Paraguay on its website as weather bulletins https://www.pol.una.py/?q=node/240, this project analyzed the available data for the years 2017 and 2018.

The analyzes of the corresponding data were carried out, firstly by downloading the data and then joining to have in a single database, cleaning and adaptation was performed to apply the algorithms of MAchine Learnig to predict the level of infection of this disease that contains two High and Low categories.

Analysis

As it was accumulated in the original base, a new variable is created to have the number of cases per week, and once that the cases variable is categorized, cases per week higher than 60 as high infection level, and less than 60 low level of infection thus leaving the response variables.

In exploratory data analysis, specific functions have been applied, to detect if there are any incomplete rows, missing quantities for variables and a function to detect if there is any empty element. With a bar graph the distribution of the response variable is observed, since it is what we are interested in predicting, showing a distribution of High 0.43 and Low 0.57 in the categories. With the car library, the behaviors of the variables are observed among themselves, detecting high correlations.

Several variables have been eliminated that do not contribute to the explanation of the model or because they are correlated between the predictor variables, it has been detected that the River Level variable is one of the variables that contributes the most to the model, that is, when the River begins to rise from level the density of the infection variable is higher, the dew temperature, humidity, and rainfall also behave in a similar way, but in a milder way with respect to the density.

With the caret library, the training observation indexes are created for the division of the base for training and test (datos_train and datos_test) obtaining a similar proportion regarding the distribution of the categories. With the function nearZeroVar (saveMetrics = TRUE) it is detected that all the variables do not have variances close to zero.

In the preprocessing of the data we use the recipes library to create a recipe () object with the response variable and the predictors, for this analysis all numerical variables are normalized and the qualitative Serotype variable is binarized, then the created recipe object is trained and apply to training and test bases with the bake function of the recipe package.

With the randomForest package, the variables that will be important when predicting the level of infection of the disease are detected a priori, resulting in the river level, dew temperature and humidity being the most important variables at the time to predict.

Models

After preprocessing the data, 6 models were trained giving the best result "C5.0Tree" in the training data with a proportion of correct classifications of 78%. To adjust and evaluate the models, the caret package has been used to evaluate multiple models with different subsets created from the training data, obtaining an estimate of the error in each repetition. Seeds have been established that will only be necessary if the reproducibility of the results is to be ensured, since cross-validation and bootstrapping involve random selection. The seeds are stored in a list with B + 1 elements where B depends on the validation method used:

To specify the type of validation, as well as the number of repetitions, a training control is created using the trainControl () function, which is passed to the trControl argument of the train () function. In this case, repeated cross-validation is used as a validation method.

Results

- K-Nearest Neighbor that has a hyperparameter K, has been trained for the values k = c (1, 2, 5, 10, 15, 20, 30, 50), the final value used for the model was k = 30, with an Accuracy of 0.7676984.
- Model for logistic regression: that it does not have hyperparameters with an Accuracy of 0.7269048.
- C5.0Tree: This model does not have hyperparameters but it is one of the models that has given the best result with an Accuracy of 0.7792063.
- Random Forest: This model has the following hyperparameters mtry = c (3, 4, 5), min.node.size = c (2, 3, 4, 5, 10, 15, 20, 30), splitrule = "gini"), it has been trained with the different hyperparameters assigned, giving as best for mtry = 5, splitrule = gini and min.node.size = 2 with an Accuracy of 0.7501984.
- Support vector machine (SVM): This model has the following hyperparameters sigma = c (0.001, 0.01, 0.1, 0.5, 1), C = c (1, 20, 50, 100, 200, 500, 700) giving the best results for sigma = 0.01 and C = 200 with an Accuracy of 0.7426190.
- Neural networks (NNET): This model has the following hyperparameters size = c (10, 20, 50, 80, 100, 120), decay = c (0.0001, 0.1, 0.5), it has been trained for the different hyperparameter values assigned giving the best result for the combination of size = 120 and decay = 0.1 with an Accuracy of 0.7415476.

Summary of Accuracy and Average Kappa values in training data.

Models	Accuracy	Kappa
1 C5.0Tree	0,779	0.550
2 KNN	0,768	0.533
3 Random Forest	0.750	0.485
4 SVMradial	0.743	0.473
5 NNET	0.742	0.467
6 Logistic	0.727	0.446

Data test

We evaluated the final model using the data from "datos_test_prep" that did not participate in the training of the model, the function extractPrediction () was used. This function returns a dataframe with the predictions of each of the models, both for training and test observations. Furthermore, it shows the true value of each observation.

Comparison of results

Models	Test	Training
1 KNN	0.833	0.757
2 Random Forest	0.833	1
3 C5.0Tree	0.8	0.878
4 Logistic	0.767	0.784
5 NNET	0.767	0.784
6 SVMradial	0.767	0.838

Conclusion

The Random Forest model is the one that obtains the best results taking into account the accuracy metric both in the test set and in the validation (repeated CV). The remaining models achieve very similar test values.

For a later study, it is recommended to obtain more data on Dengue disease, such as the history of the index of larval infestation of the mosquito (Aedes aegypti), this work had many limitations in terms of obtaining data, it could only be completed for the periods 2017 and 2018, the impact generated by this study is that the behavior that the outbreak of the epidemic can have can be predicted, taking into account variables such as the level of the Paraguay River, the Type of the disease (Serotype), and climatic factors .