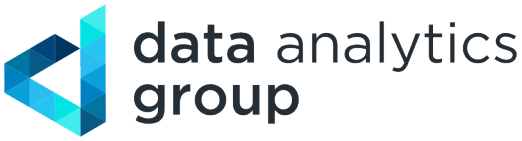

Universität Paderborn
Fakultät für Wirtschaftswissenschaften
Department Wirtschaftsinformatik

Studienarbeit

Malena Brinkmann
und
Vincent Jian Arvand
Warburger Str. 100, 33098 Paderborn
malenab@mail.upb.de
varvand@mail.uni-paderborn.de

vorgelegt bei
Prof. Dr. Oliver Müller

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Paderborn, 03. März 2023

M. Brinkmann

Contents

List of Figures

List of Tables

1 Business Understanding

1.1 Defining the Business Problem

The continuing destruction and pushback of earth's diverse biosphere is not only problematic in relation to preserving and protecting biodiversity on our planet but has also been proven to increase the risk of new zoonotic diseases spreading between global population centers (Keesing and Ostfeld, 2021). This is especially the case when the loss of biodiversity decreases the predation of hosts, thus increasing the density of infected host animals (Keesing et al., 2010). It has been estimated that about 60% of emerging infectious diseases are zoonoses, with over 30 new human pathogens having been discovered in the last three decades alone of which 71% are of zoonotic origin (Jones et al., 2008). An example of a zoonotic virus with increasing numbers of infections is the dengue virus. As the World Health Organization states, the number of around 500.000 cases in 2000 has increased to more than 5 million in 2019, and an estimated 390 million cases occur annually worldwide (WHO, 2022). The virus is transmitted by Aedes mosquitos, whose optimal temperature for living are 27 to 32 degrees (Molla, 2019). Thereby, most cases are found in the Indian subcontinent, South America and parts of Africa (see Figure 1.1).

Global warming increases dengue cases because, due to warmer temperatures, the optimal temperatures for Aedes mosquitos are met. Thus the time it takes for a mosquito to develop fully is shortened (Molla, 2019). Common symptoms of dengue fever include fever, joint pain and internal bleeding (Molla, 2019). The WHO describes dengue fever as a "global burden" (WHO, 2022).

To mitigate and prevent further detrimental pandemics, like the 2019 spreading of Covid-19 that started in Wuhan, machine learning could play a pivotal role in predicting those outbreaks before they even happen.

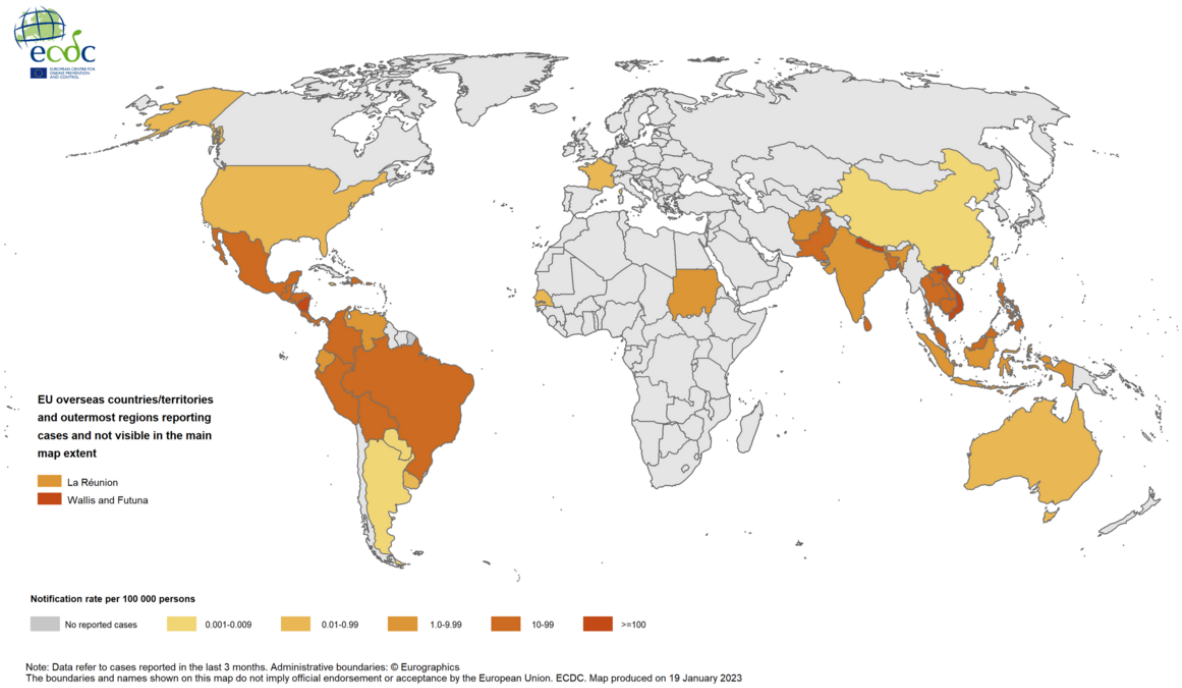

Figure 1.1: Spread of dengue fever cases in 2023, (ECDC, 2022)

1.2 Business Problem as a Data Mining Problem

The prediction of dengue fever cases can be supported by machine learning methods, which we will explore going further. Our output variable is the total cases of dengue fever in San Juan, a city in Puerto-Rico, and Iquitos, a city located in Peru, making it a regression problem. Given is historical data of past dengue fever cases in both cities. The goal of this work is to build machine learning methods that are able to predict cases in the future, given one of the two cities. In the following text we will illustrate to what extent machine learning techniques can be leveraged to predict outbreaks of the dengue virus, and which environmental variables may play a role in exacerbating this disease spreading.

1.3 Defining Success Criteria

In order to evaluate our models and their predictions, our success criteria should be established beforehand. Since we have a regression problem, some common ways to

evaluate the predictions are by calculating the Mean Squared Error (MSE), the Mean Absolute Error (MAE), and the R^2 score (Sekeroglu et al., 2022). We focus on the MAE to be able to compare our results with other models developed for the same task, as documented on the *drivendata.org* leaderboard (DrivenData, 2023). This also gives us an idea as to what is realistic to achieve, since at the moment, no MAE below 10 was reported yet. Therefore we strive to achieve a total MAE in the range of about 20-30.
Furthermore, we will plot all predictions with the actual cases and analyze the quality of the predictions made. Optimally, the best model(s) will get the average of the cases right, meaning if there are no peaks over a period of time, the algorithms will predict these cases pretty accurately. If there are higher peaks, they should predict these in rudiments at least.
Having defined our success criteria, we can now start looking at the actual data.

2 Data Understanding

In this chapter we will explore our underlying data set by examining its unaltered structural composition and contents.

2.1 Initial Data Sets

In general the given data was provided to *drivendata.org* by various U.S. institutions such as the U.S Centers for Disease Control, as well as multiple universities and the Peruvian government.
The data supplied by *drivendata.org* contains three different data sets:

- feature training data
- label training data
- feature test data

2.1.1 Feature Training data

The feature training data consists of 21 different features, with a total of 1,456 entries. Since environmental factors play an essential part in the spreading of dengue fever (Faruk et al., 2022), most of the recorded data contains environmental variables sourced from two different locations. This data set only contains numerical variables except for the categorical feature 'city' with two possible values: 'sj' for San Juan and 'iq' for Iquitos.

Timescale and Location Indicators

Our feature dataset provides information on a 'year' and 'weekofyear' timescale while also retaining the recorded location as either 'sj' or 'iq'. Structure-wise the data set only consists of discrete data points taken at regular intervals, thus representing a time series. The data set spans a total of two decades, beginning in 1990 and ending in 2010.

Data Sources

The data was gathered by multiple systems such as

- **GHCN daily climate data weather station measurements**

  The Global Historical Climatology Network is a database of connected weather stations tracking environmental variables such as the average air temperature, maximum and minimum air temperature as well as the precipitation (NOAA, 2021).

- **PERSIANN Satellite measurements**

  PERSIANN is an algorithm estimating precipitation by using infrared satellite data (lin Hsu et al., 1997).

- **NCEP Climate Forecast System Reanalysis measurements**

  The Climate Forecast System (CFS) uses multiple data sources to model global interaction between oceans, continents and the atmosphere (Saha et al., 2010).

- **Normalized difference vegetation index**

  An index depicting the amount of vegetation in a certain area. The more the value converges to 1 the higher the amount of green vegetation present (NASA, 2000).

2.1.2 Label Training Data

This contains time data such as 'year', 'weekofyear', 'city' and 'total_cases'. We merged this data set with the feature train data in order to get the total cases for each observation.

2.1.3 Feature Test Data

The file contains the same features as the feature train file. We are not using it because there is no data on the number of total cases, therefore it cannot be used to train and test machine learning models.

2.2 Exploratory Data Analysis

As already mentioned, our data set contains a collection of environmental variables recorded in two geographically distinct places. Plotting the recorded cases (see Figure 2.2) immediately reveals a few noteworthy things. Firstly, not all of our case data depicts the same time frame. Taking a look at our data shows that the collection of environmental data started in San Juan during the 1990s while Iquitos only started in 2000. Taking into account the different geographic and environmental circumstances as well as the different time frame, splitting up our data set into two may yield better results overall.

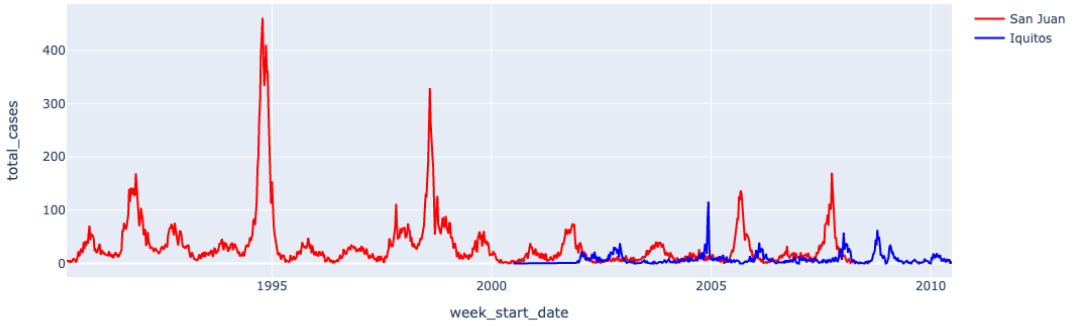

Figure 2.1: Total cases in San Juan and Iquitos

As we can see, the scope of endemic outbreaks are quite different between both cities. While both cities seem to experience periodic dengue outbreaks, San Juan has a lot more dengue cases per outbreak, even though their population is approximately equal in size (Benedum et al., 2020).
The cases in both cities seem to be independent, or only have a small correlation. For example, while the highest peak of cases in Iquitos was at the end of 2004/beginning of 2005, there was a relatively low number of cases in San Juan at the same time.

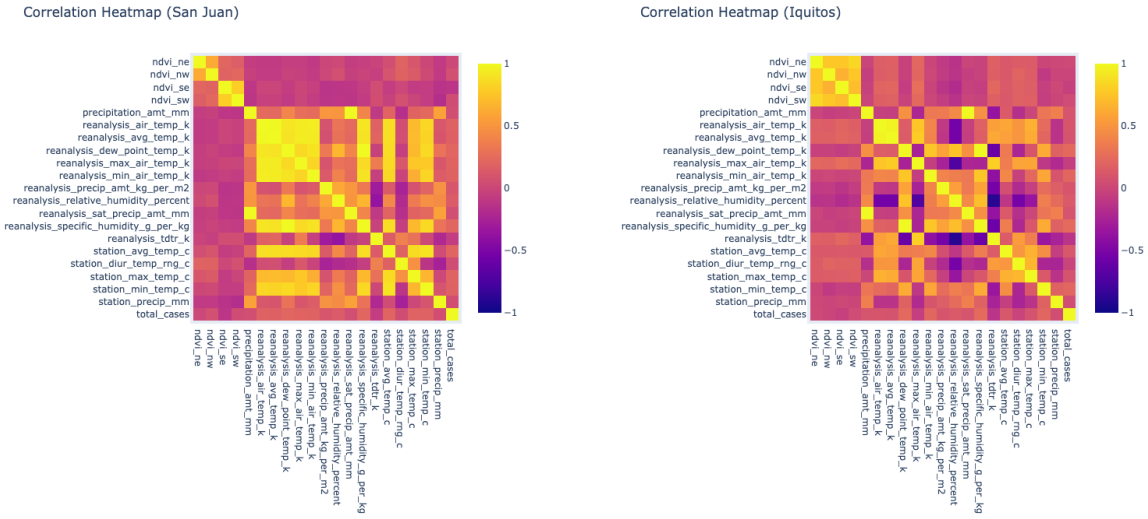

Figure 2.2: Correlation heatmaps for all features for San Juan (left) and Iquitos (right)

The heatmaps in Figure 2.2 visualize the correlation the different variables have with

each other in both cities.
In Iquitos there seems to be a positive correlation between the variables prefixed with 'ndvi'. Other than that, some strong positive and negative correlations exist sporadically. In San Juan most of the 'reanalysis' variables are highly correlated with each other, while the other variables do not show any noticeable relations.
In general most of the temperature variables correlate, which is to be expected. In both cities, all features only correlate very little to the output variable 'total_cases', which is something to keep in mind when building models.

3 Data Preparation

After exploring the general structure and characteristics, we will now have a look at the preparation of our data before we can train models with it.
As already mentioned, the dataset is split into two geographical places. The history of cases shows very different patterns, considering height of the peaks and when they occur in terms of time. This is why we decided to split the dataset into two datasets, which we will later on analyse seperately. Thus, we will train different models for each city.

3.1 Preparing Data

A similarity both feature datasets have is the absence of values in almost all columns. Most features have a missing value percentage of around 1%, which is not very drastic. The only exception are the features 'ndvi_ne' with about 15,37% missing values and 'ndvi_nw' with 3,7%. Nonetheless, these percentages are still relatively low, which is why we decided to fill them in with existing values of the feature, by always taking the last observated feature value.
Another aspect to look at is the different unit of measurements in the variables belonging to temperature, as some are in Kelvin and others are in Celsius. In order to keep the numerical values in similar ranges, we decided to transform the Kelvin temperature values into Celsius.
Now that the data is transformed, we can start training models.

4 Modeling

4.1 General Approach

4.1.1 Feature Selection

As seen before, our data set does not contain strong correlations between the different variables to the output variable. This makes selecting viable features a difficult task. As an approach to deal with this problem, we decided to use recursive feature elimination with a logistic regression as an estimator. Since a logistic regression does not assume any linear relationship between variables (Singh et al., 2016, p.02) it should be a fitting estimator for our feature selection.
The following table shows the selected features for both cities.

San Juan	Iquitos
reanalysis_precip_amt_kg_per_m2	*reanalysis_max_air_temp_k*
reanalysis_relative_humidity_percent	*reanalysis_relative_humidity_percent*
reanalysis_sat_precip_amt_mm	*reanalysis_tdtr_k*
station_max_temp_c	*station_precip_mm*

Table 4.1: Selected features for San Juan and Iquitos

4.1.2 Splitting the Data

For all models, we split the data into a training and a test set, with a ratio of 80:20. The models were trained with the training set and the performance was measured with the test set. When splitting the data, we made sure the data is not shuffled before the split. This way, the test split is still in chronological order time-wise and can be compared to the predicitions visually, using plots.

4.1.3 Calculating Results

Both datasets have different sizes; for Iquitos there are 520 observations, and for San Juan 936. To calculate the total MAE for each model, we multiplicate the indiviual MAEs with their percentage of all observations and add those two values up to receive

the total MAE.

4.2 Multiple Linear Regression

Linear Regression is the first approach we used for this problem. Due to no feature correlating noteworthy to our outcome variable, we had to use multiple linear regression to get any good results. In a study that used multiple linear regression along with other many forecasting techniques for predicting dengue fever cases, the mutliple linear regression predicted with a Mean Absolute Error of 29.8; scoring 6th best out of 13 other methods (Baker et al., 2021). Considering the simplicity of the method, this study shows that it can possibly predict pretty effectively. In addition to that, linear regression is very commonly used in spatial analysis (Hoyos et al. (2021)).

4.2.1 Implementation

For the list of features, we used the outcome of the feature selection in both cities.
In San Juan, this resulted in an Mean Absolute Error of approximately 23.7323 on our test set. In Iquitos, a MAE of 7.1227 was achieved. The following plots show the actual number of cases versus the predictions of the multiple linear regression model.

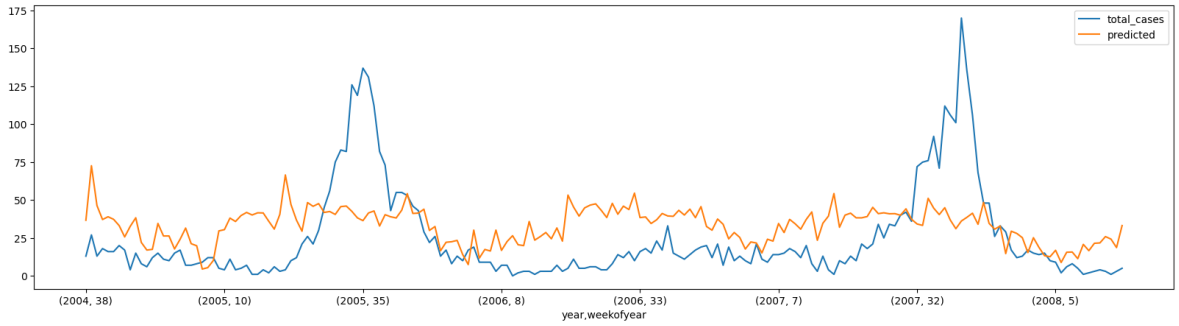

Figure 4.1: Prediction using multiple linear regression in San Juan

As one can see, the peaks are not well predicted, but rather the average of the cases in both cities. In total, the multiple linear regression can average the cases, but is not able to predict peaks very well. The resulting MAE for this method is 17.8003.

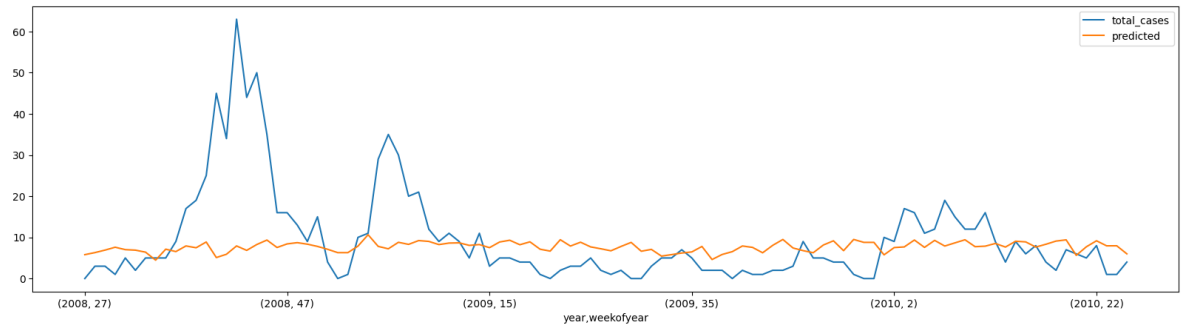

Figure 4.2: Prediction using multiple linear regression in Iquitos

4.3 Random Forest

Using random forests on a forecasting task is a very common approach (Siddiq et al., 2021; Baker et al., 2021). In a study testing different machine learning methods in a forecasting task for predicting dengue fever cases in Jeddah, a city in Saudi Arabia, a random forest regression was used for dengue fever case prediction (Siddiq et al., 2021). It scored a 55% accuracy and a Mean Absolute Error of 0.4474; coming after decision tree regression with 57%, and Support Vector Classification with 76% accuracy.
In another work (see Baker et al., 2021) a random forest regression was used among other methods to predict the number of dengue fever cases. It resulted in an MAE of 26.6, which scored third best of a total of 13 methods used. The competing methods included methods like neural networks and Support Vector Machines.
Considering the good results a random forest regression has achieved in similar problems, we decided to build a random forest regression model as well.

4.3.1 Implementation

As a first step, we built a random forest model for each city, with already good results. In San Juan the MAE was 25.7146, and in Iquitos 7.8471, which resulted in a total MAE of 18.7750.
Going further, we used hyperparameter tuning. In order to estimate good hyperparameters, we used a grid search with 5-fold cross validation with the following value range:

Hyperparameter	Value range	San Juan	Iquitos
n_estimators	[20, 50, 100, 200]	50	100
max_depth	[5, 25, 50]	5	2
min_samples_leaf	[5, 10, 2, 20]	5	20
min_samples_split	[2, 5, 10]	5	5
max_features	['auto', 'sqrt']	'auto'	'auto'

Table 4.2: Value range of hyperparameters and the selected values by grid search for random forest in both cities

With hyperparameter tuning, we were able to improve our predictions from the multiple linear regression models for both cities, reaching a total of 16.7552. This is an improvement of approximately 7%.

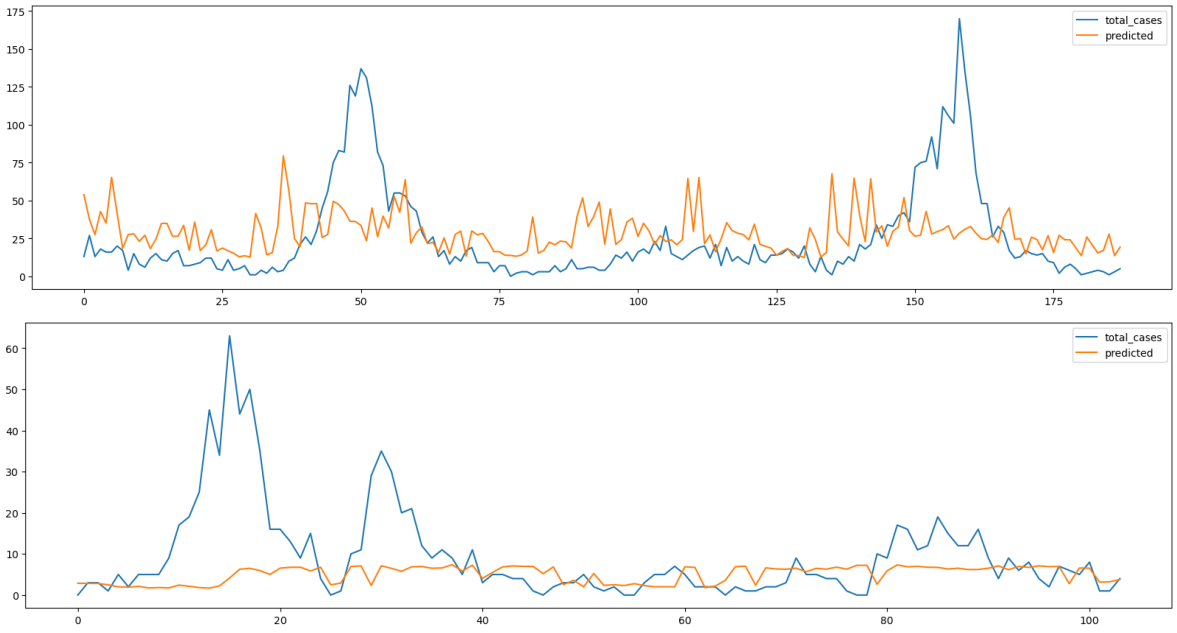

Figure 4.3: Prediction using a random forest in San Juan (above) and Iquitos (below)

As one can see (4.3), the random forest model is able to predict the peaks a little better than the multiple linear regression model. However, the higher peaks of cases are still not predicted accurately.

4.4 Time Series Analysis - ARIMA

ARIMA or 'Autoregressive Integrated Moving Average' is a technique to describe and forecast data in a given time series. It combines three different components: autoregression (AR), integrated (i), and the Moving Average (MA) (Hyndman and Athanasopoulos, 2018). Autoregression means that the value of the variable which is being forecasted depends on its past values, while the degree of dependence is specified by the order of the AR term (Hyndman and Athanasopoulos, 2018). An AR term of the order 1 for example would use the previous value of the variable to predict the current value, while an AR term of order 2 would use the two previous values. Differencing (I) is used to make the time series stationary, meaning that properties like mean and variance are constant over time. The Moving Average (MA) means that the value of the forecast is dependent on past errors of the time series, with the degree of dependence specified by the order of the MA term (MA of order one would use the previous residual value for a prediction, while MA of order two would use the previous two) (Hyndman and Athanasopoulos, 2018). It is important to note that differencing alone may not be sufficient to capture the underlying patterns of a time series (Hyndman and Athanasopoulos, 2018).

By ensuring that the time series is stationary before fitting the model, we can be more confident that the patterns captured by the AR and MA components are actually meaningful. To test the time series on stationarity, we decided to use the Augmented Dickey-Fuller test, which assumes the existence of a unit root. Running the Augmented Dickey–Fuller test on our time series reveals a p-value of approximately $5.963e-9$ which is less than the significance level of 0.01, thus rejecting the null hypothesis and inferring stationarity.

4.4.1 Implementation

As a first step we used 'auto ARIMA', which uses the Hyndman-Khandakar (Hyndman and Athanasopoulos, 2018) algorithm, to determine the best ARIMA and integration order. The algorithm returned an AR of order 2, an MA of order 6, as well as a first degree of integration (ARIMA(2,1,6)). Afterwards we trained the model on 80% of our feature data while the remaining 20% were reserved for testing the model. We kept 'week_start_date' as the index. We start forecasting from the last point of time in the training set.
In both cases, a almost straight line was predicted (4.4). This could be due to a lack of trend in the time series, which causes ARIMA to use the mean as a prediction in

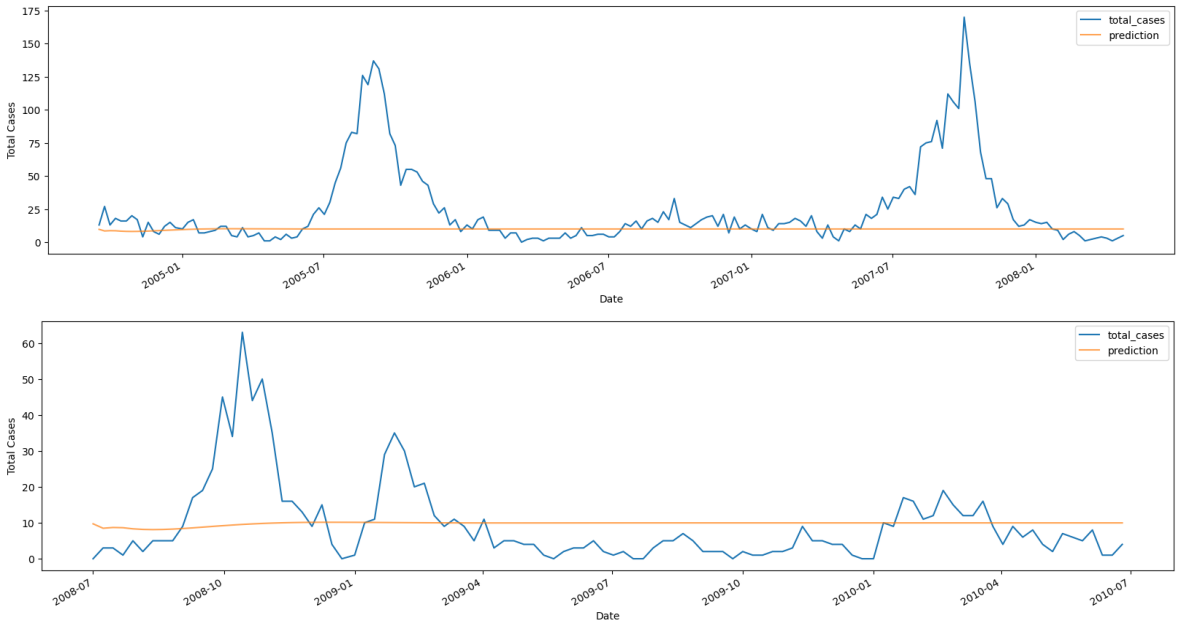

Figure 4.4: Prediction using ARIMA in San Juan (above) and Iquitos (below)

the long term. Due to the smaller time period of the test set the MAE is 18.2953, even though it is apparent that the model does not forecast the actual values well, especially in the long term. The MAE of the Iquitos model is 7.9315, which is probably due to the smaller set size. The total MAE is therefore 14,5939.

4.5 Support Vector Machines

Support Vector Machine Learning or Support Vector Regression work by finding a hyperplane which maximizes the margin between predicted and the actual target values (Noble, 2006). In SVM Regression the primary objective is to find a hyperplane which contains the maximum number of data points within a certain distance, denoted as ϵ, which is also called the epsilon-insensitive loss function (Vapnik, 1999). The SVM regression algorithm tries to minimize the loss function subject to a regularization constraint, this regularization helps to prevent overfitting. To make a new prediction, the algorithm calculates the distance between the hyperplane and the input features. If the calculated value lies in the ϵ range, it is a predicted output value (Cristianini and Ricci, 2008).

In a forecasting project for dengue fever cases in China, different forecasting methods were used to predict the number of cases in 5 different cities and other cities in Guang-

dong province, that were summarized into one variable (Guo et al., 2017). A distinction was made between two points in times: the outbreak period and the following 12 weeks. The application of Support Vector Regression has resulted in a very good perfomance, beating a few other applied methods such as linear models and gradient boosted regression tree models (Guo et al., 2017). In this study, the performance was measured by the Root Mean Squared Error (RMSE) and R^2. The Support Vector Regression had an significantly smaller RMSE in all different cities that were evaluated, and a R^2 value of at least 88% in all cases.

4.5.1 Implementation

As a first step, we scaled our data. While some machine learning algorithms such as random forests do not necessarily require feature scaling, not scaling the data when using Support Vector Machines can cause multiple issues, one of the main issue being that some features dominate the distance computation due to different value scales, which results in an inaccurate hyperplane and overfitting (Hsu et al., 2003).

As a next step we used grid search with 5-fold cross validation to test different hyperparameter combinations. Using the following value ranges:

Hyperparameter	Value range	San Juan	Iquitos
Kernel	['linear', 'rbf', 'sigmoid', 'poly']	'linear'	'linear'
C	[0.1, 5, 10, 100]	100	0.1
gamma	['scale','auto']	'scale'	'auto'
epsilon	[0.001, 0.009, 0.1]	0.1	0.001

Table 4.3: Value range of hyperparameters and the selected values by grid search for SVM in both cities

Using the hyperparameters discovered in the grid search, a MAE of 18.0869 was achieved for San Juan, while running the SVM algorithm with the default hyperparameter values yielded a MAE of 19.0880. Using SVM Regression on the Iquitos dataset yielded an MAE of 7.0865. Using the standard parameters instead of the hyperparameters only has minor effects in this case.
As one can see (Figure 4.5) the prediction seems to lay in the average of cases and is not good at predicting higher peaks. The total MAE is 14.1582. The plots show that the SVM regression has problems forecasting higher peaks, similar to multiple linear

regression and random forests. Overall the performance is still relatively good.

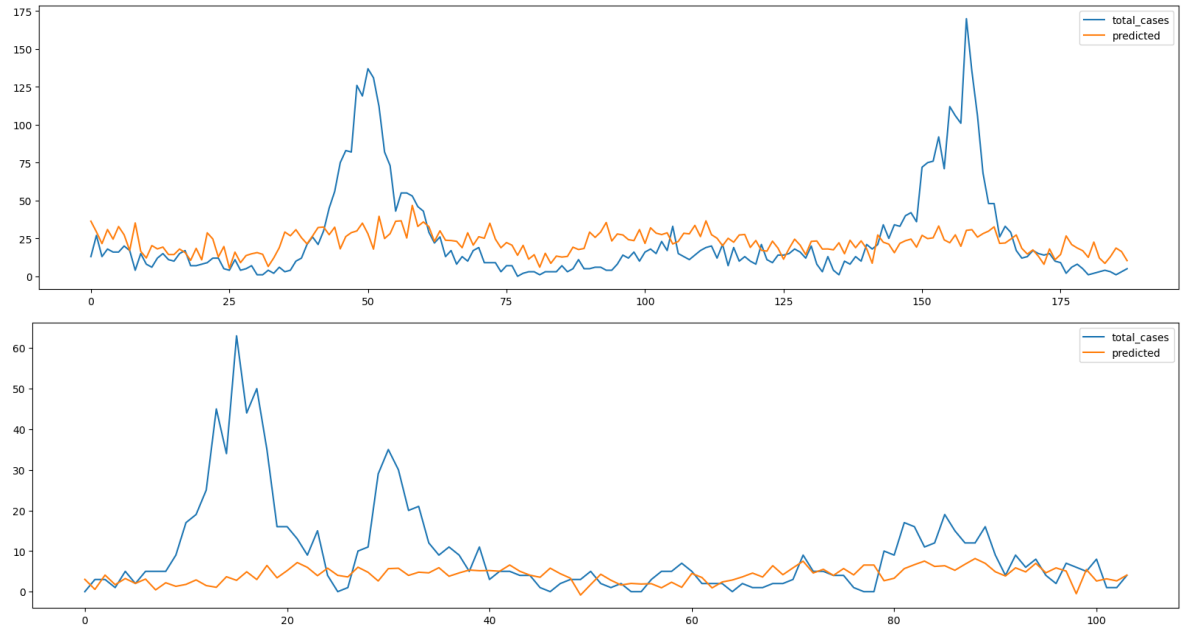

Figure 4.5: Prediction using SVM in San Juan (above) and Iquitos (below)

5 Evaluation

Having built and evaluated our different models, we will now compare them with each other and their results to our defined success criteria.
The performance of our models was not a spot-on prediction in many cases but rather a good approximation of dengue cases to be expected in the future. This approximation can be a very useful asset for governments and health organizations to get a rough idea of how strong dengue fever may affect their region in the future. As a result, this can enable the effective implementation of measures to prevent severe outbreaks of dengue fever. From a technical point of view, all of our models passed our success criterion of a MAE laying in the range between 20 and 30, and even exceeded our goals, with our worst result being a MAE of 17.8003. All models worked well for predicting a baseline average of cases, ARIMA practically only predicted a fixed value, but they all did not predict the high peaks very well. Looking at the MAEs of all methods, the best results

were achieved with SVMs, followed by ARIMA. However, looking at the plots, ARIMA performed the worst. Therefore, no model is a clear winner in this case.

5.1 Conclusion

Due to the limitation of this project, only a few different machine learning methods could be explored in this work. Even though the our performance goal was met, there are a lot more machine learning methods that have resulted in good performances in similar problems in the past, like a Naive Bayes or Poisson regression (Baker et al., 2021). Another different approach could be to use neural networks, as this has achieved good results in a study of dengue outbreak in Malaysia (Salim et al., 2021). Exploring these different approaches could result in better performances than those of our current models.
Looking at our own models, one could improve them even more by exploring different search strategies for hyperparameters than we used, for example by using a framework specifically designed to optimize hyperparameters, like 'optuna' (Kakarla et al., 2023). Even though there is much room for potential, this project has shown that the prediciton of dengue fever cases can be supported by machine learning techniques in many different ways.

6 Deployment

Now that the models have been evaluated and compared regarding to our success criteria, the question arises if the knowledge gained in the process can be used in any form in the future.
Our system contains a training and preprocessing pipeline on a server. Since the data, especially variables concerning environmental factors, might change in the future, the ability to frequently re-train the model on new data should be retained. A simple API such as REST could be used to frequently receive, as well as to automatically post data to an external source (e.g another server). A possible implementation is demonstrated in Figure 6.1.

Possible requests for the REST API would be a GET request, to get predictions for data that has already been processed, or a POST request, where the requestor sends the

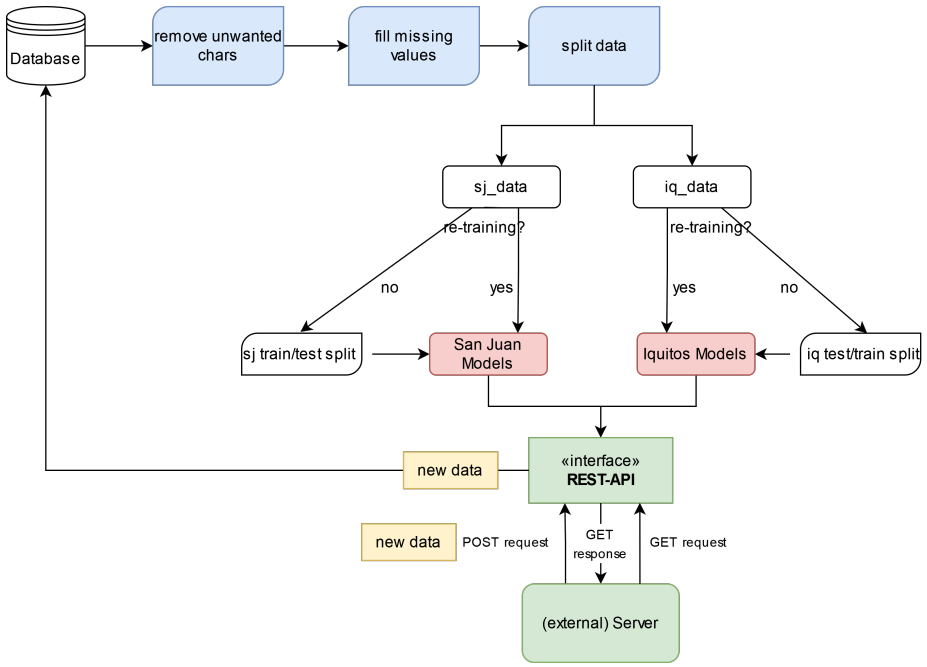

Figure 6.1: Server-side processing and re-training with external communication ability

data that they want predictions for within the request. In this case, a re-training with the new data would be executed. The condition for this data is, that it has the same structure as the data we processed in this text. Since our models have been trained only on data from Iquitos and San Juan, predictions for other locations would not be reasonable.

Bibliography

Baker, Q. B., Faraj, D., and Alguzo, A. (2021). Forecasting dengue fever using machine learning regression techniques. In *2021 12th International Conference on Information and Communication Systems (ICICS)*, pages 157–163. IEEE.

Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., and Markuzon, N. (2020). Weekly dengue forecasts in iquitos, peru; san juan, puerto rico; and singapore. *PLoS neglected tropical diseases*, 14(10):e0008710.

Cristianini, N. and Ricci, E. (2008). *Support Vector Machines*, pages 928–932. Springer US, Boston, MA.

DrivenData (2023). Dengai: Predicting disease spread. `https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/leaderboard/`.

ECDC (2022). Dengue worldwide overview. `https://www.ecdc.europa.eu/en/dengue-monthly`.

Faruk, M. O., Jannat, S. N., and Rahman, M. S. (2022). Impact of environmental factors on the spread of dengue fever in sri lanka. *Int J Environ Sci Technol (Tehran)*, 19(11):10637–10648.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., et al. (2017). Developing a dengue forecast model using machine learning: A case study in china. *PLoS neglected tropical diseases*, 11(10):e0005973.

Hoyos, W., Aguilar, J., and Toro, M. (2021). Dengue models based on machine learning techniques: A systematic literature review. *Artificial intelligence in medicine*, 119:102157.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Hyndman, R. and Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993.

Kakarla, S. G., Kondeti, P. K., Vavilala, H. P., Boddeda, G. S. B., Mopuri, R., Kumaraswamy, S., Kadiri, M. R., and Mutheneni, S. R. (2023). Weather integrated multiple machine learning models for prediction of dengue prevalence in india. *International Journal of Biometeorology*, 67(2):285–297.

Keesing, F., Belden, L. K., Daszak, P., Dobson, A., Harvell, C. D., Holt, R. D., Hudson, P., Jolles, A., Jones, K. E., Mitchell, C. E., et al. (2010). Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, 468(7324):647–652.

Keesing, F. and Ostfeld, R. S. (2021). Impacts of biodiversity and biodiversity loss on zoonotic diseases. *Proceedings of the National Academy of Sciences*, 118(17):e2023540118.

lin Hsu, K., Gao, X., Sorooshian, S., and Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, 36(9):1176 – 1190.

Molla, M. A.-M. (2019). Dengue Outbreak in South Asia: Climate change the culprit? *The Daily Star*.

NASA (2000). Measuring vegetation (ndvi amp; evi). `https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php`.

NOAA (2021). Global historical climatology network daily (ghcnd). `https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily`.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M., Sela, J., and Goldberg, M. (2010). The ncep climate forecast system reanalysis. *Bulletin of The American Meteorological Society - BULL AMER METEOROL SOC*, 91.

Salim, N. A. M., Wah, Y. B., Reeves, C., Smith, M., Yaacob, W. F. W., Mudin, R. N.,

Dapari, R., Sapri, N. N. F. F., and Haque, U. (2021). Prediction of dengue outbreak in selangor malaysia using machine learning techniques. *Scientific reports*, 11(1):939.

Sekeroglu, B., Ever, Y. K., Dimililer, K., and Al-Turjman, F. (2022). Comparative evaluation and comprehensive analysis of machine learning models for regression problems. *Data Intelligence*, 4(3):620–652.

Siddiq, A., Shukla, N., and Pradhan, B. (2021). Predicting dengue fever transmission using machine learning methods. In *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 21–26. IEEE.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

WHO, W. H. O. (2022). Dengue and severe dengue. `https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue`.