Instituto Superior Técnico Lisboa
Faculty of Mathematics
Applied Mathematics Master

Statistical Methods in Data Mining

A Dengue Infection Severity Study

Professor:
Conceição Amado

Authors:
Luis Oliva Fontecha
Mahima Raut
Miguel Lourenço Farinha
Sahil Rajesh Kumar

Lisbon, August 16, 2020

Contents

1 Introduction

This is a study which aims to classify the severity of the Dengue pandemics in the city of San Juan, Puerto Rico. Several supervised learning techniques were used to fit different classifiers such as K-Nearest Neighbors, Naïve Bayes, Linear Discriminant Analysis and Quadratic Discriminant Analysis among others. The best method appeared to be the Random Forest algorithm having an accuracy of more than 50%.

Dengue is a mosquito-transmitted viral disease which may cause nausea, vomiting, and pain in multiple parts of the body, leading to shock, internal bleeding, and even death. This terrible disease is a global challenge for humanity. Indeed, about 3 billion people, 40% of the world's population, live in areas with a risk of Dengue. [CC]

This kind of studies are key for the affected populations since these epidemics are currently unpredictable and cause major consequences. In fact, according to the Centers for Disease Control and prevention (CDC), each year, up to 400 million people get infected with Dengue, approximately 100 million people get sick from infection, and 22,000 die from severe Dengue.

Nevertheless, accurate Dengue predictions would help public health workers and people around the world take steps to reduce the impact of these epidemics. But predicting Dengue is a hefty task that calls for the consolidation of different data sets on disease incidence, weather, and the environment.

Specially, the data for this report comes from the following sources:

- Dengue surveillance data is provided by the **U.S. Centers for Disease Control and prevention**, as well as the **Department of Defense's Naval Medical Research Unit 6** and the **Armed Forces Health Surveillance Center**;

- Environmental and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA), an agency of the **U.S. Department of Commerce**.

2 Data preparation and exploratory analysis

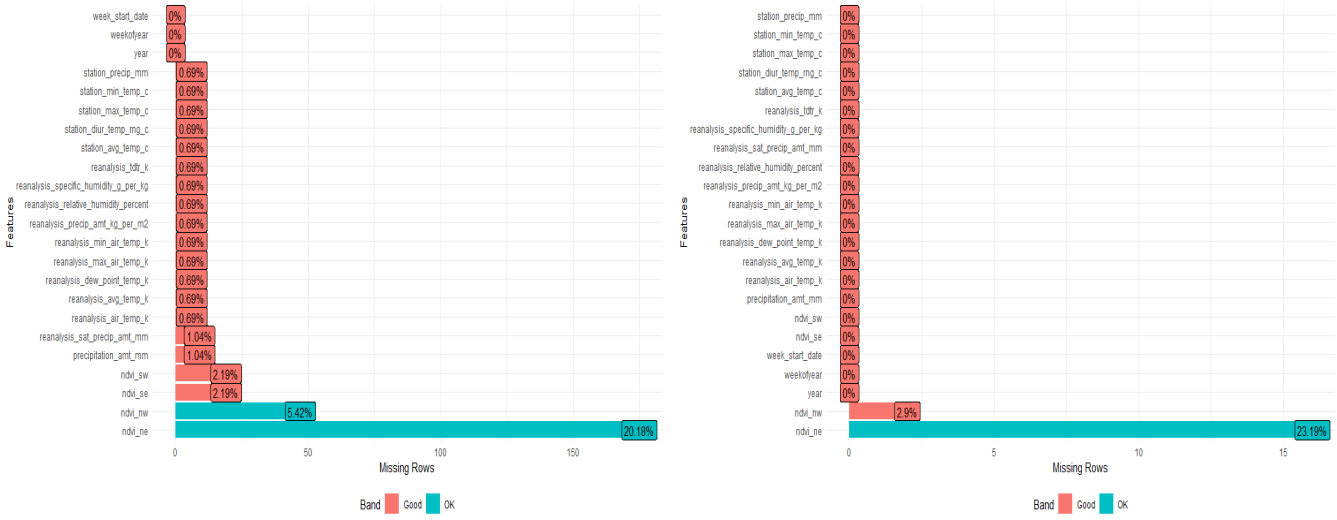
2.1 Dataset

The database provided for this study already had two separate datasets, one for the training stage and one for the testing stage. The data consists of weekly reports about the number of individuals suffering from Dengue plus environmental and climate information from 1990 to 2006 for the training set and from 2007 to 2008 for the test set. Obviously, the Dengue information for the test set is unknown. This is what this report aims to estimate. As illustrated in Table 1, some values are missing in the data.

	Training Set	Test Set
Rows	867	69
Columns	23	23
Quantitative Columns	22	22
Categorical Columns	1	1
Total Missing values	362	18
Complete rows	675	52

Table 1: Raw counts about covariates.

Deleting the instances where missing data was present was not an option since it would mean dropping more than 20% of the training and test data according to the number of *complete cases* in Table 1.
When dealing with missing data, it is necessary to know how this missing information is distributed. For instance, if all the missing values were concentrated in just one column it would be logical to consider deleting that column. The distribution of the missing values for each dataset is described in Figure 1.



(a) Training set

(b) Test set

Figure 1: Distribution of the missing values.

It is interesting to see that in both datasets, the feature having the most missing values is the same. However, the number of missing values per covariate is reasonable in both cases so choosing an imputation method for this values arises as the best method to minimize this loss of information. For instance, in this case **predictive mean matching** was the method used to overcome this difficulty. For more information about this method refer to appendix 9.1.

In accordance to Table 1, the dataset possessed 867 observations and each of the observations was measured in 23 distinct variables. The variables of this study are:

Feature	Description
year	Year of data collection
weeekofyear	Week of the year of data collection
week_start_date	Day of start of data collection
ndvi_ne	Pixel northeast of city centroid
ndvi_nw	Pixel northwest of city centroid
ndvi_se	Pixel southeast of city centroid
ndvi_sw	Pixel southwest of city centroid
precipitation_amt_mm	Total precipitation
reanalysis_air_temp_k	Mean air temperature
reanalysis_avg_temp_k	Average air temperature
reanalysis_dew_point_temp_k	Mean dew point temperature
reanalysis_max_air_temp_k	Maximum air temperature
reanalysis_min_air_temp_k	Minimum air temperature
reanalysis_precip_amt_kg_per_m2	Total precipitation
reanalysis_relative_humidity_percent	Mean relative humidity
reanalysis_sat_precip_amt_mm	Total precipitation
reanalysis_specific_humidity_g_per_kg	Mean specific humidity
reanalysis_tdtr_k	Diurnal temperature range
station_avg_temp_c	Average temperature
station_diur_temp_rng_c	Diurnal temperature range
station_max_temp_c	Maximum temperature
station_min_temp_c	Minimum temperature
station_precip_mm	Total precipitation

Table 2: Features description.

By analysing the explanatory variables one can see that some of them measure exactly the same feature, albeit being measured in different units. Therefore, a high correlation between such covariables is expected, which might result in multicollinearity problems. This is due to the fact that for each observation some of the covariables measuring the same features come from different sources. Each source has different limitations and quality issues. This topic will be further explored in the subsequent sections.

The response variable to be analysed comes from a separate dataset. It consists in the total number of cases of infection by Dengue that occurred at specific date. This variable will be object of further study in the next section.

2.2 Total cases of Dengue

Predicting the total number of people who will suffer from Dengue is key to addressing the global challenge posed by this disease. However, country leaders, who are responsible for taking the necessary measures to protect the population from these pandemics, are usually not interested in the exact number of people who will be infected, which is more difficult to estimate, but rather they are interested in determining this by intervals. Then, each interval represents a class and the challenge becomes a classification problem.

Observation of Table 3, leads to the conclusion that the total number of cases varies greatly. More specifically, one can see that, in San Juan, there were observations at given dates with no infections by Dengue whereas other observations had extremely high numbers of infected people. The average total number of infections was approximately 35, which shows that for most of the recorded data the total number of infections was not severe in comparison with the total number of inhabitants of San Juan in 2006 which was about 425.000 people [Nat]. These assertions are corroborated by Figure 14 in the appendix.

Min.	Q_1	Median	Mean	Q_3	Max.
0.00	9.00	19.00	34.51	37.00	461

Table 3: Descriptive statistics of the variable *total cases*.

In order to obtain some insight about the "shape" of the response variable, as a kind of continuous replacement for the discrete histogram, the kernel density estimation (KDE) method was utilized. The KDE is a non-parametric way to estimate the probability density/mass function of a random variable which can mathematically be expressed as follows:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K_h \left(\frac{x - x_i}{h} \right) \tag{1}$$

where K is the kernel (which is chosen to be a symmetric probability density function) and $h > 0$ is a smoothing parameter designated bandwidth. The bandwidth changes the shape of the kernel, *i.e.*, a lower bandwidth means only points very close to the current position are given any weight, which leads to the estimate looking squiggly whereas a higher bandwidth means a shallow kernel where distant points can contribute. [Con]

In Figure 2 an estimate of the underlying distribution is shown as well as the response variable's histogram. To obtain such an estimate a Gaussian kernel was utilized. Indeed, the Gaussian kernel appeared to be an appropriate choice by testing several kernels without specifying any parameters. It produced the best approximation of the shape of the total cases variable. Moreover, the obtained AMISE value was considered acceptable (low). Then distinct parameters were implemented and applied to get the best possible approximation. Finally, an unknown probability mass function (variable total cases is discrete) was approximated using a Gaussian density. However, it is relevant to point out that the probability density function of the total cases variable is not Gaussian.

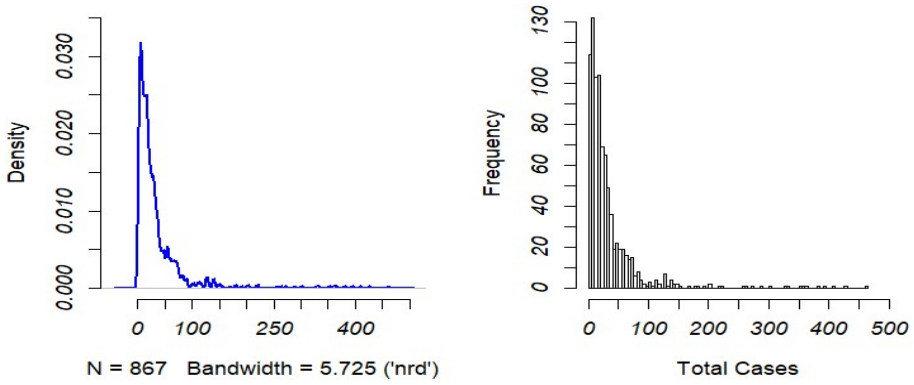


Figure 2: Kernel density estimation and histogram of the variable *total cases*.

Evidently for some observations the number of total cases is atypical. Thus, those observations might be considered as outliers regarding the referred variable. Therefore, an identification of the possible outliers was carried out using the method proposed by John Tukey [Lit03]. This method is based on the IQR (Interquantile Range) which can be

written as

$$IQR = Q_3 - Q_1 \tag{2}$$

where Q_1 and Q_3 are, respectively, the first and third quantiles of the ordered sample. Tukey suggested that a given observation should be considered an outlier if it was not contained in the following interval:

$$\Big[Q_1 - k \times IQR, Q_3 + k \times IQR\Big] \tag{3}$$

with $k = 1.5$.

Having this in mind the boxplot of the variable total cases was obtained and is presented below in Figure 3. According to the Tukey criterion, 67 possible outliers were detected which correspond to times when the number of total cases of infection by Dengue was extremely high.

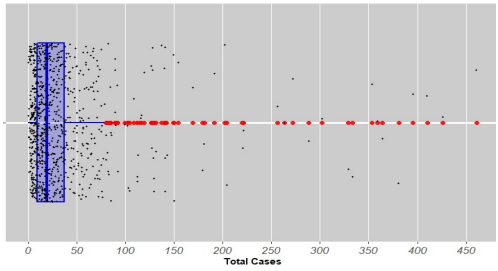


Figure 3: Boxplot of the variable *total cases*.

It is of interest to split the discrete total cases variable into suitable classes in order to have a feasible classification problem. When splitting a set of quantities into intervals, the first idea that comes to people's mind is to use quantiles and split the data in four balanced sets. Nevertheless, this is not the best idea in this case since, as illustrated in Figure 3, the outliers would provoke one of the classes to be much more spread than the others. Hence, this would carry a loss of information because belonging to the class containing the values between the 3^{th} and 4^{th} quantiles would result in having from 40 to 400 cases of Dengue.

In this study, class assignment was done as follows. Firstly, the outliers were assigned to one class, labeled as *"Very High"* number of total cases. Then, the quantiles of the *total cases* variable were re-calculated without considering these outliers, as illustrated in Table 4.

Min.	Q_1	Median	Q_3	Max.
0	8	17	31	78

Table 4: Quantiles of the *total cases* variable without outliers.

Capitalizing on the obtained quantiles, a class was assigned to each 25% interval. This process is illustrated in Table 5.

Class Assignment				
Low	Low-Medium	Medium	High	Very High
$x \in [0, 8]$	$x \in (8, 17]$	$x \in (17, 31]$	$x \in (31, 79]$	$x > 79$

Table 5: Class assignment rule for the *total cases* values.

The frequencies of each class are shown in Figure 4. An analysis of Figure 4 showed that the class frequencies were very similar in the case of the four *"lower"* [1] classes (around 200 occurrences per class) and the number of instances in the *"Very High"* class was 65% less (around 70 occurrences). This was expected since the latter only contained the observations considered outliers according to the Tukey criteria. The reduced number of instances in

[1] *"Lower"* in terms of the number of dengue total cases they represent: "Low", "Low-Medium", "Medium" and "High".

the *"Very High"* class might pose some limitations to the classifier's ability to predict the total number of infections in extreme cases. This is due to the fact that the utilization of fewer instances to build the classifier translates into a worse learning procedure.

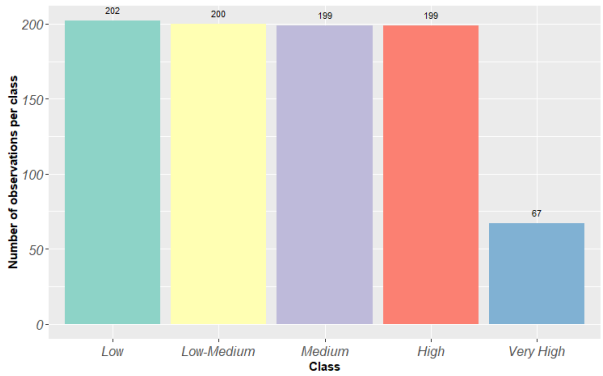


Figure 4: Class frequencies for the *total cases* labels.

2.3 Study of the explanatory variables

As a means to better understand the data of the problem the study of the explanatory variables was carried out. This study enabled the identification of covariables to which some given transformation should have been applied, the identification of atypical observations which could be inconsistent with the remainder of the observations and the identification of possible explanatory variables to be removed so as to avoid multicollinearity problems as discussed in section 2.1. This analysis took into consideration the distinct sources used to obtain the measurements for each observation on each of the features.

2.3.1 Environmental data sources

The data utilized throughout this study came from multiple sources. Dengue surveillance data was provided by the CDC, as well as the **Department of Defense's Naval Medical Research Unit 6** and the **Armed Forces Health Surveillance Center**, in collaboration with the U.S. universities. Environmental and climate data was provided by the NOAA, an agency of the **U.S. Department of Commerce**, from a variety of sources such as, ground observations, remote sensing, and reanalysis.

The utilized environmental sources are presented below:

- NOAA's GHCN daily climate data weather station measurements

  The data provided by the GHCN (Global Historical Climatology Network) stations consists in the explanatory variables *station_max_temp_c*, *station_min_temp_c*, *station_avg_temp_c*, *station_diur_temp_rng_c* and *station_precip_mm*. The temperature values were measured in degrees *Celsius* and the precipitation values were measured in *millimetres*.

- NOAA's CDR PERSIANN satellite precipitation measurements

  PERSIANN is a global climatological data record of precipitation from remote sensing information using an artificial neural network. PERSIANN was used to measure the explanatory variable *precipitation_amt_mm* in *millimetres*. The resolution of the data is in a 0.25×0.25 degrees scale.

- NOAA's NCEP Climate Forecast System Reanalysis measurements

  Climate Forecast System Reanalysis is a global, high-resolution, coupled atmosphere-ocean-land surface-sea ice. The values of the covariables with the prefix *reanalysis* were measured using this system. The temperature values were measured in *Kelvin*, the precipitation values were measured in mm and kg/m^2 and the values for humidity are presented as a percentage and in g/kg. The resolution of the data is in a 0.5×0.5 degrees scale.

- NOAA's CDR Normalized Difference Vegetation Index measurements

  NDVI CDR is a global climatological data record of vegetation. Four pixels closest to the city centroid are provided for evaluation of vegetation change corresponding to the covariables *ndvi_se*, *ndvi_sw*, *ndvi_ne*, *ndvi_nw*. The resolution of this data is in a 0.5×0.5 degrees scale.

It is possible to withdraw some conclusions by taking into account the individual characteristics of each one of the measuring sources. More specifically the resolution of each measuring system directly affected the accuracy of each measure and consequently it might influence the variance of a given covariable. Taking this into account, for the features which were measured using different sources (*e.g "total precipitation"* and *"average temperature"*), it is reasonable to expect that the explanatory variable with the lowest variance will be the one measured using a source with higher resolution. This criterion will be of utmost importance in the process of the covariables selection.

2.3.2 Analysis of the relations between covariables

As mentioned previously, the dataset contains explanatory variables which represent the same phenomenon. To reveal potential relations between the covariables and obtain some insight about possible multicollinearity problems an analysis of the correlation between the explanatory variables is of interest.

In Figure 5 the correlation matrix is presented. On the left of the diagonal of the correlation matrix the scatterplots of the covariables are presented, the histogram for each covariable is presented in the diagonal and on the right the correlations between covariables are shown.

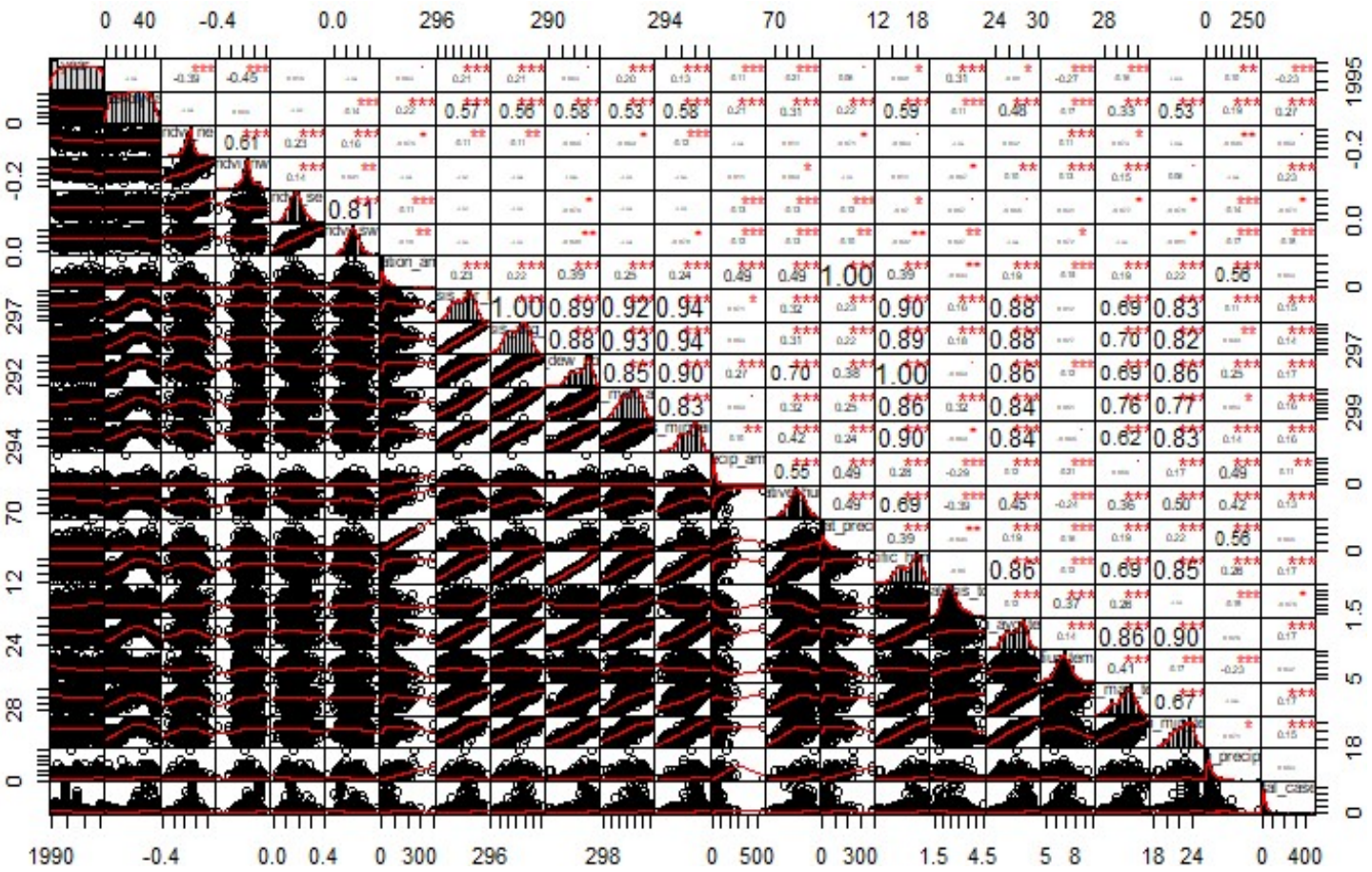


Figure 5: Correlation matrix between covariables.

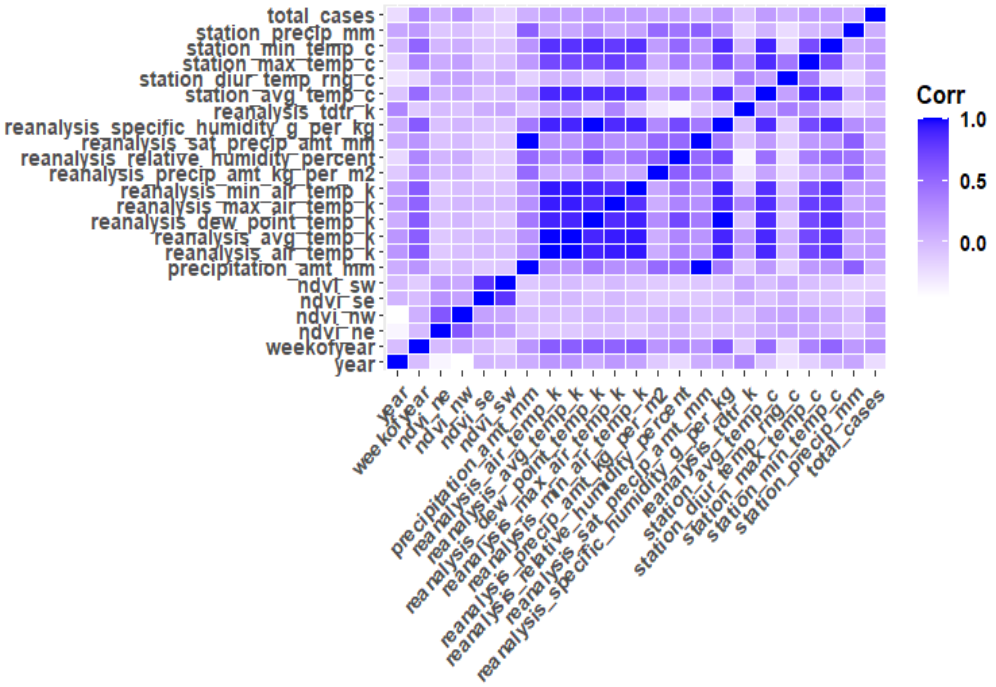


Figure 6: Heatmap of the correlation matrix.

Notice that the covariable *week_start_date* was removed from the dataset. This decision was taken based on the fact that having the covariables *year* and *weekofyear* would already provide sufficient information about the date/time at which the measurements were taken. By dropping this covariable the exact day at which the measurements were recorded was lost. However, the mosquito's life cycle is based on the time of the year and its associated weather conditions rather than on a specific day.

Inspection of Figure 5 confirmed the previous assertions, *i.e.*, the explanatory variables from the different sources measuring the same feature have a high correlation which means they provide the same information. Therefore, it seemed reasonable to consider only one of such covariables. Table 6 summarizes some of the information contained in the correlation matrix.

Feature Description	Covariates	Variance	Correlation
Minimum temperature	reanalysis_min_air_temp_k	1.671	0.828
	station_min_temp_c	2.334	
Maximum temperature	reanalysis_max_air_temp_k	1.576	0.757
	station_max_temp_c	2.851	
Average/Mean temperature	reanalysis_avg_temp_k	1.455	0.997
	reanalysis_air_temp_k	1.501	
Average/Mean temperature[a]	reanalysis_avg_temp_k	1.455	0.877
	station_avg_temp_c	1.968	
Diurnal temperature range	reanalysis_tdtr_k	0.243	0.367
	station_diur_temp_rng_c	0.706	

Table 6: Correlation between covariables measuring the same feature and respective variances.

[a]There are two entries for the average temperature since there are three covariates describing this quantity. Firstly, one is removed and then the two remaining are evaluated.

To proceed with the choice of which of the covariables should be kept some criterion must be applied. The chosen criterion to select between covariables measuring the same phenomenon is based on the comparison of the variances of such covariables. More specifically, the covariable to be retained is the one which has the lowest variance. However, there will be some exceptions to this criterion which will be explained in due time.

Considering only the first four rows of Table 6, which correspond to the features *Minimum Temperature*, *Maximum temperature* and *Average/Mean Temperature*, according to the values of the variance for each explanatory variable representing each given feature the covariables *reanalysis_min_air_temp_k*, *reanalysis_max_air_temp_k* and *reanalysis_avg_temp_k* were retained since they have the lowest variance. The retention of the covariables obtained resorting to the NOAA's NCEP Climate Forecast System Reanalysis was expected since this system has a high resolution in contrast to the NOAA's GHCN daily climate data weather station whose data was obtained from ground weather measurements.

Inspection of the correlation between the covariables measuring the feature *Diurnal temperature range*, presented on the fifth row of Table 6, revealed a low correlation between the two explanatory variables which was an unexpected result. Since the explanatory variables *reanalysis_tdtr_k* and *station_diur_temp_rng_c* both measure the same feature, one would expect a very high correlation between them. Moreover, given that the considered feature is a range of temperatures the units in which the covariables were measured (*Kelvin* and *Celsius*) do no longer influence the recorded values which supports the prior belief that a high correlation between these covariables should have been obtained. Nevertheless, since the measured feature was the same only one of the covariables was retained. Namely, the covariable *reanalysis_tdtr_k* since it has the lowest variance.

Up until this point in the explanatory variables selection process only covariables measured by the system NOAA's NCEP Climate Forecast System Reanalysis were retained. The retention of only covariables measured by one system might raise some limitations and influence the results obtained by the classifiers. This fact will be taken into account in the future selection of explanatory variables.

There is one feature that hasn't been analysed yet: the *Total precipitation*. According to Table 2, there are up to four variables measuring this quantity which makes its analysis more difficult. None of the covariables was removed at this stage, since they will be object of further investigation.

Further analysis of Figure 5 shows that there are other highly correlated explanatory variables which do not exactly measure the same phenomenon/feature, although being dependent of one another. Consider the covariables *reanalysis_dew_point_temp_k* and *reanalysis_specific_humidity_g_per_kg* whose information is summarized in Table 7.

Feature Description	Covariates	Variance	Correlation
Mean dew point temperature	reanalysis_dew_point_temp_k	2.558	0.998
Mean specific humidity	reanalysis_specific_humidity_g_per_kg	2.509	

Table 7: Correlation of the covariables measuring the features *mean dew point temperature* and *mean specific humidity* and respective variances.

According to Table 7, the correlation between the two covariables is 0.998 which was already expected since increasing the barometric pressure, *i.e.*, increasing the humidity at a specific location in this case measured in g per kg by the covariable *reanalysis_specific_humidity_g_per_kg*, leads to an increase in the dew point temperature, *i.e.*, an increase in the values of the covariable *reanalysis_dew_point_temp_k*. Figure 7 illustrates the referred relation between these covariables.

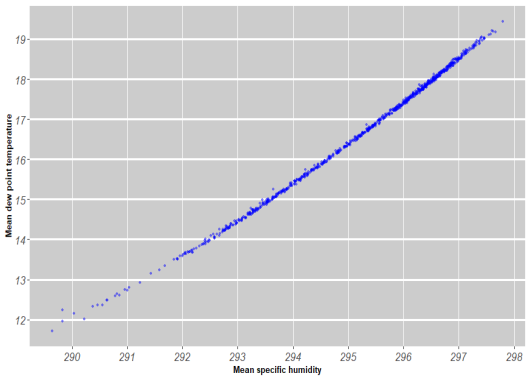


Figure 7: Scatterplot of *mean specific humidity* against *mean dew point temperature*.

The values of the variance of both covariables are practically identical as shown in Table 7. Therefore, the choice of one of the covariables should not be based on the variance criterion utilized so far. Furthermote, the scatterplots computed for each covariable against the number of *total cases* are practically identical which was expected, since both have a correlation of approximately 0.17 with the latter. Given that the *mean dew point temperature* is measured in *Kelvin* (which is in accordance with the majority of the selected covariables) and research about this topic revealed that the *mean dew point temperature* is a more accurate measure of humidity the covariable *reanalysis_dew_point_temp_k* was retained. Having removed the covariable *reanalysis_specific_humidity_g_per_kg* there remains the covariable *reanalysis_relative_humidity_percent* which also measures humidity. Comparison of the latter with *reanalysis_dew_point_temp_k* allows the construction of Table 8.

Feature Description	Covariates	Variance	Correlation
Mean dew point temperature	reanalysis_dew_point_temp_k	2.558	0.697
Mean relative humidity	reanalysis_relative_humidity_percent	11.433	

Table 8: Correlation of the covariables measuring the features *mean dew point temperature* and *mean relative humidity* and respective variances.

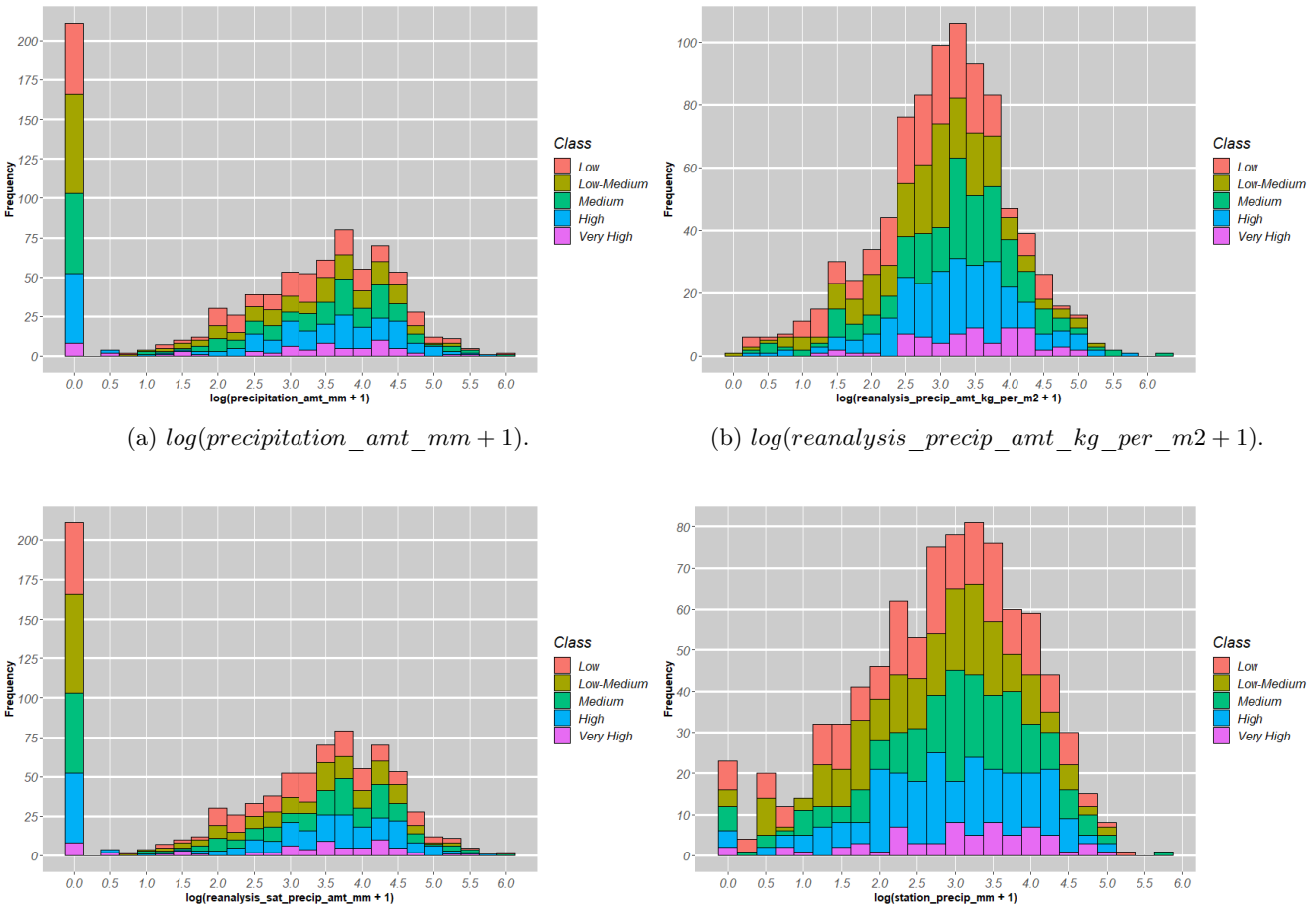
Applying the variance criterion (retain the covariate with the lowest variance) resulted in the retention of the explanatory variable *reanalysis_dew_point_temp_k*.

To sum up, inspecting Figure 15 one concludes that only 15 explanatory variables were retained. However, further study of the covariables measuring *total precipitation* is required since their distribution appears to be different from the remainder of the covariables. Moreover, retaining four covariables describing this feature is redundant and might provoke multicollinearity problems. Therefore, in the next section the transformation of the referred covariables was addressed.

2.3.3 *Total precipitation* covariables transformation

The selection of which of the covariates describing the feature *total precipitation* to be retained is not an obvious task. All covariables are approximately equally correlated with the variable *total cases*. Moreover, all their distributions were skewed to the left and the tails observed in the *QQ-plots* were not straight[2]. The *QQ-plot* and histogram for each of the four covariables were computed. The latter is presented in Figure 16 of appendix 9.2.2. In order to try to correct this problem the function $f(x) = log(x + 1)$ was applied to each covariable yielding the histograms presented in Figure 8.

[2]Heavy tails.



(a) $log(precipitation\_amt\_mm + 1)$.

(b) $log(reanalysis\_precip\_amt\_kg\_per\_m2 + 1)$.

(c) $log(reanalysis\_sat\_precip\_amt\_mm + 1)$.

(d) $log(station\_precip\_mm + 1)$.

Figure 8: Histograms for the transformed covariables measuring the feature *total precipitation*.

As observed in Figure 8 and in the computed *QQ-plots* there was a reduction of the skewness of the covariables, which could lead to an improvement of the classifier's performance and aid in the covariable selection. Furthermore, a significant reduction of the variance of each of these covariables was also achieved. Thus, an analysis of the correlation matrix was carried out hoping that the transformation applied to these covariables would help in the selection of the most appropriate ones to continue the study of the dataset.

2.3.4 *Total precipitation* selection

In section 2.3.2 some of the covariables were removed since they provided the same information regarding a given phenomenon. Capitalizing on the analysis in section 2.3.2 and on the transformation applied to the covariables describing the feature *total precipitation* in section 2.3.3, the heatmap of correlation matrix illustrated in Figure 9 was obtained.

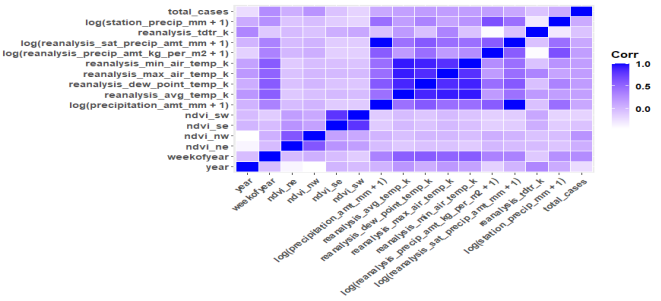


Figure 9: Heatmap of the correlation matrix with transformed covariables.

The correlation between the covariables $log(precipitation\_amt\_mm+1)$ and $log(reanalysis\_sat\_precip\_amt\_mm+1)$ as well as their variances are summarized in Table 9.

Feature Description	Covariates	Variance	Correlation
Total precipitation	$log(precipitation\_amt\_mm+1)$	2.931	0.999
	$log(reanalysis\_sat\_precip\_amt\_mm+1)$	2.939	

Table 9: Correlation of two of the covariables measuring the feature *total precipitation* and respective variances.

The application of the variance criterion was not appropriate in this situation as both covariables' variances are very similar. Recall that most of the retained variables were measured by the system NOAA's NCEP Climate Forecast System Reanalysis. Therefore, it was considered appropriate to retain the covariable $log(precipitation\_amt\_mm+1)$ since its recorded values were measured using a different source. By retaining the mentioned covariate it was possible to mitigate some potential measuring errors from each of the recording systems. This is due to the fact that the dataset to be analysed will not be constituted solely on values from one measuring system, which could lead to gross errors if the equipment of the measuring system would have been, for example, decalibrated.

Consequently, there are still three covariables describing the feature *total precipitation* which is not optimal due to multicollinearity issues. To aid in deciding which covariables should be removed the correlation among each combination of these explanatory variables was computed and is presented in Table 10.

Feature Description	Covariates	Variance	Correlation
Total Precipitation	$log(station\_precip\_mm+1)$	1.290	0.459
	$log(precipitation\_amt\_mm+1)$	2.931	
	$log(reanalysis\_precip\_amt\_kg\_per\_m2+1)$	0.918	0.576
	$log(precipitation\_amt\_mm+1)$	2.931	
	$log(reanalysis\_precip\_amt\_kg\_per\_m2+1)$	0.918	0.645
	$log(station\_precip\_mm+1)$	1.290	

Table 10: Correlation between remaining covariables measuring the feature *total precipitation* and respective variances.

The choice of which covariables to retain is not direct. According to the variance criterion applied previously, the most suitable variable to describe the *total precipitation* would be *log(reanalysis_precip_amt_kg_per_m2 + 1)*. However, this decision would mean that the information coming from NOAA's CDR PERSIANN satellite and NOAA's GHCN daily climate data weather station wouldn't be used at all. Besides, the values of the variances for each covariable are of the same order of magnitude so their differences may not lead to better or worst performance of the models. Selecting the covariable *log(precipitation_amt_mm + 1)* would suppose omitting completely the information provided by NOAA's GHCN daily climate data weather station but selecting the covariable *log(station_precip_mm + 1)* would provoke the same effect with the NOAA's CDR PERSIANN.

Therefore, three distinct datasets were created with each covariate describing the *total precipitation*.

2.4 Final datasets

Finally, three datasets were created and its features are listed in Table 11. The $*$ means that the transformation $f(x) = log(x + 1)$ was applied to the referred covariable.

Dataset 1	Dataset 2	Dataset 3
year	year	year
weekofyear	weekofyear	weekofyear
ndvi_ne	ndvi_ne	ndvi_ne
ndvi_nw	ndvi_nw	ndvi_nw
ndvi_se	ndvi_se	ndvi_se
ndvi_sw	ndvi_sw	ndvi_sw
reanalysis_avg_temp_k	reanalysis_avg_temp_k	reanalysis_avg_temp_k
reanalysis_dew_point_temp_k	reanalysis_dew_point_temp_k	reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k	reanalysis_max_air_temp_k	reanalysis_max_air_temp_k
reanalysis_min_air_temp_k	reanalysis_min_air_temp_k	reanalysis_min_air_temp_k
reanalysis_tdtr_k	reanalysis_tdtr_k	reanalysis_tdtr_k
precipitation_amt_mm*	reanalysis_precip_amt_kg_per_m2*	station_precip_mm*

Table 11: Features of the analysed datasets.

Each of the three datasets was utilized to construct several classifiers to predict the total number of infections by Dengue. The methodology used to choose the dataset and its respective model that best predicts the severity of future pandemics is described in the next section.

3 Methodology

This study aimed to classify the severity of the dengue pandemics in San Juan. To do so, several types of classifiers were used: K-Nearest Neigbors (KNN), Naïve Bayes, Linear Regression of an Indicator Matrix, Linear Discriminant Analysis (LDA), Robust LDA, Quadratic Discriminant Analysis (QDA), Robust QDA, Logistic Regression, Decision trees, Conditional trees, Random Forests and Support Vector Machine (SVM). The performance of these classifiers was compared and the best one was selected and tested on the test dataset.

The techniques used when fitting the models and evaluating their performance were the Hold-out method and repeated K-fold cross validation. The latter method appeared to be the most suitable one to this study. Obviously, repeated K-fold cross validation gets more accurate estimations for the prediction error than simply K-fold cross validation or re-substitution. In addition, according to the literature, leave-one-out and bootstrap validation methods are rather time consuming and are more suitable when working with smaller datasets [Del08].

The procedure to select the most appropriate model capable of classifying the severity of the Dengue pandemics was the following.

Firstly, a classifier of each type was fitted for each one of the three datasets presented in Table 11 and their performances were compared by measuring their accuracy values. Moreover, if a classifier depended on hyper-parameters, multiple models were fitted using different values for these hyper-parameters with the purpose of determining the most suitable one.[3]

Secondly, the dataset which yielded better performance results than the other two was selected to further continue the study of the problem. In order to improve the performance of the classifiers applied to the selected dataset, a thorough analysis of the latter was carried out. Specifically, principal component analysis (PCA) was applied to the selected dataset and a classifier of each type was fitted to a *"new"* dataset given by the scores of each observation on the selected principal components. Furthermore, possible atypical observations on the selected dataset were detected applying the method ROBPCA and were, subsequently, removed to understand their influence on the classifier's performance.

Thirdly, further performance measures for each model were obtained by application of the Hold-out method and were, posteriorly, evaluated and compared using the ROC Space. The best model was selected and trained over the whole training set utilizing repeated K-fold cross validation.

Finally, the model obtained after training was used to classify each observation of the test dataset. The results were registered and critically analysed.

[3]For instance, multiple models are fitted using KNN algorithm varying k in order to find its optimal value.

4 Dataset selection based on classifier's performance

From the exploratory data analysis resulted the three different datasets presented in Table 11. As stated, the choice of which of the covariables, describing the feature *total precipitation*, to be picked was not obvious. Therefore, each referred dataset was tested on each of the classifiers referred in section 3. This set of tests aimed to find which was the dataset for which the classifiers yielded the best results. As an extensive analysis of the response of each classifier to the different datasets would be unfeasible, only the classifier's accuracy was computed. The results obtained using repeated K-fold cross validation on the non-standardized datasets are presented in Table 12.

Classifier		Acccuracy		
		Dataset 1	Dataset 2	Dataset 3
KNN		0.451	0.470	0.455
Naïve Bayes		0.360	0.359	0.369
LDA		0.322	0.327	0.324
LINDA[a]		0.358	0.355	0.349
QDA		0.328	0.331	0.341
QdaCov[b]		0.344	0.357	0.363
Indicator Matrix		0.253	0.284	0.249
Logistic Regression		0.327	0.321	0.323
Decision Tree		0.410	0.413	0.413
Conditional Tree		0.446	0.452	0.453
Random Forest		0.536	0.537	0.551
SVM	Linear Kernel	0.326	0.334	0.333
	Radial Kernel	0.376	0.379	0.384
	Polynomial Kernel	0.394	0.391	0.407

Table 12: Classifiers' accuracy results for each dataset.

Analysing Table 12 one can check that no dataset was substantially better than the remaining ones. Thus, choosing the dataset which yields the best results is not straightforward. Moreover, the obtained accuracies for all classifier algorithms on each dataset can be considered quite low, *i.e.*, the classifiers have a reduced ability to assign observations to the corresponding true class.

Further inspection of Table 12 revealed that there was an increase in the accuracy when the robust version of the Linear and Quadratic Discriminant Analysis (LDA and QDA) classifiers were applied, *i.e.*, LINDA and QdaCov. Therefore, their application was more suited to the analysis of the datasets. The analysed datasets possibly contain some atypical observations which might have influenced the acquired results.

Furthermore, the highest values of accuracy were obtained when classification tree methods were applied. Decision and Conditional Tree algorithms are prone to overfitting, especially because the computed trees were particularly deep, which could have led to a false increase in the values of accuracy. However, Random Forest algorithms are a strong modelling technique and much more robust than Decision and Conditional Trees. The Random Forest algorithm aggregates many decision trees to limit overfitting as well as error due to bias and therefore yields more accurate results. Consequently, the values of accuracy obtained by applying this method may be regarded as valid.

The decisions taken during the exploratory data analysis and the datasets obtained as its end result seemed suitable. However, possible mistakes could have been committed when choosing which covariables to retain. In order

[a] Robust version of Linear Discriminant Analysis [TF09].
[b] Robust version of Quadratic Discriminant Analysis [TF09].

to check that possibility and assess if the classifiers would yield more satisfactory results each classifier was applied to the original dataset. The obtained results were extremely similar to those presented in Table 12. Hence, the usage of the datasets from Table 11 is appropriate and advantageous since it was possible to reduce the dimensionality of the problem and still obtain similar results as if the original dataset had been analysed, *i.e.*, the loss of information caused by the removal of some covariables did not affect the performance of the considered classifiers.

The low values of accuracy might have derived from the fact that the measured values for each observation on each covariable were measured in different units. In order to test such hypothesis, the three datasets presented in Table 11 were standardized and the classifiers were implemented again using the referred datasets. The obtained results for the accuracy of each classifier on each standardized dataset were extremely similar to those obtained in Table 12. Therefore, the heterogeneity of measuring units does not seem to have a significant impact on the classifiers ability to predict the class of an observation.

Alternative problems might be the cause of the low values of accuracy. More specifically, such problems might be:

- The classifier algorithms might not be able to correctly predict and learn due to the large number of covariates of the analysed datasets;
- The presence of possible outliers and lack of robustness of some of the considered classifiers to such atypical observations;
- Multicollinearity between some covariables as observed in Figure 9.

In the following sections each of these problems will be addressed. Distinct techniques will be utilized to understand the influence of such problems on the results and ultimately try to mitigate them. The first problem will be handled by using Principal Component Analysis in order to reproduce the total or most of the system's variability with a small number of principal components. The second problem will be addressed using the method ROBPCA proposed by Hubert et al. [HRB05], which will be utilized to determine which observations might be potential outliers.

Nevertheless, an extensive and unfeasible analysis would result if all datasets presented in Table 11 were to be analysed. Therefore, only one of such datasets will be chosen and analysed. By analysing Table 12, it appears that the classifiers prediction ability was best when they were applied to Dataset 3 which is the dataset that contains the covariable *station_precip_mm*. Therefore, the subsequent analysis on the following sections will only take into consideration Dataset 3. Aside from seemingly producing the best results the choice of Dataset 3 is appropriate due to other reasons. More specifically, by retaining the covariable *station_precipitation_mm* not only covariables measured with the NOAA's NCEP Climate Forecast System Reanalysis are considered, although the information from NOAA's CDR PERSIANN is disregarded. Moreover, it was possible to corroborate that the differences in the variances of the covariables measuring the feature *total precipitation* do not lead to better or worse performances of the model. So, although the covariate *station_precipitation_mm* does not possess the smallest variance this fact is not expected to influence future results.

5 Dataset treatment and classification performance

In order to improve the performance of the classifiers on the previously selected dataset, two techniques were applied to the referred dataset. Posteriorly, the obtained results were compared to the ones presented in Table 12.

5.1 Principal Component Analysis and classification performance

As a result of the explanatory data analysis it was possible to obtain a dataset with lower dimensions by selecting the covariables considered as most relevant to the analysis of the dataset. In order to check if further dimensionality reduction of Dataset 3 would improve the accuracies of the classifiers obtained in Table 12, Principal Component Analysis (PCA) was applied. This dimensionality reduction through PCA was concerned with explaining the variance-covariance structure of the covariables of Dataset 3 through a few linear combinations of these covariables.

The main goal was to find the PC's that explained the variability of the dataset. Since the variables were measured in different scales, standardised variables were utilized. Consequently, the PC will depend on the correlation matrix of the variables. Only the function *prcomp* [RCo] was tested since, according to the literature, it is more robust to outliers.

	λ_i	$\frac{\lambda_i}{12}$	$\sum_i \frac{\lambda_i}{12}$
PC_1	4.215	0.352	0.352
PC_2	2.150	0.179	0.531
PC_3	1.871	0.156	0.687
PC_4	1.246	0.104	0.790
PC_5	0.766	0.064	0.854
PC_6	0.538	0.045	0.899
PC_7	0.469	0.039	0.938
PC_8	0.376	0.031	0.969
PC_9	0.170	0.014	0.984
PC_{10}	0.105	0.009	0.992
PC_{11}	0.064	0.005	0.998
PC_{12}	0.027	0.002	1.000

Table 13: Variance of the i^{th} PC and cumulative variance of the i^{th} PC.

The variance of the i^{th} PC, the proportion of the variance explained by each PC and the cumulative proportion explained up to the i^{th} PC, are presented in Table 13. On one hand, by analysing Table 13, it was possible to conclude that the first five PC's explained approximately 85% of the total variability of the data. On the other hand, only the first 3 PC's possessed a variance greater than 1. Since the first 3 PC's only accounted for approximately 68% of the total variability, the consideration of the first 5 PC's appeared to be appropriate. However, since during the exploratory data analysis some covariables were removed there was some loss of information. Therefore, to avoid analysing an extremely uninformative dataset the first 9 PC's will be utilized to generate a new dataset, given by the scores for each observation on each of the first nine principal components. As observed in Table 13, the first 2 PC's explain approximately 53% of the total variability, which can be considered a high value. Thus, the projection of the data onto the 2 first PC's could be appropriate.

Following the previous analysis the loadings for the first 9 PC's were computed. In Table 14 the loadings for the first 2 PC's are presented.

	PC_1	PC_2
year	0.102	0.330
weekofyear	0.332	-0.089
ndvi_ne	-0.088	-0.493
ndvi_nw	-0.046	-0.480
ndvi_se	-0.058	-0.444
ndvi_sw	-0.075	-0.417
reanalysis_avg_temp_k	0.466	-0.075
reanalysis_dew_point_temp_k	0.457	-0.085
reanalysis_max_air_temp_k	0.446	-0.094
reanalysis_min_air_temp_k	0.459	-0.064
reanalysis_tdtr_k	0.035	0.019
$log(station\_precip\_mm + 1)$	0.152	0.110

Table 14: Loadings for the first two principal components.

The interpretation of the first two principal components is as follows:

- The first principal component exhibits higher absolute values for the covariables *reanalysis_avg_temp_k*, *reanalysis_dew_point_temp_k*, *reanalysis_max_air_temp_k* and *reanalysis_min_air_temp_k*. This PC can be interpreted as a weighted average of the mentioned variables, *i.e.*, this PC is concerned with analysing each observation with respect to its measured values with the NOAA's NCEP Climate Forecast System Reanalysis (excluding the mean diurnal temperature range feature). Moreover, observations with high values of mean average temperature, mean dew point temperature, maximum temperature and minimum temperature will have high values in this PC.
- The second PC admits larger absolute weights to the covariables *ndvi_ne*, *ndvi_nw*, *ndvi_se* and *ndvi_sw*. Therefore, this PC can be interpreted as a weighted average of the mentioned variables, *i.e.*, this PC is concerned with analysing each observation with respect to its measured values with the NOAA's CDR Normalized Difference Vegetation Index measurements.

The scores for each observation were computed for the first 2 PC and projected into the referred components resulting in Figure 10.

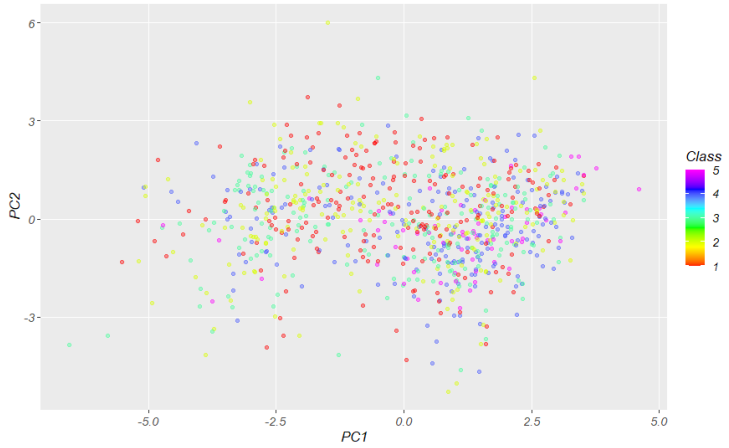


Figure 10: Plot of the first two principal components.

Inspection of Figure 10 revealed that there is no apparent separation between observations belonging to distinct classes, *i.e.*, observations belonging to the same class did not form groups/clusters characterized by a certain set of features/values. This observation is consistent with the low values of accuracy obtained for the different classifiers. Since there is little separation between classes the classifiers were not able to comprehend which were the set of features that characterized a specific class and, consequently, their prediction ability was impaired.

By retaining the first 9 PC's approximately 98% of the total variability of the dataset was explained and, consequently, approximately 2% of the dataset's information was lost, which was considered to be a reasonable value. Moreover, it was possible to reduce the dataset's dimensionality by three dimensions which might prove advantageous, provided that the values of the classifier's accuracy increase.

Therefore, the scores of each observation on the first 9 PC's were utilized to construct a new dataset. Each of the classifiers mentioned in the section 3 was then applied to this new dataset and their accuracies were measured allowing the compilation of Table 15.

Classifier		Accuracy
KNN		0.362
Naïve Bayes		0.377
LDA		0.342
LINDA[a]		0.364
QDA		0.353
QdaCov[b]		0.364
Indicator Matrix		0.237
Logistic Regression		0.376
Decision Tree		0.331
Conditional Tree		0.343
Random Forest		0.413
SVM	Linear Kernel	0.333
	Radial Kernel	0.395
	Polynomial Kernel	0.413

Table 15: Classifiers' accuracy results on scores of Dataset 3.

Observing the values of the accuracies of each classifier presented in Table 15 one verifies that they are relatively similar to those shown in the third column of Table 12. Therefore, the application of PCA does not seem to provide many advantages to the analysis of the problem, except for the possibility of further dimensionality reduction. With the application of PCA one would expect an increase in the values of the accuracies for each classifier. This is due to the fact that with fewer variables it was possible to retain approximately all of the dataset's information. Consequently, with fewer dimensions it would be possible to explain most of the dataset's variability which would, in principle, enhance the classifier's ability to understand its structure.

5.2 Outliers and classification performance

In this section the influence of possible outliers on the classifier's performance was studied. In order to do that Hubert et al. defined a diagnostic plot which helps to distinguish between the regular observations and the different types of outliers [HRB05]. This plot is based on the score distances and orthogonal distances computed for each observation. Utilizing the default cutoff values for the score and orthogonal distances it was possible to obtain the diagnostic plot presented below in Figure 11.

[a] Robust version of Linear Discriminant Analysis [TF09].
[b] Robust version of Quadratic Discriminant Analysis [TF09].

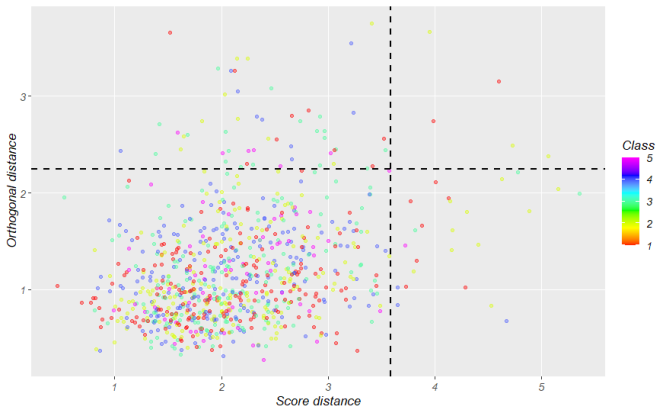


Figure 11: Outlier map obtained using the ROBPCA method.

In the context of PCA three types of outliers can be distinguished. Nonetheless, such distinction was not considered in this report since it would result in an extensive analysis out of the scope of this study. According to Figure 11, 82 observations were considered as potential outliers. Moreover, the atypical observations came from all the five different classes, although mostly from classes Low-Medium and Medium.

It was of interest to check the possible influence of such outliers in the classifier's performance. Consequently, the referred outliers were removed from Dataset 3 generating a *"new"* dataset. Moreover, PCA was posteriorly applied to the latter dataset. Finally, all the mentioned classifiers in section 3 were implemented using both referred datasets and the value of each classifier's accuracy was measured. It is important to refer that the relative proportions of instances belonging to each class remained approximately similar.

After applying each classifier to each of the datasets without outliers the accuracy results were compared to those obtained in the third column of Table 12. Such comparison led to the conclusion that the influence of the potential outliers was negligible. Although the majority of the classifiers' performances increased, that increase was residual. Moreover, such increase did not compensate for the fact that possibly important information was lost with the removal of the outlying observations. Therefore, the removal of the outliers did not appear to be appropriate and necessary to obtain the best possible classifier.

6 Selection of classifier

The aim of section 5 was to apply dimensionality reduction techniques and determine the influence of outliers on Dataset 3 with the purpose of obtaining the most satisfactory classification performances. Comparison of the results yielded in section 5 with the ones obtained when utilizing the original Dataset 3 allowed to conclude that neither further dimensionality reduction nor the removal of outliers was required to improve classification performance. As a consequence, on the subsequent sections the classifier's performance measures will be acquired using the original Dataset 3.

6.1 ROC Space and classification performance

The ROC Space is a coordinate system used for visualising the classifier's performances. It is defined by plotting the false positive rate (FPR) in the x axis and the true positive rate (TPR) in the y axis. Moreover, the TPR is equivalent to the sensitivity and the FPR is equivalent to 1 – specificity. Since the total cases variable was split into five classes, a multiclassification problem is being addressed. As a result, the values of the sensitivity and specificity are specific and different for each class. Therefore, the TPR and the FPR were calculated as follows:

$$TPR = \sum_{i=1}^{5} (sensitivity_i \times p_i) \tag{4}$$

$$FPR = \sum_{i=1}^{5} \Big[(1 - specificity_i) \times p_i \Big] \tag{5}$$

where p_i is the probability that a given observation belongs to class i which, in this case, is given by the relative proportions of each class.

In order to avoid computationally complex and time-consuming computations the sensitivity and specificity of each class for each classifier were obtained using the Hold-out method. Then equations (4) and (5) where applied to compute the TPR and FPR for each classifier, respectively. All the results were then utilized to construct the ROC Space presented in Figure 12.

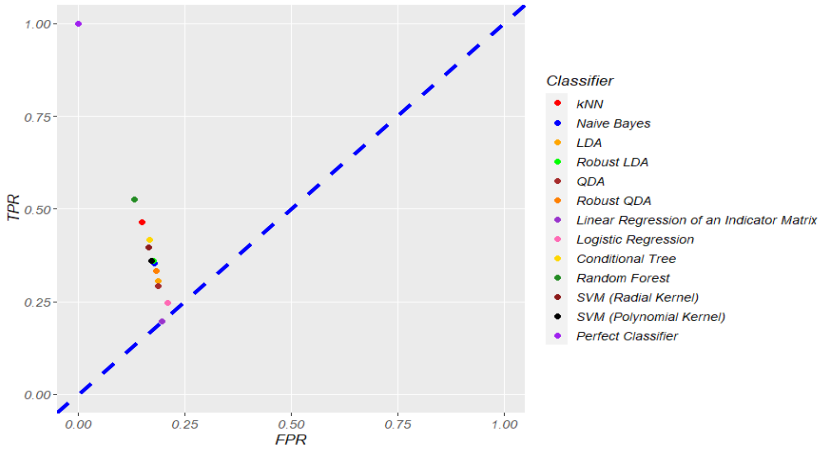


Figure 12: ROC Space with predictions for each classifier.

There are some features of interest in the ROC Space. Namely, the best possible prediction method would yield a point with coordinates $(0, 1)$ in the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). Therefore, the point $(0, 1)$ would be occupied by a perfect classifier as seen in Figure 12. A random classifier would give a point along the blue diagonal line. The blue diagonal line divides the ROC

space, *i.e.*, points above the diagonal represent *"good"* classification results (better than random) and points below the line represent *"bad"* results (worse than random).

Analysing Figure 12 one may check that, with the exception of the classifier obtained with Linear Regression of an Indicator Matrix, all classifiers had a *"good"* performance, *i.e.*, the assignment of classes was better than if it was done randomly. Furthermore, all points representing each classifier were quite distant from the point $(0, 1)$ meaning that none of the obtained classifiers could be considered a perfect classifier, which was something already expected given that the values of the accuracies obtained in the previous sections were not close to one.

A thorough inspection of Figure 12 revealed that the Random Forest classifier had the highest TPR score and lowest FPR score. Therefore, it may be considered as the best classifier for the analysed dataset, which is in agreement with its value of accuracy presented in the third column of Table 12. In the next section, a careful analysis of the Random Forest classifier was carried out.

6.2 Analysis of the best classifier

Taking into consideration the values of accuracy and the analysis of the ROC Space it was determined that the Random Forest provided the best classifier for the problem's dataset. Different parameters of the classifier were analysed and studied in order to obtain the best possible classifying results. Such study can be consulted in appendix 9.3. The evaluation of the classifier's performance was accomplished by utilizing repeated K-fold cross validation. More specifically, the dataset was split into $K = 5$ blocks and the number of repetitions was 3. Thus, the obtained classifier's performance measures may be considered validated. Such performance measures are summarized in Table 16.

	Low	Low-Medium	Medium	High	Very High
Sensitivity	0.698	0.443	0.434	0.634	0.786
Specificity	0.920	0.833	0.830	0.902L	0.959
Accuracy			0.551		

Table 16: Performance measures for the Random Forest model.

Analysing the sensitivity performance measure, one may verify that the obtained values were mediocre. In fact, one may state that there were a lot of observations which belonged to a certain class which were labelled to a different class. Namely, the values of sensitivity for classes *Low-Medium* and *Medium* may be regarded as quite low which was surprising. Since these classes had a large number of instances one would expect the classifying algorithm to correctly assign the majority of this observations to the respective correct class. However, that was not the case which was an unexpected fact. In contrast, the sensitivity value obtained for class *Very High* was very high taking into account the fact that this class had a reduced number of instances when compared to the remaining ones. This fact was also unexpected. Nevertheless, it provides assurance that if an observation belongs to class *Very High* it will, most certainly, be assigned to class *Very High*. Hence, one may state that the classifier tends to predict better if, at a given time, a very high number of Dengue infections should be expected.

Inspection of the specificity performance measure reveals that the obtained results were very satisfactory. Indeed, for classes *Low*, *High* and *Very High*, the specifity values were above 90%. Thus, the majority of observations belonging to a certain class were classified to that same class. This is of utmost relevance since for a given time if the severity of infections by Dengue is classified into one of the five levels (*Low*, *Low-Medium*, *Medium*, *High*, *Very High*) then with very high confidence one may say that the severity of infections will, in fact, be at the predicted level.

Finally, the value obtained for the accuracy performance measure was mediocre. Although, an extremely high accuracy level was not achieved, it was considered satisfactory taking into consideration the complexity of the problem in hand.

7 Application of best classifier on the test dataset

To sum up, in the previous sections it was possible to select the classifier which yielded the best performance measures and, simultaneously, validate its results. Finally, the referred classifier was utilized to predict the severity of Dengue infections of each observation of the test dataset.

Therefore, the classifier obtained by repeated K-fold cross validation was applied to the test dataset. Consequently, a class was assigned to each of the 69 observations of the test dataset. The number of observations assigned to each class is presented in Figure 13.

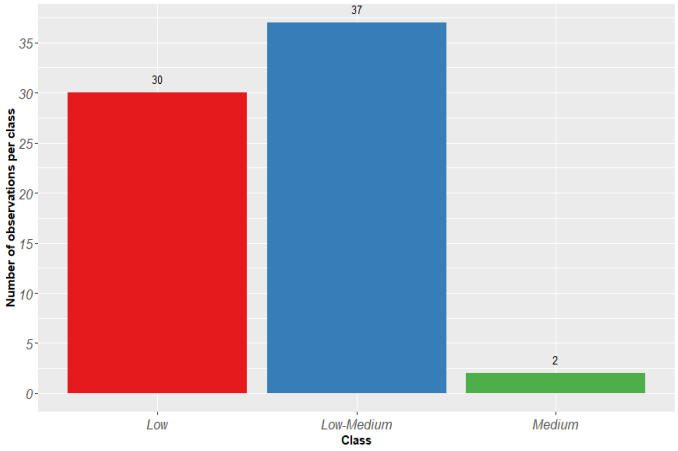


Figure 13: Class frequencies for the test dataset.

Analysis of Figure 13 revealed that the majority of the observations were assigned to class *"low"*, low severity of infections by Dengue, and class *Low-Medium*, low to medium severity of infections by Dengue. Taking into consideration the values of the sensitivity obtained in Table 16, one may conclude that, probably, some of the observations which were assigned to class *Low-Medium* should not in fact have been classified to that class due to the low level of sensitivity of the classifier for that class. The same reasoning may be applied to the assignment of observations to class *Low*. However, since the classifier has a higher sensitivity to this class the labelling of instances is less prone to errors.

Moreover, one may check that no observations were assigned to classes *High* and *Very High*. The fact that no observations were assigned to classes *High* or *Very High* does not necessarily mean that the classifier does not have the ability to predict such cases. In fact, considering the values of sensitivity and specificity obtained for class *Very High* in Table 16, one may conclude that if no observations were labelled to this class then, with a high level of confidence, it was due to the fact that no observations carried a high or very high severity of Dengue infections.

8 Conclusions and discussion

In this project a classification study of the Dengue infection severity in the population of San Juan was carried out.

Firstly, an exploratory data analysis was performed. Due to the fact that the datasets possessed missing values an imputation method was applied. Following this procedure, a study of the response variable corresponding to the total number of infections by Dengue was conducted. The analysis of the total number of infections by Dengue suggested that five classes describing different levels of the severity of infections should be considered. Thus, each observation was assigned to one of five levels of Dengue infection severity *"Low"*, *"Low-Medium"*, *"Medium"*, *"High"*, *"Very High"*. Posteriorly, a meticulous inspection of the explanatory variables was performed. Such inspection led to the removal of several covariables. Concretely, covariables which measured the same feature and had high correlation with the remaining covariates were subject to analysis and selection. The performed selection process aimed to mitigate possible multicollinearity issues and eliminate possibly irrelevant covariables to the classification procedure. Ultimately, three distinct datasets were obtained each containing a different covariate measuring the feature total precipitation.

Secondly, given that the choice of one of the datasets was not straightforward each one was tested on several classifiers. Therefore, it was possible to select the dataset which would, most probably, produce the most accurate classifier. Nevertheless, the performance of the classifiers on each dataset was extremely similar. Therefore, the decision to select the dataset containing the covariable *station_precip_mm* was taken due to previously explained arguments.

Thirdly, two methods were applied to the dataset to check if an increase in classification performances would occur. Namely, principal component analysis was applied to the dataset. Further dimensionality reduction was achieved since by retaining only nine principal components it was possible to explain approximately all of the dataset's variability. However, the newly found linear combinations of the covariables and associated dimensionality reduction did not lead to an increase in the performance of the classifiers. Subsequently, a robust principal component analysis as proposed by Hubert et al. was applied to the dataset in order to detect possible outlying observations. The detection and posterior removal of the outliers enabled the construction of a new dataset. The new dataset was then tested on the different classifiers yielding results similar to those obtained previously. Principal component analysis was also applied to the outlier free dataset resulting in a new dataset, which when tested on the different classifiers did not lead to an improvement of their performances. Therefore, neither the dataset obtained after PCA nor the dataset obtained after ROBPCA were selected for further study.

Fourthly, other relevant classifier performance measures were studied, such as the sensitivity and specificity. The comparison between such measures for each method was made by resorting to the ROC Space. It was concluded that none of the classifiers could be considered a perfect classifier. Thus, regardless of the selected classifier, the predictions would never be completely accurate, *i.e.*, for a new instance the classification results should be critically analysed. Following the referred comparison, the best classifier was determined to be given by the method Random Forest. Therefore, the parameters of the model were studied in detail in order to obtain the best possible classifier. The performance measures of the best classifier were then studied and interpreted. An accuracy of 0.551 was obtained. Some peculiarities of the dataset, which will be explained below, might explain the low level of accuracy. Analysis of the sensitivity and specificity of the classifier led to the conclusion that the predictions for classes *"Low"*, *"High"* and *"Very High"* are more accurate than for classes *"Low-Medium"*, *"Medium"*. Hence, if a new instance was, for example, classified as being characteristic of a Dengue severity level *"Low-medium"* or *"Medium"* the results should be confronted with other suitable classifiers and double-checked. Therefore, the level of confidence one may place in the predictions of the referred classifier should never be extremely high.

Finally, the best Random Forest classifier was utilized to predict the level of severity of the new observations given by the test dataset. No instances were labelled as belonging to a *"High"* or *"Very-High"* level of Dengue severity which is, most likely, an accurate assumption. Nonetheless, the instances assigned to a level of severity *"Low-medium"* should be meticulously reanalysed.

Naturally, the present study had some limitations. The analysed dataset possessed a temporal component of

correlation, *i.e.*, the covariables describing the time measures were intercorrelated. Specifically, the features unrelated with time were measured for different days of different weeks for some given year. Therefore, the measured features on one day will obviously influence the measures taken in the next days that is, the observations are not independent. The Dengue's mosquito lifecycle and ultimately the severity of infections will be determined by several factors and specificities which occur as the days go by and influencing one another and, as such, being highly correlated. However, the tested classification methods were not able to capture the correlation in time of the covariates. Therefore, this fact might be the basis for the mediocre results. Moreover, the analysed dataset should be analysed as a prediction problem. In order to model a response such as the severity of Dengue infections (without transforming it into a categorical variable) one should apply a Poisson regression. To sum up, for future work it is advised to apply regression models, such as, Poisson or Fused Lasso regression models, in order to predict the response variable.

In conclusion, having in mind the explained limitations and restrictions, the obtained model yielded satisfactory results, without prejudice to the existence of other models more appropriate to the study of the problem.

9 Appendix

9.1 Predictive Mean Matching

Predictive mean matching is a way to do multiple imputation for missing data. Especially, PMM is an attractive method for imputing quantitative variables that are not necessarily normally distributed.

PMM stands out above the standard methods based on linear regression and the normal distribution because it produces values that are much more like real values. Indeed, if the original data is skewed, discrete or bounded, the imputed values will also be skewed, discrete or bounded. Below is a brief description of how PMM works.

Let X be a variable that has some missing data, and a set of variables Z (with no missing data) that are used to impute X. Then, the method consists in following the steps shown below. **[PMM]** Let $n, k \in \mathbb{N}$.

1. For cases with no missing data, estimate a linear regression of X on Z, producing a set of coefficients β.
2. Make a random draw from the posterior predictive distribution of β, producing a new set of coefficients $\widetilde{\beta}$.
3. Using $\widetilde{\beta}$, generate predicted values for X for all cases, both those with data missing on X and those with data present.
4. For each case with missing X, identify a set of k cases with observed X whose predicted values are close to the predicted value for the case with missing data.
5. From among those close cases, randomly choose one and assign its observed value to substitute for the missing value.
6. Repeat step 2 to 5 n times and, for each case with missing X, take the mean of the values determined in step 5.

To implement this method, the R MICE package was used with $k = 5$ and $m = 3$.

9.2 Figures

9.2.1 Total cases of Dengue

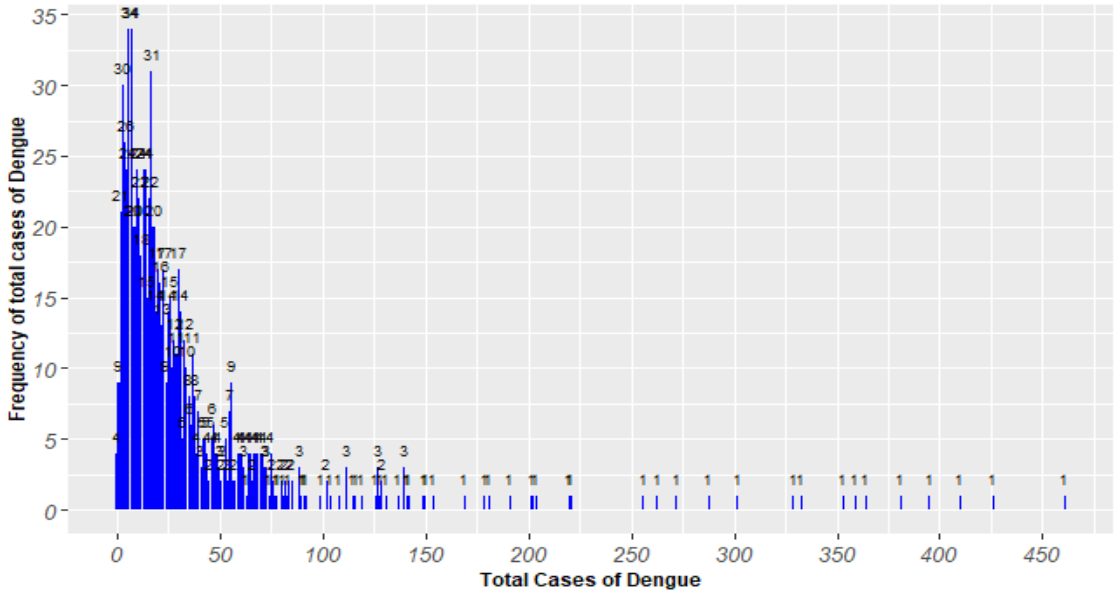


Figure 14: Frequency of the *total cases*.

9.2.2 Study of the explanatory variables

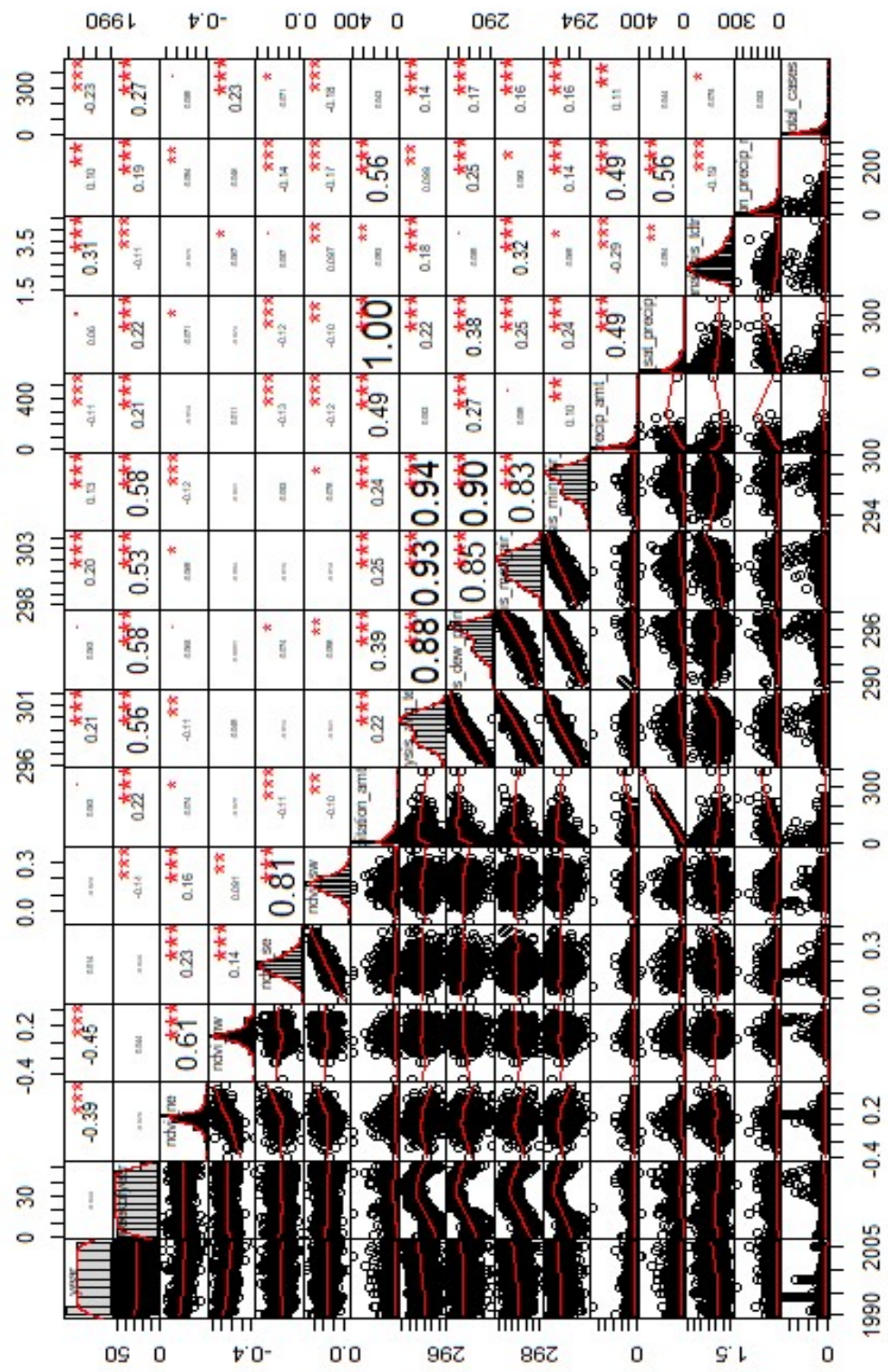
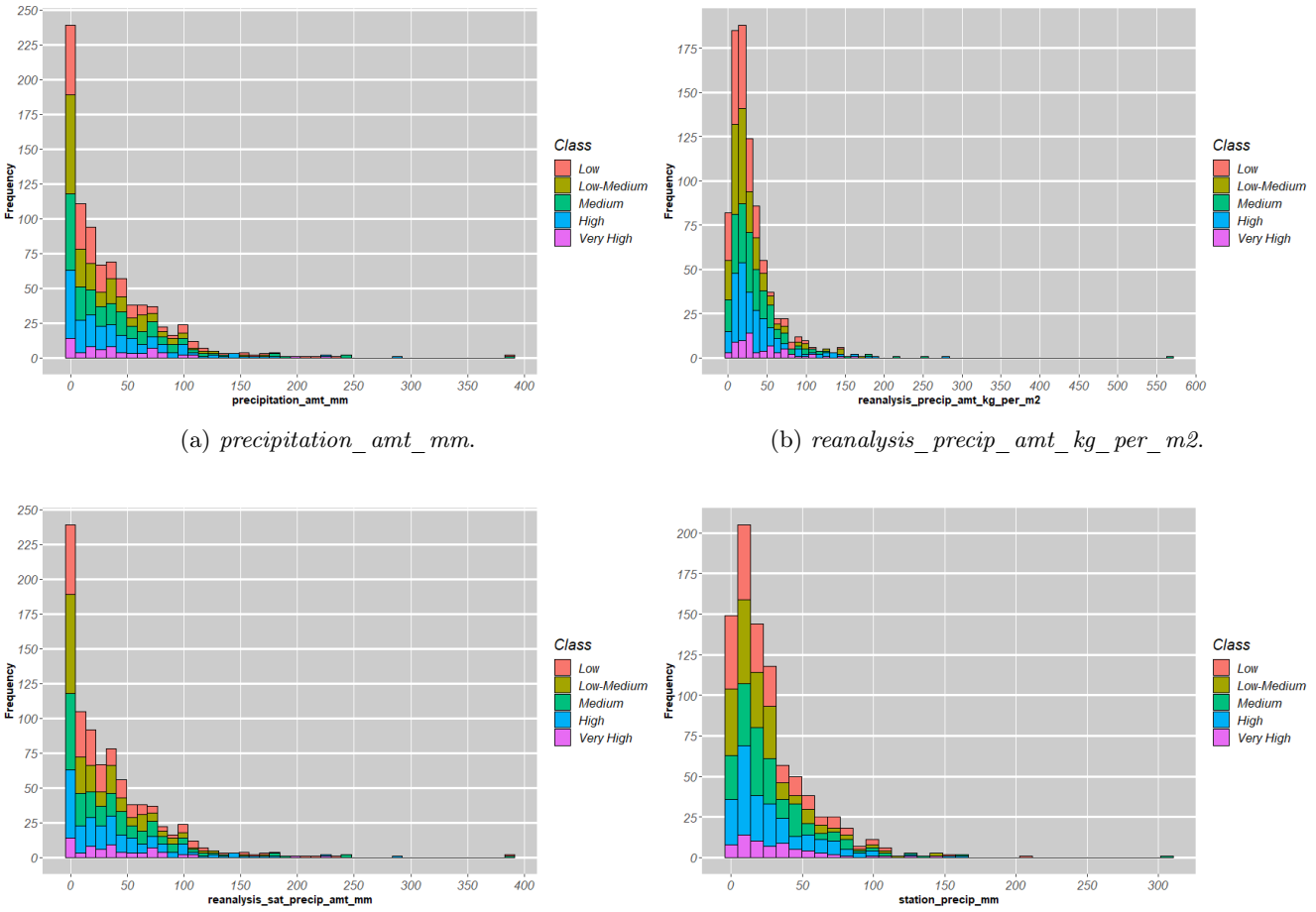


Figure 15: Correlation matrix between the retained covariables.



(a) *precipitation_ amt_ mm.*

(b) *reanalysis_ precip_ amt_ kg_ per_ m2.*

(c) *reanalysis_ sat_ precip_ amt_ mm.*

(d) *station_ precip_ mm.*

Figure 16: Histograms for the covariables measuring the feature *total precipitation*.

9.3 Random Forests models

The random forest model is a powerful tool to reduce overfitting in decision trees. The basic idea of the random forest algorithm is to combine the predictions of multiple decision trees to create a more accurate final prediction. Indeed, the final prediction is determined by majority voting, each decision tree classifier gets a "vote" and the most commonly voted value for each row "wins". [Dat20]

The main strengths of a random forest are that it generally leads to very accurate predictions and is resistant to overfitting. The main weaknesses of using a random forest are that they are time-consuming[4] and are difficult to interpret. Since they average the results of many trees, it can be hard to figure out why a random forest is making predictions the way it is.

Variation in the random forest will ensure each decision tree is constructed slightly differently and will make different predictions as a result. Bootstrap aggregation and random forest subsets are two main ways to introduce variation in a random forest.

With bootstrap aggregation, one trains each tree on a random sample of the data. When doing this, one performs sampling with replacement, which means that each row may appear in the sample multiple times. With random forest subsets, however, only a constrained set of features that is selected randomly will be used to introduce variation into the trees.

[4]Making two trees takes twice as long as doing one, making three trees takes three times as long, and so on.

Consequently, when building a random forest classifier one can specify how many trees to build and how many predictors to use. While adding more trees usually improves accuracy, it also increases the overall time the model takes to train. In this report, Figure 17 shown below illustrates the process and shows that computing more than 500 trees does not improve the accuracy. Besides, using all the covariates in the exploratory analysis yields the best classifier.

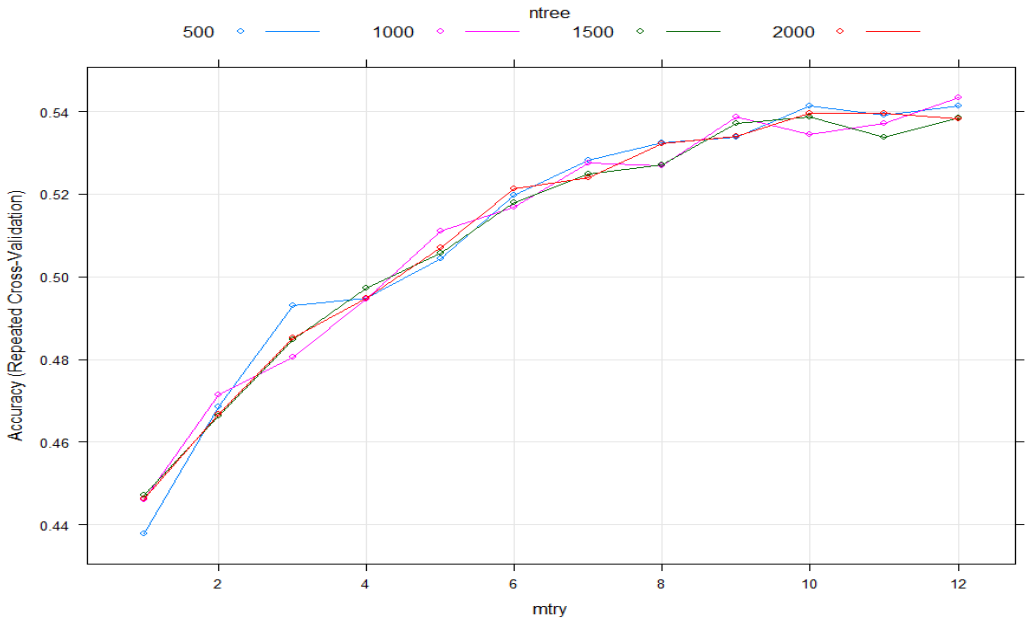


Figure 17: Random Forest fitting stage summary.[a]

9.4 Support Vector Machine models

The Support Vector Machine (SVM) classifier is a supervised learning algorithm used in classification which consists in finding the best hyperplanes separating the classes of the data in a high dimensional space.

Consider that the data has only two classes. Then, by *best* hyperplane, it is understood the hyperplane whose distance to the nearest element of each class is the largest. Therefore, only the closest points from different classes are considered when fitting a SVM classifier. This fact represents a great advantage of the SVM algorithm in memory efficiency since only a subset of the data points have to be stored. In many cases, the classes are not linearly separable and this is why it is necessary to map the data points to a higher dimensional space in which the classes are linearly separable. Doing this for every point might be computationally expensive and this issue is overcome using the well known kernel functions[5]. This "kernel trick" enlarges the feature space in order to accommodate a non-linear boundary between the classes. Common types of kernels used to separate non-linear data are polynomial kernels, radial basis kernels, and linear kernels. One of the main disadvantages of the SVM classifier is that they are very sensitive to the choice of the kernel parameters. [MBi06]

For more than two classes, let's say k, there are multiple ways of addressing the problem. In this report, the *one-vs-one* approach was used. That is, $\frac{k(k-1)}{2}$ binary classifiers are trained and the appropriate class is found by a majority voting rule as explained in the previous section.

[a] *mtry* denotes the number of variables randomly sampled as candidates at each split. *ntree* denotes the number of trees implemented to apply the random forest algorithm.

[5] No deeper information will be provided about how the kernel functions are used in SVM since this would need to explicitly talk about the mathematics of the method which is out of the scope of this report.

Bibliography

[Lit03] John B. Little. *Modeling and Data Analysis: An Introduction with Environmental Applications*. American Mathematical Society, 2003. ISBN: 9781470448691.

[HRB05] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. "ROBPCA: A New Approach to Robust Principal Component Analysis". In: *Technometrics* 47.1 (2005), pp. 64–79. eprint: https://doi.org/10.1198/004017004000000563. URL: https://doi.org/10.1198/004017004000000563.

[MBi06] Christopher M.Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 978-0387-31073-2.

[Del08] David L. Olson; Dursun Delen. *Advanced Data Mining Techniques*. Springer Science and Business Media, 2008. ISBN: 9783540769170.

[TF09] Valentin Todorov and Peter Filzmoser. "An Object Oriented Framework for Robust Multivariate Analysis". In: *Journal of Statistical Software* 32.3 (2009), pp. 1–47. eprint: https://www.jstatsoft.org/v032/i03. URL: https://www.jstatsoft.org/v032/i03.

[Dat20] Inc. Dataquest Labs. "Introduction to Random Forests". In: (2020).

[Con] Matthew Conlen. *Kernel Density Estimation*. URL: https://mathisonian.github.io/kde/. (accessed: 14.04.2020).

[CC] Centersfor Disease Control and prevention (CDC). *About Dengue: What You Need to Know*. URL: https://www.cdc.gov/dengue/about/index.html. (accessed: 12.04.2020).

[Nat] United Nations. *United Nations Data: A World of Information*. URL: http://data.un.org/Data.aspx?d=POP&f=tableCode:240. (accessed: 27.04.2020).

[RCo] RCommunity. *RDocumentation*. URL: https://www.rdocumentation.org/. (accessed: 24.04.2020).