Predicting Local Epidemic Spread of Dengue Fever in San Juan and Iquitos by Using Machine Learning Methods

BRENO ABERLE JAVIER PEÑA
breno | javierpr @kth.se

October 6, 2020

Abstract

Dengue fever causes thousands of deaths every year and the number of new infections is steadily increasing. Underdeveloped regions with a lack of medicine and know-how are affected the most. These regions have a deficiency in the health care system and therefore outbreaks result in a collapse in the emergency system. Consequently, helping institutions are overwhelmed and cannot handle the situation. Therefore, we want to predict the number of dengue cases reported each week in San Juan, Puerto Rico, and Iquitos, Peru. Based on our predictions, authorities can prepare for the outbreak to actively contain the spread. They can allocate medicine, equip hospital facilities and organize auxiliary staff. With these predictions, we can have a positive impact on society by addressing ethical and sustainability issues. Additionally, fewer people could get affected and consequently the living quality in the region will be improved. As a result of our analysis, Support Vector Regression outperformed Random Forest with a slightly lower mean absolute error. Additionally, the most important features are week of the year, temperature, rainfall and vegetation.

Contents

1 Introduction

1.1 Aim, Objective and Goals

Dengue is a severe illness that causes thousands of deaths and infections per year and is increasing steadily [1]. Its spread depends heavily on climate [2] because temperature affects the physiology of the Aedes aegypti (mosquito vectors). The aim of this project is to predict the number of dengue cases reported each week in San Juan (Puerto Rico) and Iquitos (Peru), by using meteorological and vegetation data. Since the predictions rely on the weather forecast, we can only make predictions as far ahead as weather forecast data is available. The reason why we chose two cities instead of focusing on just one, is that, in theory, the causes of dengue to spread should be the same in both cities, when the optimal conditions are met. This means that building a model with two cities implies having more instances and information to build the model. Nevertheless, in section 4 this hypothesis will be tested.

Additionally, these two cities are well known for having many dengue cases and they both have deficiencies in the health care system, since, during outbreaks, the emergency system collapses. Being able to predict this information will help the sanitary professionals to get ready for treatment measurements and anticipate the outbreak [3]. Based on our predictions people in charge can take precautions, for instance, getting more medicine, vaccination, hospital facilities and organize the appropriate number of staff.

The goal of this project is to use Random Forest (RF) and Support Vector Regression (SVR) in order to predict the number of dengue cases in a certain week. Furthermore, based on the outcome of our analysis we want to determine which are the most important factors for the predictions in order to understand the causes better.

1.2 Background and Rationale

1.2.1 Dengue fever

Dengue is a mosquito-spread disease that has rapidly spread throughout the tropics due to rainfall, temperature and unplanned fast urbanization. We can see a close relation between climate and dengue in places like Puerto Rico where dengue's rate increases during the rainfall season from May to November [4].

Dengue spread has grown drastically in recent years. Estimating the number of cases is complicated because some are asymptomatic. According to The World Health Organization, one estimation reflects 2.2 million infections in 2010 and 3.34 million in 2016, and 2.5 billion people at risk of infection. Not only there has been an increase in the number of infections, but also the number of countries affected. Being only 9 in 1970 and now is considered an epidemic in 60 countries. Studies on dengue fever estimates 25 thousands deaths and about 390 million infections per year [1].

One of the biggest problems with dengue is the lack of a vaccine for prevention or treatment. The only available treatment is to visit the doctor, get supportive measures like pain killers and fluid therapy. For this reason, during epidemics, hospitals get collapsed with a huge number of patients. This, with the fact that dengue is present in countries with limited resources, makes it difficult for doctors to give a personal service to every person.

The current methods that have been used to reduce dengue spread include removing breeding places, use larvicides and insecticides. But these techniques have been rather inefficient as mosquitoes can breed even in tiny spaces with water such as between the trunk and the leaves of plants with large foliage [5].

Currently, there is a great interest in the study of the relationship between climate and vegetation with dengue. For instance, there are some studies that suggest that the geographic area where dengue fever is present, can be modeled only using average vapor pressure [6]. Another study carried out in Colombia from 2001 to 2010 [2] claimed that air temperature and precipitation have a huge influence in dengue, to be precise, in the cities where the study took place, when the temperatures were higher after rainfall the number of cases increased [2]. Vegetation density (NDVI) also was taken into consideration in an study carried out in Costa Rica [7], showing that there is an inverse relationship between the number of dengue cases, and the vegetation density since dengue is more prominent in cities, where vegetation is lower. Nevertheless,

the most significance relationship between dengue and vegetation is during the dry season where plants store water in their leaves, where mosquitoes can breed. Also, the population habits affect mosquitoes since during dry season they develop in water containers or discarded tires [8] [7].

1.2.2 San Juan and Iquitos

Iquitos is a city located in the north of Peru, known as the capital of the Peruvian Amazon. Dengue outbreaks occur from October to April. There are three seasons in Iquitos [9]. In the first, temperatures are warm, rainfall is elevated, the level of the Amazon river is increasing and dengue cases start to descend. In the second, conditions are relatively cooler and drier, the river begins to subside, and there are not many dengue cases [9]. In the third, temperatures are their warmest and precipitation increases, the river subsides to its lowest levels, begins to rise again, and dengue transmission picks up [9]. Being able to predict the number of cases per week will be especially helpful for this city as the emergency services usually collapse during outbreaks [10]. Moreover, the location of Iquitos makes it the largest city in the world that cannot be accessed by road [11], this means that in the case of an unplanned outbreak the access to extra medicines will be limited.

San Juan is Puerto Rico's capital, the rainfall season is produced from May to November when the majority of cases are reported [12]. it has suffered recent epidemics in the years 1998, 2007, 2010 [12]. In 2010 27,000 infections and 128 deaths were reported [12]. A study conducted from 1979 to 2005 in Puerto Rico confirmed that precipitation and temperature have a big influence on the number of dengue case [13].

1.2.3 Machine Learning

Support Vector Regression (SVR) is a supervised learning model for regression analysis. The strength of SVR is the kernel trick. With an appropriate kernel function, it is possible to solve complex problems in high dimensional spaces. It also has a low risk of over-fitting due to generalization, and is relatively memory efficient [14].

Random Forest (RF) is also used for classification and regression problems. RF usually can achieve high performance and accuracy. The results often tend to be better than polynomial regression and are competitive to Artificial Neural Networks (ANN). Random Forest handles missing values while maintaining the accuracy of large data. It is also able to handle large datasets with high dimensions and is therefore good at learning complex, highly non-linear relationships.

1.3 Ethics & Sustainability

This project aims to predict the number of dengue cases in a specific week of the year, using a dataset from various U.S. Federal Government agencies. This means that we will not use participants for carrying out this project.

This research project will also address some ethical and sustainability issues. Dengue fever is an illness that is generating millions of infections per year. A bigger effort is needed to prioritize research and financial resources [15]. This last point is vital, as dengue is present in poor countries with inefficient public health care infrastructure. Furthermore, dengue has a high disease burden, as patients infected with dengue require hospitalization. Using disability-adjusted life years (DALY) [16] to estimate the number of years lost due to disability, delivers an estimation of 2.153 DALYs/million people in 1994 in Puerto Rico [16]. Nowadays the illness is wider spread, we can assume that this number will be greater.

Another important fact is that DALY for dengue, only in Puerto Rico, is the same as malaria or tuberculosis in the whole of South America [16]. That shows that dengue is a severe illness and governments as wells as international organizations should pay attention to it.

Within our sustainability issues, we are approaching 2 of the goals proposed by the United Nations [17].

Goal 3, Good health and well-being: If we can foretell the spread, we can take precautions in advance. For instance, if we know that the number of cases are going to increase, we can obtain more medicines and hire more doctors for that specific period.

Goal 9, Industry, innovation, and infrastructure: The aforementioned algorithm can help the region to start using innovating systems to be more efficient.

2 Theoretical framework

There exist several research papers that apply Machine Learning models in order to predict the dengue fever outbreak in specific regions. They have mostly similar temperatures, vegetation and humidity conditions. Furthermore, these papers have a high importance for our academic paper because their structure is similar to ours. They provide us indications on how to conduct our research. The papers provide background about conditions in the regions and the impact of dengue fever. Moreover, they make use of several Machine Learning methods in order to compare the results. Based on the results, we can identify which methods are suitable for that kind of prediction. These may be a good starting point for our research.

In the paper 'Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach' from the year 2011 the authors Yuhanis Yusof and Zuriani Mustaffa introduced a prediction model that incorporates Least Squares Support Vector Machines (LS-SVM) in predicting future dengue outbreak in Selangor, Malaysia. The goal is to achieve high accuracy in prediction. That is important in order to plan precaution measures. The problem is that different training and learning techniques approaches may lead to different prediction accuracy. Therefore the suitability of different Machine Learning methods needs to be assessed and compared [18]. In Malaysia there exists crucial public health concerns. Dengue fever is widely spread in that country and the number of cases increases steadily. Under certain conditions including the right temperature and humidity, dengue fever spreads out easily. Furthermore, global warming and unpredictable rainfalls are also important factors. The dataset contains data on dengue cases and rainfall levels collected in five districts in Selangor. The research method is to apply Machine Learning methods on a dataset to make predictions of future dengue cases. As part of the research, the two Machine Learning methods Least Squares Support Vector Machines (LS-SVM) and Artificial Neural Network (ANN) are applied, assessed and compared. The LS-SVM prediction model outperformed the Neural Network model in terms of prediction accuracy and computational time. One reason is that the structure of ANN is very complex and needs many tuning parameters. Moreover, it is difficult to select the network architecture and determine the number of hidden neurons. Additionally, the learning speed of ANN is slower and gets easily stuck in local minimum. LS-SVM, on the other hand, applies Structural Risk Minimization (SRM) and improves both training time and accuracy. Several numbers of factors like humidity, temperature and cloudiness have influence on the dengue outbreak prediction. The determination of tuning parameters is also an important part of the research work [18].

In the research article 'Developing a dengue forecast model using machine learning: A case study in China' from 2018, the authors compare the predictive accuracy of the temporal pattern of Dengue incidence in Metropolitan Manila, Philippines, influenced by meteorological factors from four modeling techniques. The applied methods are General Additive Modelling, Seasonal Autoregressive Integrated Moving Average with exogenous variables, Random Forest and Gradient Boosting. The goal is to not only use one statistical method but use several and compare them to each other. The dataset can be categorized into the two types observed meteorological factors and its corresponding delayed or lagged effect. It includes data about dengue incidences and meteorological data of Metropolitan Manila from January 1, 2009, to December 31, 2013, and that data is retrieved from respective government agencies of the Philippines. Meteorological data includes information about floods, precipitation, temperature, southern oscillation index, relative humidity, wind speed, and direction. The predictive accuracy and importance of features were calculated and evaluated. The analysis showed that Random Forest had the best predictive accuracy. The best dataset to use was the delayed or lag effects of the meteorological variables. Furthermore, relative humidity, rainfall and temperature are important meteorological factors [19].

3 Hypothesis and Research Question

The project is data-driven. We will analyze a dataset and try to predict future development by using two Machine Learning methods (SVR and RF), and all the information mentioned in section 1.1 that help to predict dengue spread.

Based on the research in section 1.2 about dengue fever, the most relevant factors for dengue fever spread are temperature, rainfall and vegetation. An objective of this paper is to investigate the following hypothesis: are temperature, climate and vegetation as important factor as we expect to predict dengue spread?

Another hypothesis that is going to be investigated is related with the dataset. Two approaches are possible, as we have one dataset with two cities, we have the option of training one model with the whole dataset, or train two different models with each city. The hypothesis is that the model trained with the whole dataset will perform better than the two separate models due to the lager number of instances.

4 Research Methodology

Our goal is to predict the local epidemic dengue fever outbreak in San Juan and Iquitos and to determine the most important factors. For this purpose we will use SVR and RF. The analysis of similar academic papers (see section 2) confirms the suitability of this choice. Their research has shown that SVR, as well as RF, have performed with high accuracy compared to other methods. It makes sense to use two methods and compare the results with each other. It is not possible to determine that one specific method is the most suitable one for this research topic since the methods behave differently in different situations. Identifying a few suitable methods and comparing the results is, therefore, a reasonable approach.

4.1 Data Collection, Analysis and Pre-processing

The first step of our project was to get the data. The dataset was obtained from various U.S. Federal Government agencies*. Dengue surveillance data is provided by the "Center for Disease Control and Prevention" (CDC) in collaboration with the Peruvian government, and the environmental and climate data is provided by the "National Oceanic and Atmospheric Administration" (NOAA). Regarding the source, we can conclude that this dataset is trustworthy for our purpose. This data includes parameters mentioned in section 1.2 as crucial factors for dengue spread. The features include:

- Week, year (from 2000 to 2010 for Iquitos and from 1990 to 2010 for San Juan) and city where the data is taken.
- Climatic information measured in local weather stations during that week (Maximum temperature, Minimum temperature, Average temperature, Total precipitation, Diurnal temperature range).
- Satellite measured precipitations and NOAA's weather forecast (Total precipitation, Mean dew point temperature, Mean air temperature, Mean relative humidity, Mean specific humidity, Total precipitation, Maximum air temperature, Minimum air temperature, Average air temperature, Diurnal temperature range).
- Vegetation index, this information is measured with the normalized difference vegetation index (NDVI). It is a graphic indicator used to measure the amount of green areas in satellite pictures. This information in our dataset is measured from the southeast, southwest, northeast and northwest from the city centroid.

The dataset contains 1457 instances (937 from San Juan and 520 from Iquitos). To carry out the analysis of the features, first, we plotted some key attributes to see how they distributed overtime. These are the features that, according to the investigations explained in section 1.2, are the most informative for dengue

* Dengue surveillance, weather forecast and vegetation data: https://dengueforecasting.noaa.gov/

prediction (temperature, precipitation, and humidity). The chosen time period is 2005, 2006 and 2007 in San Juan. As we can see in Figure 1, the number of dengue cases is seasonal. The outbreaks tend to occur in the second half of the year, as explained in section 1.2.2, dengue outbreaks happens from the mid to end of rainfall season, from May to November.

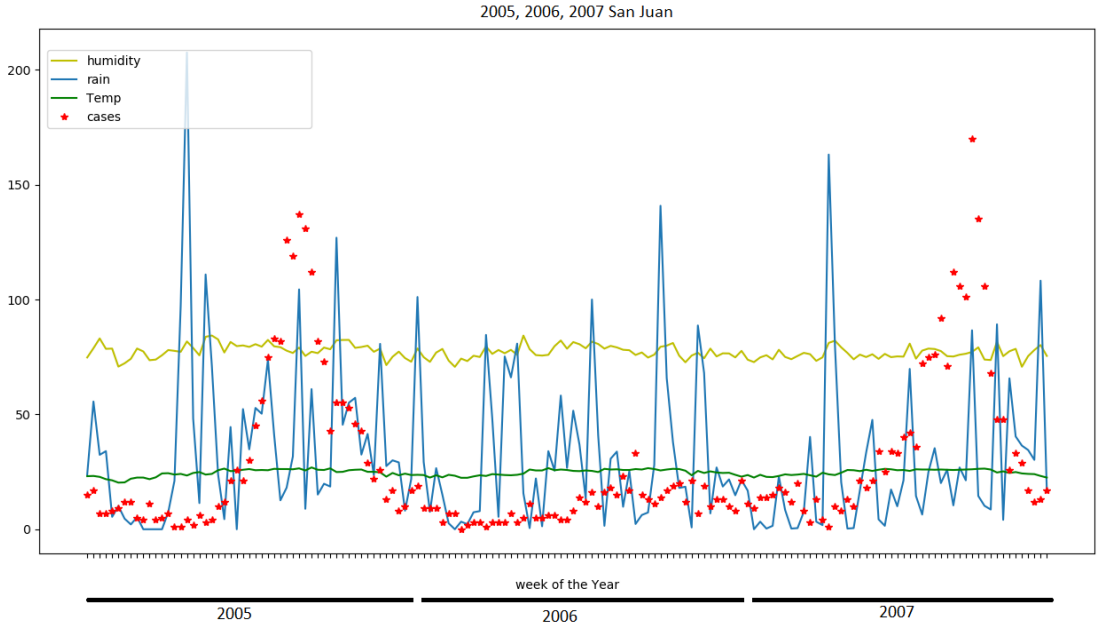

Figure 1: Number of dengue cases per week in 2005, 2006, 2007 in San Juan, Puerto Rico with temperature, humidity and precipitation

The next step is to analyze the relationship between the features themselves. For this purpose, we checked the feature correlation. With this information, we can control if one attribute depends on another. After checking all the attributes, we found out that the features showed in Figure 2 are highly correlated. This means that these attributes do not contribute with any extra information while increasing the complexity of the model because of a higher number of features.

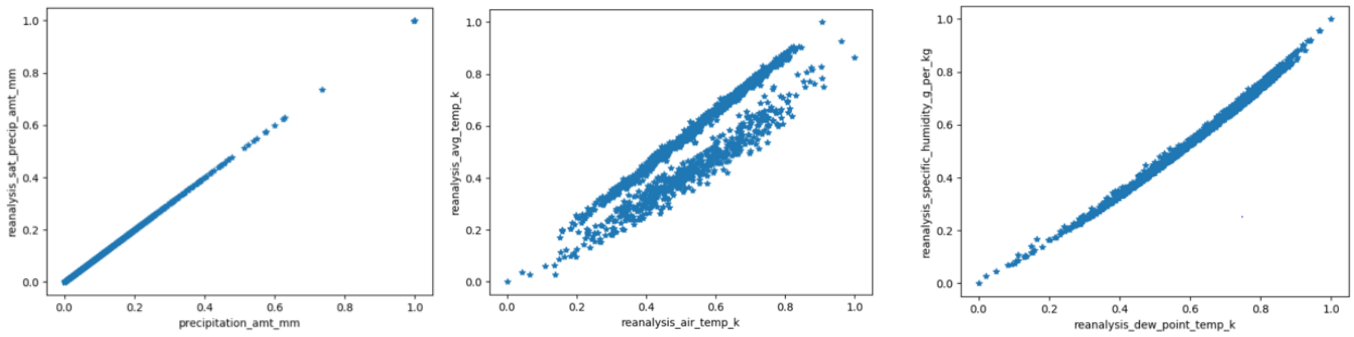

Figure 2: From left to right. Total precipitation measure in weather station and total precipitation forecasted by satellite. Mean air temperature forecasted and average air temperature in Kelvin. Dew point temperature and Mean specific humidity. These are samples taken out of A.1.

For doing the data preprocessing, first, the highly correlated features were discarded, as they do not carry any additional information. Dropping them will reduce the dimensionality and, as a consequence, make the model simpler. Also, imputation was applied by filling missing values with the mean of the column. At this point, we had to decide to use one model per city or train a model with both cities' data. For this reason,

we kept the original dataset and applied One-hot encoding to the categorical feature that represents the city. With this data, we created a model. Besides, we split the dataset into two, having in each one the data of each city. With this, we created one different models per city.

4.2 Modelling approaches

As explained in section 4.1, two different approaches will be taken into consideration, first, we will train a model with the whole dataset and then we will create one model per city. In both of the cases, we will use a 5-fold cross-validation. The source code for the applied ML models can be accessed through the link provided in appendix A.3.

Random Forest

RF is a model combining many decision trees into one model. The problem with Decision Trees is the sensitivity to training data. If the training data gets changed, the resulting Decision Tree can be significantly different and as a consequence the prediction also deviates. Furthermore, Decision Trees tend to overfit and get stuck in local minima. Whereas, RF uses ensemble learning methods combining multiple trees to solve regression and classification problems. Ensemble learning combines the prediction of multiple machine learning methods in order to make more accurate predictions than individual models. Random Forest uses the ensemble learning algorithm Bootstrap Aggregation, also called Bagging. Bagging involves random sampling of small subset of the dataset with random replacements. As a result, the variance gets reduced. It is very important that each Decision Tree is independent. The trees in RF are run in parallel and independently. The RF Regressor will output the mean prediction of the individual trees.

RF is created using the sci-kit learn library in python. Three models were created, one containing the whole dataset with 120 estimators and max depth of 500, one model containing only San Juan data with 75 estimators, and one last model trained with Iquitos data. To prevent overfitting and due to the small dataset in the single-city model, we used a 5- fold cross-validation. In Figure 3 is shown the different way the models fit the test data.

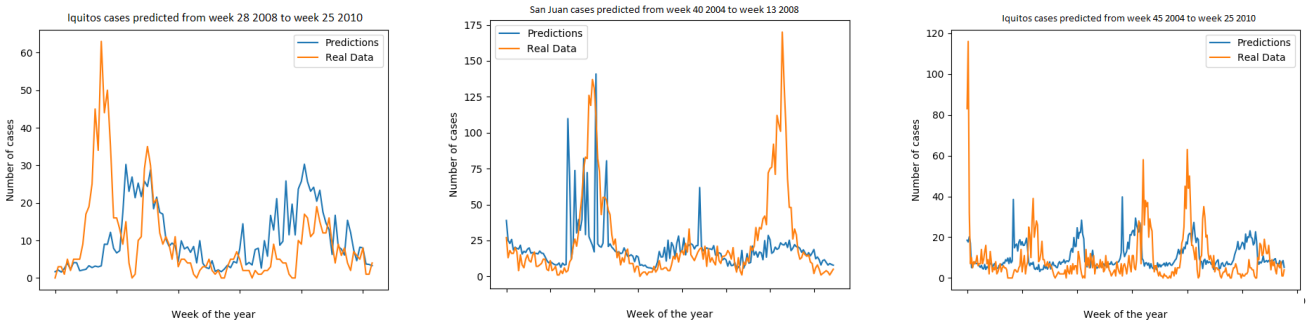

Figure 3: From left to right. Prediction made with the model trained just with Iquitos data. Predictions made with model trained just with San Juan dataset. Predictions made with model trained with the whole dataset (Iquitos + San Juan).

Support Vector Regression

To implement the Support Vector Regression (SVR) model we used the scikit-learn library*. The main idea is to find the most suitable hyperplane representing the data within a decision boundary. The hyperplane will help us predict the continuous value. We are trying to determine a decision boundary with a margin of tolerance ε to the original hyperplane. The best hyperplane maximizes the amount of data points within a specified threshold. The Support Vectors are the points closest to the boundary [14].

When the model is created we can modify parameters in order to improve the predictions of our model. Our margin of tolerance ε determines how wide the margin of the decision boundary is, and makes sure

* Scikit-learn, Support Vector Regression: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

that only the points with least error rate are considered. So ε is one parameter we can adjust. We can decide if we want to allow a higher or lower error rate. We need to take into account, that the model tends to overfit and generalizes poorly the more outliers we involve. On the other hand, if we set a very low error rate, the model may not be really representative because many important data points are not taken into consideration. Therefore, the performance also decreases. As a result, we should choose the parameters carefully [20]. C is a penalty parameter of the error term. With a higher value of C we have a smaller margin. Consequently, with a smaller value of parameter C the margin gets larger. SVR uses a kernel, a function used to map a lower dimensional data into a higher dimensional data. The incentive is to make use of the kernel trick, because you can use this method to apply a linear classifier to non-linear classifiable data. With the SVR library of scikit-learn we can select the linear, polynomial, sigmoid and radial basis function kernel. Depending on which kernel we choose we get different accuracies.

4.3 Evaluation

As evaluation metric for our results we have chosen Mean Absolute Error (MAE). MAE returns the difference between the predicted and observed values. The equation is illustrated in the following:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

While n is the number of instances, y_i the observed value and $\hat{y}_i$ the predicted value. MAE and Mean Squared Errors (MSE) are common evaluation metrics for regression models. The advantage of MAE is that we get the absolute difference of predicted and observed value which is the wrongly predicted amount of Dengue cases. In the case of this research project, MAE is more suitable. MSE makes in another context sense, for instance differentiability problems.

5 Results and Analysis

We expect to identify some factors which have an important impact in predicting the dengue fever outbreak. These factors could be humidity and temperature as spotted in section 1.2 as crucial factors for mosquitoes to breed and dengue to spread. Moreover, we expect to have predictions which are close to the observed data.

Random Forest

To decide which model performs the best, we are going to calculate 20 times the MAE for the three models and do the error average to obtain a reliable result. In Figure 4 we can observe the different MAE each model scores.

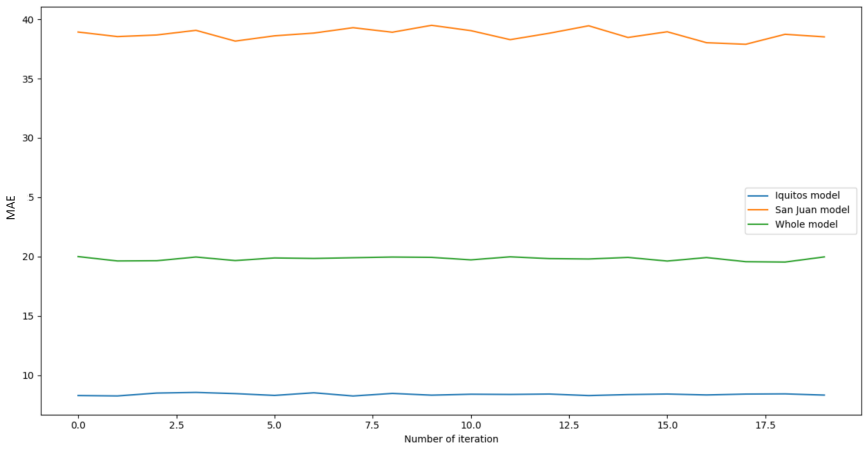

Figure 4: Model performance comparison.

As we can observe the best performing model is the one trained only with Iquitos dataset, with an average MAE of 7.56. The second-best model is the one created with the whole dataset, scoring 19.64. The third is the model just created with San Juan data, which scores 38.43. Taking these numbers into account we can conclude that there is not big difference between using one model for both cities or using a model per city, as Iquitos model outperforms the whole model but this last outperforms San Juan's model. For simplicity reason to compare with SVR we choose the model trained with the whole data.

Support Vector Regression

The best result could be achieved with $\varepsilon = 0.1$, $C = 0.5$, $Kernel =$ "*polynomial*". Figure 5 displays the predicted number of dengue cases for a SVR model with polynomial degree of 4 and 5 as well as the actual test labels.

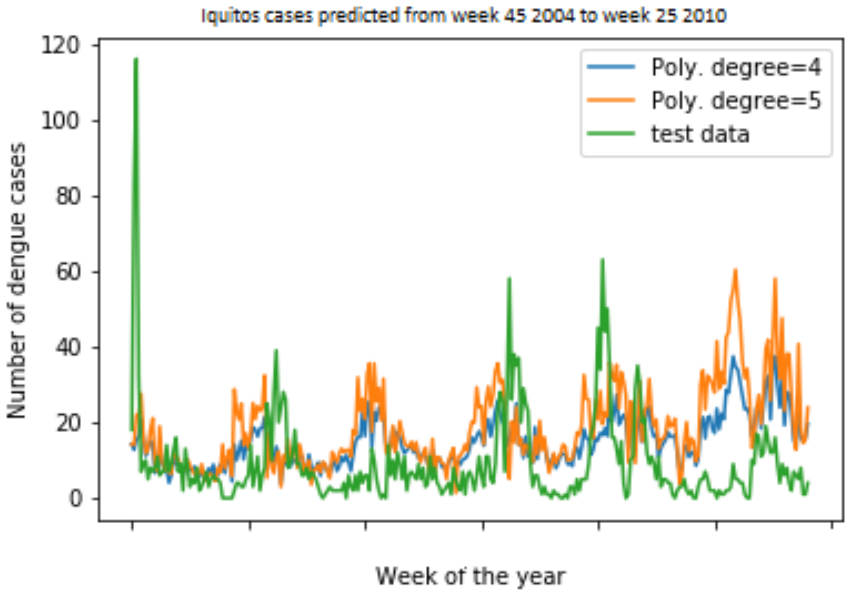

Figure 5: Predicted number of dengue cases with Support Vector Regression and different polynomial kernels.

The mean average error (MAE) of the SVR model with polynomial degree of 4 is 18.38 and with the degree of 5 is 19.19. Moreover, the predictions of the SVR model with polynomial degree of 4 also predicts the number of the dengue cases closer to the test labels.

5.1 Conclusion

Based on the results in section 5 the SVR model outperformed the RF model. With a MAE of 18.38 SVR is slightly better than RF with a MAE of 19.64. SVR is known to perform well for these regression tasks. Thus, it is not surprising that it performed here also well. Moreover, it is worth mentioning that hyperparameter tuning has also a considerable contribution to the result. In SVR we have great flexibility of adjusting the parameters. Besides the margin of tolerance ε and the penalty parameter of the error term C, we can also modify the kernel. For each kernel, we have additional and different parameters to tune. For instance, for the polynomial kernel we can also change the polynomial degree. In RF we adjust the depth of trees, number of trees, the minimum number of splits at each node and if bagging is included or not. It could be that out of that huge variety of possible parameter constellations not the optimal was chosen. A MAE of 18.38 tells us that the predictions are very likely to have an error. Since the MAE is low it is not that severe. When authorities are preparing precautions based on our predictions they should put that error into consideration. Therefore, it makes sense to organize a few more additional medicine. It is better to overestimate the prediction. Leftover medicine can be stored and used in the future. But in case there is not enough medicine that could have a severe impact.

As a hypothesis we formulated that climate aspects will have a big impact on the prediction of dengue cases, this statement is proven right if we check the feature importance determined by random forest in A.2 where we can see that the most determinant feature is vegetation (ndvi_nw and ndvi_sw). Moreover, we can identify how an also important feature is the week of the year since dengue is a seasonal illness. Temperature is also considered crucial to determine the output, this makes sense, as explained in section 1.2.1 temperatures are higher after the rainfall season when the majority of cases occur. However, precipitations are not as important as expected. Another remarkable fact is how unimportant the cities are, the reason for this may be that the conditions for dengue in both cities are considerably similar, this information also supports the idea of using the whole dataset to train the model instead of using two different models.

Another formulated hypothesis is that the model trained with the whole dataset would perform better than the two separate models, this hypothesis is not supported in figure 4 where we can see that the model trained with Iquitos data performs considerably better than the one trained just with San Juan data. However, in between, we can see the error of the model trained with the whole dataset. The reason for this could be that the number of errors for San Juan is higher than the one for Iquitos, additionally, the number of samples from San Juan (937) is higher than the number of instances from Iquitos (520), what makes the whole model error balance in the middle. The good results for Iquitos balances the bad results of San Juan, making the whole model an average between both individual models. This makes us decide that both approaches are equally precise, also the feature importance graph in A.2 supports this approach, since knowing the city does not provide much information. For simplicity reason to compare with SVR we choose the model trained with the whole data.

We have seen that we can predict dengue spread in a reliable way with a low MAE. Authorities can take our predictions into consideration while keeping in mind that the predictions contain errors. Therefore, it should be carefully examined how much additional medicine and other resources are going to be allocated. For further research, additional ML methods may be applied. It also make sense to apply Neural Networks to make predictions. Due to the limited time, applying more models would exceed the scope of this paper.

5.2 Limitations

A bad dataset can lead to poor results. Since our data is from governmental organizations we assume the quality of our data is high. But that does not mean that all important information is included. In addition, a high-quality dataset does not mean that all information is useful for our research.

The literature review has shown that some Machine Learning methods return good results and outperform other methods in prediction accuracy and computing time. We conducted research on the respective Machine Learning methods and in which context they should be used. As a result, we chose Random Forest and Support Vector Regression. Maybe there exist more suitable algorithms but we decide for them based on our research. Furthermore, our model could be bad. In order to improve the results of our model we used hyperparameter tuning.

One severe risk to our research could be that we have not considered all important factors. Based on the literature review in section 2, we identified important factors like precipitation, temperature and vegetation. Their analysis has shown that these factors had a great impact to the results. The feature importance of our analysis has shown that the mentioned factors are also important for San Juan and Iquitos. Moreover, we included a meaningful amount of features provided by our dataset. Nevertheless, we can not say for sure that we did not miss any factor, because the circumstances differ from location to location.

It is impossible to make a completely accurate prediction and the variance of the result can also differ strongly. Further, a wrong prediction can have a great impact on the situation. If the wrong place is predicted resources will be allocated to the wrong location. People on other locations will suffer because they are in need of treatment and are missing medicine and professional help. Additionally, the spread of the disease cannot be contained. The same problem applies for the wrong prediction of time.

Moreover, there can also arise general problems that are associated with the usage of Machine Learning methods. For instance, with Machine Learning methods it is very difficult to track what actually happened to have a certain output.

References

[1] W. H. Organization, "Dengue and severe dengue," 2019, accessed October 09, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[2] A. Meza-Ballesta and L. Gónima, "The influence of climate and vegetation cover on the occurrence of dengue cases (2001-2010)," *Revista de salud pública (Bogotá, Colombia)*, vol. 16, pp. 293–306, 04 2014.

[3] R. Lowe, A. Stewart Ibarra, D. B. Petrova, M. Garcia-Diez, M. J. Borbor-Cordova, R. Mejía, M. Regato, and X. Rodo, "Climate services for health: predicting the evolution of the 2016 dengue season in machala, ecuador," *The Lancet Planet Health*, vol. 1, pp. e142–e151, 2017. doi: 10.1016/S2542-5196(17)30064-5

[4] M. Jury, "Climate influence on dengue epidemics in puerto rico," *International journal of environmental health research*, vol. 18, pp. 323–34, 2008. doi: 10.1080/09603120701849836

[5] S. Rajapakse, C. Rodrigo, and A. Rajapakse, "Treatment of dengue fever," *Infection and drug resistance*, vol. 5, pp. 103–12, 2012. doi: 10.2147/IDR.S22613

[6] S. Hales, N. Wet, J. Maindonald, and A. Woodward, "Potential effect of population and climate changes on global distribution of dengue fever: An empirical model," *Lancet*, vol. 360, pp. 830–4, 10 2002. doi: 10.1016/S0140-6736(02)09964-6

[7] A. Troyo, D. Fuller, O. Calderón-Arguedas, M. Solano, and J. Beier, "Urban structure and dengue incidence in puntarenas, costa rica," *Singapore Journal of Tropical Geography*, vol. 30, pp. 265 – 282, 07 2009. doi: 10.1111/j.1467-9493.2009.00367.x

[8] R. Barrera, M. Amador, and A. Mackay, "Population dynamics of aedes aegypti and dengue as influenced by weather and human behavior in san juan, puerto rico," *PLoS neglected tropical diseases*, vol. 5, p. e1378, 12 2011. doi: 10.1371/journal.pntd.0001378

[9] S. Stoddard, H. Wearing, R. Reiner, A. Morrison, H. Astete, S. Vilcarromero, C. Alvarez, C. Asayag, M. Sihuincha, C. Rocha, E. Halsey, T. Scott, T. Kochel, and B. Forshey, "Long-term and seasonal dynamics of dengue in iquitos, peru," *PLOS Neglected Tropical Diseases*, vol. 8, 2014. doi: 10.1371/journal.pntd.0003003

[10] G. Muñoz, "Hospitales de iquitos colapsan ante epidemia de dengue," 2011, accessed November 28, 2019. [Online]. Available: https://larepublica.pe/archivo/515000-hospitales-de-iquitos-colapsan-ante-epidemia-de-dengue/

[11] D. MacFarlane, "Iquitos, peru: The largest city in the world that can't be reached by road," 2018, accessed November 28, 2019. [Online]. Available: https://weather.com/travel/news/2018-12-05-iquitos-largest-city-not-accessible-by-road

[12] D. Noyd and T. Sharp, "Recent advances in dengue: Relevance to puerto rico," *Puerto Rico health sciences journal*, vol. 34, pp. 65–70, 2015.

[13] M. Jury, "Climate influence on dengue epidemics in puerto rico," *International journal of environmental health research*, vol. 18, pp. 323–34, 11 2008. doi: 10.1080/09603120701849836

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013. ISBN 9781441931603

[15] A. Wilder-Smith and P. Macary, "Dengue: Challenges for policy makers and vaccine developers," *Current infectious disease reports*, vol. 16, p. 404, 2014. doi: 10.1007/s11908-014-0404-2

[16] M. Meltzer, J. Rigau-Perez, G. Clark, P. Reiter, and D. Gubler, "Using disability-adjusted life years to assess the economic impact of dengue in puerto rico: 1984-1994," *The American journal of tropical medicine and hygiene*, vol. 59, pp. 265–71, 1998. doi: 10.4269/ajtmh.1998.59.265

[17] U. G. Assembly, "Transforming our world : the 2030 agenda for sustainable development," accessed 11 October 2019. [Online]. Available: https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

[18] Y. Yusof and Z. Mustaffa, "Dengue outbreak prediction: A least squares support vector machines approach," *International Journal of Computer Theory and Engineering*, vol. 3, pp. 489–493, 2011. doi: 10.7763/IJCTE.2011.V3.355

[19] T. Carvajal, K. Viacrusis, L. F. Hernandez, H. Ho, D. Amalin, and K. Watanabe, "Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan manila, philippines," *BMC Infectious Diseases*, vol. 18, 2018. doi: 10.1186/s12879-018-3066-0

[20] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress, 2015, pp. 67–80. ISBN 978-1-4302-5990-9

A Appendix

A.1 Feature correlation

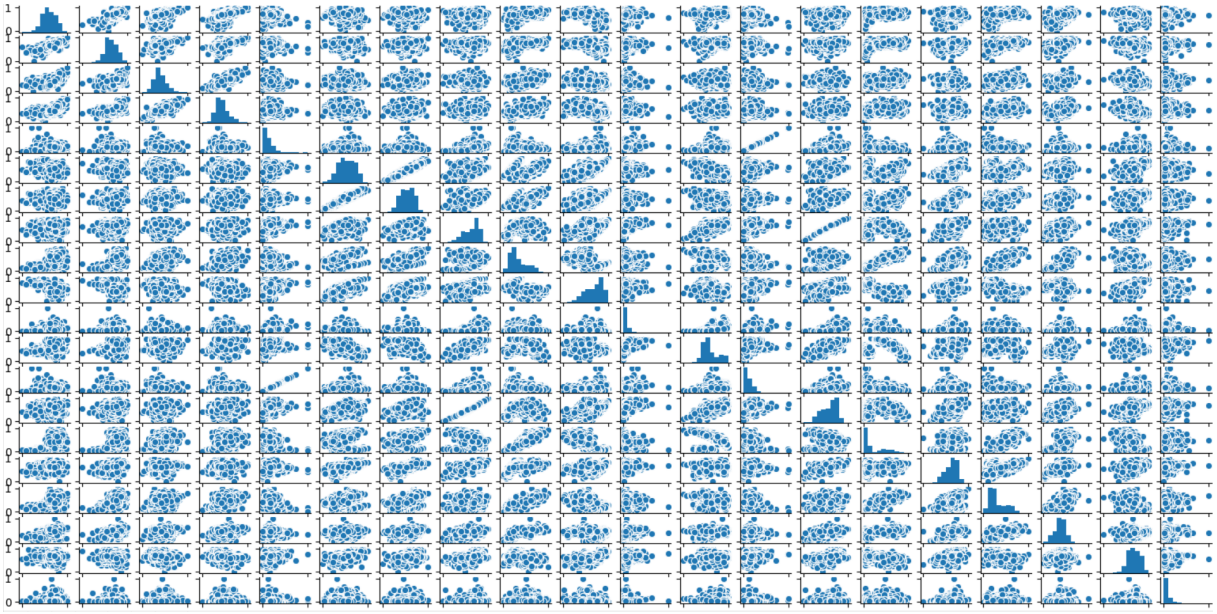

Figure 6: Whole feature correlation matrix

A.2 Feature Importance

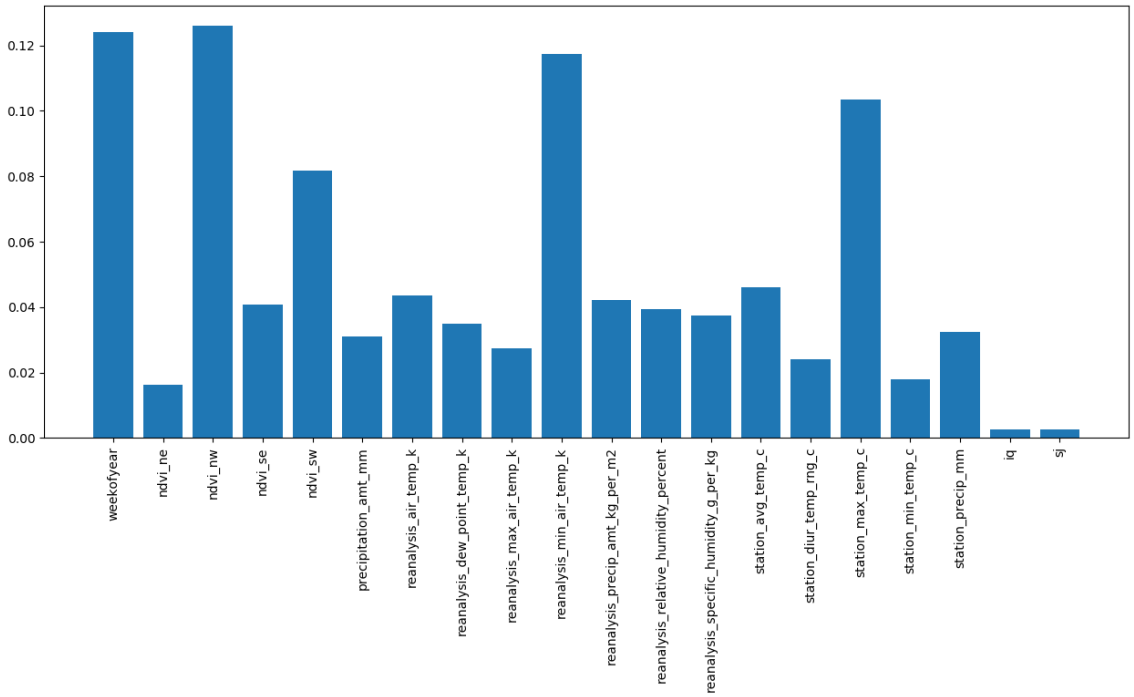

Figure 7: Feature Importance by random forest

A.3 Source code

The source code for our RF and SVR analysis can be accessed through the following GitHub repository: `https://github.com/javierpe27/denguePrediction`