ROCHESTER INSTITUTE OF TECHNOLOGY

MASTER'S THESIS

Predicting Cholera Positive Cases in Haiti

Author:
Jessica Young

Advisor:
Dr. Ernest Fokoué

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

in the

Schoool of Mathematical Sciences
College of Science

May 3, 2017

Committee Approval:

_____

Dr. Ernest Fokoue Date
Associate Professor
Thesis Advisor

_____

Dr. Steven Lalonde Date
Director, Graduate Applied Statistics
Committee Member

_____

Dr. Joseph Voelkel Date
Professor
Committee Member

Declaration of Authorship

I, Jessica Young, declare that this thesis titled, "Predicting Cholera Positive Cases in Haiti" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

ROCHESTER INSTITUTE OF TECHNOLOGY

Abstract

Dr. Ernest Fokoué
College of Science

Master of Science in Applied Statistics

Predicting Cholera Positive Cases in Haiti

by Jessica Young

While Western countries typically run census surveys frequently, poorer countries such as Haiti do not have the money to do so; thus research into how Haitians live is severely lacking. Furthermore, studies that do exist tend to be not only old and outdated, but also lacking in depth. Using new census data recently collected from Haiti, I attempt to predict if certain behaviors and living situations can be used as indicators for determining if someone has cholera. Challenges for exploring this data center on getting the surveys into a format suitable for analysis and the severe class imbalance between the number of cholera positive people and cholera negative people. Numerous solutions to this problem are attempted including using different sampling techniques, using ensembles with models like CART and SVM, and Bayesian model averaging. Better survey designs and questions to add to future surveys are also discussed.

Imbalanced Data; Imputation; CART; SVM; Novelty Detection; R; Survey Design

Acknowledgements

I would like to thank everyone that made the completion of this thesis possible: Dr. Fokoue, Dr. LaLonde, and Dr. Voelkel as well as Tyler who graciously helped me talk through numerous problems.

Contents

List of Figures

List of Tables

List of Abbreviations

CART	**C**lassification **A**nd **R**egression **T**rees
SVM	**S**upport **V**ector **M**achines
BMA	**B**ayesian **M**odel **A**averaging
SMOTE	**S**ynthetic **M**inority **O**versampling **T**echnique
KNN	**K** **N**earest **N**eighbors
ADASYN	**A**daptive **S**ynthetic
AUC	**A**rea **U**nder the **C**urve
G-mean	**G**eometric **M**ean

Chapter 1

Introduction

Poorer countries tend to be the same countries that suffer from large scale health problems. Understanding how and why these diseases sprout up and spread so quickly in these countries is of utmost importance in saving lives there. One of the ways to do this is to collect socio-economic data on the citizens and attempt to create a model to accurately predict who will get sick. While this thesis will be using newly collected census data from Haiti, no attempt will be made to extrapolate any predictive findings to the state of Haiti nor will the data be used to answer any socio-economic questions. This data is going to be used in the thesis because it is new and has not yet been analyzed, thus there is not existing knowledge on how to handle the data. This thesis only attempts to analyze the data in the capacity of being able to find models that can accurately and precisely reflect the data provided.

There are significant problems with the data that arose over the course of this thesis. The first and most immediate problem was that the survey from which the data came from was poorly created and saved. This meant that a large portion of time had to be dedicated to cleaning the data, properly extracting, and trying to make sense of it all. Nearly all of the predictors recorded in the data were also categorical which required extra care when modeling was done. However the most significant problem overall was that the response variable was severely imbalanced. The response takes on two values that indicate whether or not the person has cholera; 94% are cholera negative and only 6% are cholera positive. This imbalance presents a large challenge since traditional techniques fail to accurately predict the minority class (Grzyma\la-Busse, Stefanowski, and Wilk 2005). Thus the goals of this thesis are to use relational and socio-economic

data from Haiti to try to predict Cholera positive cases with traditional and new methods as well as explore approaches to dealing with imbalance in the data.

Preprocessing will be used since the survey data is not fully filled out. Imputation will be used to fill missing values in when necessary. Some questions are numerical (ie age and number of children) which will initially be left out to make the first pass at the analysis focused on only binary variables. Some of the questions are unstructured questions (fill in the blank) which will be left out. This is due to the nature of the answers which vary greatly and need to be put into categories. For example, some answers could say 'back' and others 'back pain' which should be the recognized as the same, but without text mining to create categories for the data, could not be viewed as the same in a prediction model.

Methods that will be attempted will focus on trying to properly predict cholera positive cases more so than predicting cholera negative cases since only about 6% of the data is cholera positive. These methods can be broken down into two categories: sampling techniques and model modifications. Sampling techniques include over-sampling, under-sampling, synthetic minority over-sampling technique (SMOTE), borderline-SMOTE, and adaptive synthetic sampling (ADASYN). Together with these five sampling techniques, models will be created with a baseline version of the data set to show the improvement these techniques make as well as an augmented version of the data set.

Since survey sections mostly include categorical or binary data, methods like trees and SVM are obvious techniques to with which to start. SVM is tested with two different kernels: the Gaussian kernel and the Laplace kernel which is thought to be better for binary data. A novelty detection method is also used (one-class SVM) with two different kernels: the Gaussian kernel and the Laplace kernel. Bagging, boosting, and random subspace learning is also implemented with a mixture of the aforementioned methods to try to increase the predictive rates. Bayesian model averaging (BMA) on trees, aka Bayesian Aggregate Regression Trees (BART), is compared to non-Bayesian model averaged trees (random forest). The results of this comparison suggested random forest had potential to be even better and thus weighted random forest, balanced random forest, and

weighted & balanced random forest are also run.

Overall there are 246 different models created (five basic models: trees, SVM, novelty detection, model averaging, and random forest; four ensemble augmentations: bagging, boosting, random subspace learning, and none; and six sampling techniques: over-sampling, under-sampling, SMOTE, borderline-SMOTE, ADASYN, and none). The results from these models suggest that the classification problem for this data set is very difficult. This could suggest that the data is poor and little to no signal can be determined from it or that the problem has a naturally high Bayes' risk.

This thesis intends to explore a novel data set from Haiti and methods that can be employed to properly predict the minority class when the data is severely imbalanced.

Chapter 2

Literature Review

2.1 Dealing With Imbalanced Data

In general, imbalanced data can be described a few ways: intrinsic imbalance, extrinsic imbalance, relative imbalance, or absolute rarity. Intrinsically imbalanced data is data that has class imbalance due to the nature of the data. On the other hand, extrinsically imbalanced data is data that is not actually imbalanced but appears to be so because where/when the data was gathered makes the data imbalanced. Relative imbalances mean that even if more data is added, the proportional imbalance remains the same. Similarly, absolute rarity occurs when the minority class is limited and/or rare to actually occur. In these cases, where the minority class is truly rare and hard to record, subconcepts of the minority class become much harder to capture which results in making classification that much harder. Subconcepts exist when a class can be broken down into smaller more distinct sections which are called subconcepts. If these subconcepts actually exist within the minority class, then another type of imbalance called within-class imbalance is present. This is commonly created by noise which makes determining these subconcepts difficult to do. Furthermore, things like data complexity and small sample sizes makes each type of imbalance that much harder to predict (He and Garcia 2009).

For heavily imbalanced data, there are two general methods for dealing with the imbalance. The cost sensitive learning approach deals with giving a high cost to misclassifying the minority class that you're trying to predict. In this way, the goal becomes minimizing the overall cost by minimizing the cost of misclassifying the minority class. The other approach is to use a sampling technique (Chen, Liaw, and Breiman 2001).

Sampling methods require that you modify the data in order to created a balanced distribution before modeling. This is typically done when modeling with the imbalance is difficult, which may not always be true. The most common methods are to either down-sample the majority class (use fewer observations from the majority class to make the size of the class smaller) or to over-sample the minority class (reuse the same observations from the minority class to make the class size larger). By increasing the class size of the minority class, you increase their weight but do not increase information about the class. Thus while random oversampling appends more data to the original dataset and can lead to overfitting, random under sampling removes data and thus removes information. A third sampling technique called SMOTE, synthetic minority over-sampling technique, combines over-sampling and down-sampling but instead of bootstrapping, creates synthetic examples of the minority class. In general, prior research has shown that for trees, coercing the data to have equal class priors is effective and that over-sampling does worse than under-sampling. In both cases, the testing error tends to be far worse than the training error (He and Garcia 2009) (Chen, Liaw, and Breiman 2001). Empirically, both under-sampling and over-sampling have shown to improve the accuracy of the minority class (Hernandez, Carrasco-Ochoa, and Martínez-Trinidad 2013).

Under-sampling can be modified to help prediction power and to avoid losing information. One example of this is called EasyEnsemble which is a supervised learning approach to under-sampling. In this case, multiple subsets of the majority class are taken and randomly combined with the minority class and fit with classifiers. BalanceCascade is another supervised learning algorithm but created an ensemble of classifiers to determine which majority class examples to undersample. For our purposes, simple under sampling will be used for a better comparison against simple over sampling. There is also a multitude of KNN under sampling methods (NearMiss-1, NearMiss-2, NearMiss-3, and the most distant method) which work fairly well but are computationally expensive and therefore will not be considered (He and Garcia 2009).

Synthetic sampling is also commonly used to combat imbalance in data. The most common is called SMOTE which creates artificial

data from the existing data based on the similarities between minority class observations. Although SMOTE is powerful, it tends to over generalize and create large variance. Part of this problem is due to SMOTE creating new synthetic points without consideration to overlapping between the classes. Thus the lines that separate one class from another become blurred and hard to pinpoint (He and Garcia 2009). SMOTE has some very nice theoretical properties when dealing with high dimensional data. While SMOTE adds data to the minority class, the overall expected value of the minority class is not changed and the variability is reduced. Furthermore, there is no correlation introduced between variables, however some is introduced between observations. Since SMOTE modifies the Euclidean distance between test samples and the minority class, the test samples tend to be more similar to the SMOTE samples. Lusa concludes that SMOTE reduces bias towards the majority class in KNN, SVM, CART, and random forest (Lusa and Blagus 2013).

Borderline-SMOTE works like SMOTE, except that it determines 'danger' points which are points on the borders that can be easily misclassified. The algorithm takes these points and creates synthetic observations for the minority class to clean up the border. Empirically, borderline-SMOTE tends to produce high true positive rates and F-values (Han, Wang, and Mao 2005). On the other hand, ADASYN (adaptive synthetic) uses a density distribution and changes the weight of different minority examples to make up for the imbalance (He and Garcia 2009). Empirically, ADASYN improves the accuracy of both the minority and the majority classes without a preference for either one (He et al. 2008).

Data cleaning techniques can also be used help clean up the imbalance. One method called Tomek links helps to get rid of any overlapping between the majority and minority classes. It find the smallest distance between two neighbors of opposite classes and if they are closest to each other and no other point (from either class) is closer to either one then the pair of points is considered to be either noise or near a border. Either way all Tomek links are removed until all nearest neighbors are of the same class. For the sake of this research, Tomek links will not be considered because removing so much data can lead to a serious loss of information (He and Garcia 2009).

While resampling at the data level is popular, this often comes with numerous problems. These techniques try to make each class have equal prior probabilities even though the optimal distribution is not actually known. Furthermore, if the resampling is ineffective or not done properly, then information about the majority class can be lost and the minority class being resampled so often could produce overfitting. There is an extra learning cost associated with processing the data for resampling that is usually unavoidable (Sun et al. 2007).

Regular classifiers tend to do poorly on imbalanced data because they focus on minimizing the overall error which neglects the minority class that one is trying to properly predict. Thus when these methods are used, the accuracy may be high, but the minority class tends to be misclassified more than the majority class. When considering disease diagnostic data in particular, the preferred classification gives more weight to properly predicting the disease than not (Sun et al. 2007).

Other methods for dealing with imbalanced data are numerous. Cost-sensitive decision trees that are not pruned can be used to easily see what would determine a class. On the other hand, cost-sensitive neural networks have also been used when the problem is too complex for other methods. However neural networks are very complex and computationally expensive, thus they will not be considered. Kernel-based methods are also numerous for dealing with imbalanced data such as kernel SVM. Additional methods include one-class SVM and novelty detection which both focus on learning the minority class instead of trying to learn both classes at once (He and Garcia 2009).

In Chen, Liaw, and Breiman 2001, sampling techniques are compared with balanced random forest and weighted random forest. For balanced random forest, down-sampling is combined with the CART algorithm to create a tree based on a bootstrapped sample of the minority class and the same size sample of the majority class. Each tree creates splits that search through a pre-specified number of variables. After all the trees are created, the predictions are then aggregated together. By using the down-sampling technique, the majority class losses some information. However this is rectified by creating a multitude of trees from different samples of the majority class. On the other hand, weighted random forest does not up

nor down-sample the data at any point. Instead the minority class is given a much higher weight than the majority class at the beginning. The weights are used when finding splits for the tree; the Gini criterion is weighted by the specified class weights. Further more, each terminal node gives a prediction based on the specified class weights. When aggregating the trees together, the final class prediction is given based combining the weighted votes of each tree per class. For the data sets in the paper, weighted random forest and balanced random forest seem to perform better than the other sampling techniques and methods tested and the two seem to do equally well (Chen, Liaw, and Breiman 2001).

In Wang and Japkowicz 2010, Wang and Japkowicz chose to explore using SVM on imbalanced data sets. They decide to not use AUC or ROC as measurements of goodness for their models since they are explicitly interested in determining if the minor class is properly predicted. Instead they use the geometric mean, true negative rate, and true positive rate as measures of fit. As they state, previous papers using SVM on imbalanced data overfit the data. The authors attempt to mitigate the overfitting by modifying the distribution used in SVM to make a more balanced distribution. To do this they considered different classifier combinations methods: bagging, boosting, and stacking. Bagging requires weak unstable learners to work properly, which meant that small changes in the training set would largely affect the classifier produced which is not reliable. On the other hand, stacking does not usually combine models of the same type and thus had to be ignored. This left boosting which could combine models of the same type and did not require unstable weak learners.

Other methods of modifying SVM to deal with imbalanced data included using kernel transformations and biased penalties. The kernel transformation method requires using the kernel on the spatial distribution instead of the input space with an RBF distance. While the kernel transformation method is highly effective, it is also very complex and hard to use correctly. On the other hand the biased penalty method adjusts the cost factor of false positives and false negatives directly in SVM. While this method does well at controlling the true positive rate, it cannot control the true negative rate. Instead

a combination of methods was created to boost SVM while also using an asymmetric misclassification cost in an attempt to control both the true negative and true positive rates. This algorithm is similar to the regular SVM algorithm with the exception that each SVM created is boosted according to a predetermined weight that varies by class. Each SVM model is only used in the final aggregation if it's geometric mean is better than 0. Overall, the boosted SVM classifier outperforms numerous other methods including boosted SMOTE, weighted random forest, and other versions of SVM by a landslide. In fact, this method is always better than SVMs with an asymmetric cost and L1 norm considered (Wang and Japkowicz 2010).

Using boosting for imbalanced data comes with quite a few benefits. Not only can it be used with most classification methods, but it automatically eliminates the extra learning cost that other methods induce like needing to learn what the border points are between two classes and synthetically creating more when using SMOTE. Boosting can also be implemented with sampling techniques to further combat the imbalance. SMOTEBoost creates artificial samples after each boosting iteration so that the minority class is focused on even more. Whereas DataBoost-IM also creates artificial samples, it does so based on the observations that are difficult to classify. However both of these methods are computationally expensive and complex so they are not often used. A less computationally expensive algorithm that still uses boosting is JOUS-Boost, which jitters the oversampled data so that the replicates created are not exactly the same. JOUS-Boost was shown to provide efficient results with its smaller runtime. While up-sampling and down-sampling can lead to overfitting and loss of information respectively, AdaBoost does not have these problems. Furthermore, AdaBoost can also potentially reduce the bias of some classification methods. However, without modifying AdaBoost improved prediction performance of the minority class is not guaranteed since AdaBoost is accuracy-oriented. A modified version of Adaboost that is auc-oriented is used in this thesis instead of JOUS-Boost and SMOTEBoost (He and Garcia 2009) (Sun et al. 2007).

One can introduce cost items into the AdaBoost algorithm in order to modify it to predict the minority class better. This paper creates 3 different ways of updating the algorithm which are referred to

as AdaC1, AdaC2, and AdaC3. Another variation of the AdaBoost method is AdaCost which is also a variation of AdaC1 but requires a more ad hoc way of selecting the cost adjustments. RareBoost is another example of a variation of Boosting being modified by a cost item. However RareBoost is not used when focusing on the minority class because it requires that the true positive rate is larger than the false positive rate which is the problem the minority class faces in imbalanced data. Besides AdaC1, AdaC2, AdaC3, and AdaCost, CSB2 is another boosting method that can be used that follows a similar cost scheme. All 5 variations of AdaBoost increase the weights of false negatives more than false positives in order to favor the minority class, however each scheme differently weights true positive and true negative predictions. With the exception of AdaC1, all of these methods can achieve a higher recall value than precision value. Both AdaC2 and AdaC3 are sensitive to their cost setups where making the cost of the minority class smaller makes the recall dramatically worse. On the other hand, AdaC1 and AdaCost are insensitive to the cost setups, but AdaCost has higher values of recall in most cases. Overall though, AdaC2 produces better results than the other methods considered. These versions of Adaboost are ignored for this thesis though since the focus is on comparing different methods instead of numerous similar methods (Sun et al. 2007).

Besides not being able to use most of the same classification techniques as non-imbalanced data, imbalanced data also needs to be evaluated differently. Some regular model evaluation methods like accuracy and overall error favor the majority class since they are considered with getting the highest overall prediction rate which tends to rely on the majority class. One can instead use the F-measure to determine if the classification method used is properly predicting the minority class which takes into account recall and precision equally. Alternatively, one can use the geometric mean (G-mean) which gives the average between the true positive rate and the true negative rate in cases when predicting both classes properly in a concern. On the other hand, the ROC curve can also be used to show how well the classifier does. In this case, a pictorial representation gives a better idea of how the classification method does compared to random guessing and one can use it to easily compare multiple methods. The ROC curve does not allow for a single overall winner unless the

curve is clearly dominating over the other curves at every aspect of the graph (Sun et al. 2007).

2.2 Model Averaging

Model averaging is a method designed to use multiple models when predicting. It works using a set of weights such that every model is assigned a specific weight where the sum of all weights considered must be 1. When predicting, average across all weighted model predictions to determine a single prediction.

$$\sum_{i=1}^{m} w_m = 1$$
$$\sum_{i=1}^{m} w_m f_m(x)$$

Model weights can be determined in numerous ways, but Bayesian Model Averaging(BMA) is the most common (Hansen 2007). BMA is used when there are numerous models that are reasonable, but there is no one clear winner among them. Similar to how lasso drives some variables to 0 so that the more relevant variables have a larger impact on the prediction, BMA drives some variable's weights closer to 0 through model weights(Amini and Parmeter 2011).

Using BMA, one can avoid the unnecessary risk of selecting a single model by using multiple models. In general, a posterior distribution for data is generically give as:

$$\begin{aligned}
\text{posterior} &= \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \\
pr(M_k|D) &= \frac{pr(M_k)pr(D|M_k)}{pr(D)} \\
&= \frac{pr(M_k)pr(D|M_k)}{\sum_{l=1}^{K} pr(D|M_l)pr(M_l)} \\
&= \frac{pr(M_k)\int pr(D|\theta_k, M_k)pr(\theta_k|M_k)d\theta_k}{\sum_{l=1}^{K} pr(D|M_l)pr(M_l)}
\end{aligned}$$

Averaging over all of the models available will on average give a better prediction than using any single model available. While Bayesian

model averaging seems useful because of this, there are a few setbacks for it that make it difficult to implement in most cases. Determining which class of models to average over, integrals implicit in the posterior, determining how many models to average over, and specifying the prior distribution of models are all ways in which BMA is difficult (Clyde 2003) (Hoeting et al. 1999).

Determining which class of models over which to average over is up to the researcher who can average all models available for use or a subset using Occam's window. Using Occam's window (and Occam's razor) to determine which models to use works fairly well in making sure there are not too many models to average over. The integrals in the posterior can be simplified using methods like Markov chain Monte Carlo method, MLE approximation, or even simpler the BIC approximation. Using MLE approximation would mean that the posterior no longer relies on the prior. Using the BIC tends to work well in most cases and is an easy way to approximate the posterior of each model. The prior distribution of models can be determined using a uniform distribution on the model space that gets updated with imaginary data from the domain expert (Clyde 2003) (Hoeting et al. 1999)(Wasserman 2000).

The weights in BMA can be determined with the BIC as:

$$w_m = \frac{\exp(-.5BIC_m)}{\sum_{j=1}^{M} \exp(-.5BIC_j)}$$

where BIC_m is the BIC of model m. It is possible to use the AIC instead of the BIC if getting an estimator with low loss is important(Hansen 2007).

2.3 Survey Design

When designing a survey, determining the method for collecting information is one of the crucial initial steps to undergo. This should be decided based on the type of information being sought as well as the people being surveyed. In general, some survey formats are mail questionnaires (which are cheap and easy but have a low response rate) or interviewing (which takes a long time to do and depends on the skill of the interviewer). Developing countries tend to use interview surveys because a majority of the population could

be illiterate and numerical questions may confuse the interviewees. For interviews, one could write up a standard set of questions to be asked at each house, or write out a list of information needed at each house but leave the exact wording of the question to the interviewer. On the one hand, using all of the same exact questions is reliable in that everyone will know the same information when asked, but may not produce valid results if anyone is confused on what the question means. On the other hand, if you allow the interviewer to phrase the questions as they like then the interviewer could phrase the questions differently so that each person fully understands what is being asked, however by doing this responses are no longer comparable. More often than not, a combination of the two is used (Raj 1971).

A clean and nice looking form should always be used when it is given to the respondent. Questions should be simple and clear and be kept to a minimum. When the form is not given to the respondent then attractiveness is no longer a concern, instead the form should be easy to fill out and obvious if/when different parts need to be filled out by different people. In general, forms can usually be divided into three parts: respondent identification questions, classification questions, and the survey topics to be answered. The order questions are asked is also very important and should be logical so as to not confuse the respondent. Questions should begin interesting so that the respondent is more likely to continue answering and they should make the respondent feel comfortable so that they are more inclined to answer more intimate questions (Raj 1971).

When creating the question content, it is important to make sure that respondents could accurately answer them without bias or guessing on their part and to make sure that they are willing to answer them. It is also imperative that questions are not biased in one direction. Furthermore, it is important to make sure the wording of each question is clear and that respondents will know each word of the question. Good question wording will use the simplest words to show the exact what you want to ask. Two different questions types to consider for surveys are closed questions (ie multiple choice) or open questions (ie short/long answer). In the case of closed questions, analysis is easy since everything is coded already, but people may be forced into a category that they do not belong. more often than not, when given a range of options, people tend to choose the

middle road option even if it is not true. Open questions allow the respondent to answer exactly how they want to, but coding the answers becomes tricky (Raj 1971).

Sensitive questions should generally be placed last in a survey since respondents make react negatively and stop answering the survey. This way, if they do quit the survey, there is still enough information gathered that it could be considered completed. Furthermore, a rapport could be built between the interviewer and the respondent by this point in the survey so the respondent could be more comfortable and answer the question. Unlike sensitive questions, related questions should not be placed last unless it makes sense. These questions should always follow the question they relate to and should make clear that they are related to the previous question so that respondents do not get confused. It is important that these questions follow a logical sequence to 'evoke reflexive responses' (Rea and Parker 1992).

Some questions could be considered screening questions meaning that they are placed in the survey to determine if the respondent is qualified to answer the following questions. Thus if they are not, then one can clearly mark the question telling them to skip whatever follow up questions are stated. When a question is truly important or sensitive, a reliability check may be needed in the survey to determine consistency. In these cases, a second question is asked later in the survey that aims to measure the same response as an initial question. Both questions ask the same thing but in different ways (ie two different multiple choice questions where the answers could be a range of likeliness for one question and a range of emotions for another). The reliability check would be if the respondent answered each question in a specific way then you can assume they pass the check. In general, there should not be many open ended questions and they should be placed late in the survey. Sometimes a survey can end in an open-ended venting question to ascertain how the respondent feels overall about specific subject matters or the survey in general. Surveys should try to stay short and include questions that only gleam the necessary information the researcher needs. Extraneous questions can distract from the primary focus of the survey goals. More importantly, as more questions are added and as questions become more complex, the survey can be seen as even longer

and prevent many from completing it (Rea and Parker 1992).

Survey nonresponse can often occur when collecting data for a variety of reasons. The problem with survey nonresponse is that it could be entirely random or there could be a systematic reason for the nonresponse. While it is entirely possible that nonresponse could be random, it is rare for it to be so. Nonrespondents tend to differ systematically from respondents thus imputing on these instances is highly desirable for pattern recognition (Rubin 1987).

Single imputation is the most common method for dealing with missing values. This would fill in a single value for each missing value. By using single imputation, complete-data methods can then be used to analyze the data, such as SVM. Furthermore, if using single imputation then the data collector could fill in the missing values themselves if they knew what the values should be (ex the data collector could be a member of a small town and know the ages of townsfolk that had not filled out their age). Single imputation however comes with the downside of assuming that the single answer you fill in correctly represents that observation when in reality, it could be entirely off (Rubin 1987).

Multiple imputation comes with all the advantages of single imputation but also has the advantage of reflecting sampling variability. Instead of replacing the missing values with only one value, m multiple values are imputed where m is usually between 2 and 10. Multiple imputation is designed for when there is a modest amount missing values, but not an overall large amount. The disadvantage to multiple imputation is that it requires more storage space, more work in imputing, and is also more difficult to analyze. This extra work and storage can be modest if the amount missing is also modest (around 20%) and reaps benefits such as sensitivity to models and valid inferences for models (Rubin 1987).

Chapter 3

Data Description

3.1 Survey Design

The Haiti census survey was created in Haitian Creole so that natives could fully understand the questions. The survey was created and collected in Microsoft Word and implemented in different villages on different days throughout 2015. In general, sections are broken down into the developments, and then areas. Surveys in the same development were completed within a week of one another whereas sections were completed weeks apart. After each development completed the survey, they were all saved in the same Microsoft Word file with the name of the section and the date it was completed as the title; for developments that needed more than one day to complete the survey, the surveys are stored in different files based on the day they were completed . Surveys in different areas and sections, however, were not all saved together. In total the survey consists of six distinct sections. To see a blank survey that excludes the name questions, see A. The survey responses analyzed in this thesis come from the following sections and developments:

TABLE 3.1: Survey Sections and Developments

Section	Development	Number of Files
Boukan Michel	Basen Kayiman	1
Boukan Michel	Chapel	3
Chanpay	Marikongo	2
Chanpay	Savi/Sent Mari	2
Chanpay	Vedrin	1
Mago	Bado	2
Mago	Dolan	1
Mago	Fon Milo	2
Mago	Ibo	2

The first section consists of fourteen unstructured fill in the blank style questions that help determine what the respondent's household looks like for family members and where it is located. The date and survey code are first recorded, followed by the section, area, and development they live in. The number of yards that their house has is also recorded. This section also asks how many homes they own, and how many significant others, boys/men, girls/ladies, total people, adults/elders, and children live with the respondent. The section ends by asking for any additional information the respondent wants to share at this point.

After the first section concludes, the respondent then gives information about family members (including themselves) in the next three table sections. The first table section asks for information on any family member that is currently living with the respondent. The basic name, age, and sex questions are asked first followed by asking what the highest grade level completed by the person was and what their current job is. From there, ten health and behavioral questions are asked that require yes or no answers. These questions ask if the person has had/currently has cholera, if they drink coffee, if they drink hard water, if they drink alcohol, if they smoke, if they chew tobacco, if they've had an operation, if they have dizziness, if they have headaches, and if they have hypertension. A follow-up question asking the respondent to detail any other diseases their family members have is asked and then the section ends with by asking if each family member goes to the hospital.

The second and third table sections ask the same questions as the first, but each asks less questions than the section before it. The second table section asks about family members that live elsewhere in Haiti whereas the the third table section asks about family members that are currently living abroad. The second table section asks the age, sex, and highest grade level completed of the family member as well as where they are currently living and what their current job is. The health/behavioral questions asked include: if the person has had/currently has cholera, if they drink coffee, if they drink alcohol, if they smoke, if they chew tobacco, if they've had an operation, if they have headaches, and if they have hypertension, other diseases they have, and if they go to a hospital. The third table section asks for the name, age, sex, highest grade level completed, how long they've

been out of the country, and where they are currently. The only health/behavioral questions this section asks about are whether or not the person has cholera, if they drink coffee, if they drink alcohol, if they smoke, what other diseases they have, and if they go to a hospital.

The fourth table section is the shortest and asks the respondent about recently deceased (within the past 5 years) family members. The only information recorded for these family members is age, sex, where they lived when alive, previous illnesses, what they died of, and where they died.

The final section of the survey consists of eighteen questions that are a mix of fill in the blanks and categorical questions. The aim of this sections was to determine the living situation of the respondent. Questions start by asking about the type of house, type of roof, and type of flooring in the house and then move into sanitary measures. These measures include asking if water is treated (yes or no), what type of treatment is used (categorical), if there is a bathroom at the house (yes or no), and what how they go to the bathroom (categorical question about if they dig a hole, have a toilet, etc.). After the sanitary questions, the survey moves on to asking general household questions: if the house cuts trees for firewood (yes or no), where they breed animals (fill in the blank), if there's a garden (yes or no), what the land is used for (categorical), if they have specific animals (a list of fourteen common farm animals and common pets is given), if there are any pregnant women in the household (yes or no), their ages (space is alloted for up to six answers), whether these women will go to a hospital (yes or no), and how sicknesses/illnesses are handled in the house (categorical). The survey ends by asking the respondent to list three development issues they'd like worked on, if they have any advice for A.S.B (the creators of the survey), and if there's anything else they have to add to the investigation.

3.2 Data Transcription

Extracting the data from the survey proved fairly difficult since the data was saved in numerous tables and paragraphs in multiple Microsoft Word documents. Initial attempts to extract the data included converting the Word documents to plain text files and attempting to

read in different sections separately with preexisting Word to CSV functions in R. The method that worked the best was eventually found to be the following: first copy the entire file over without any images (there's a header image that was not needed for the extraction) into another Word file, next save the file as HTML, and lastly using the XML library in R to read and parse the HTML file with the htmlTreeParse function. From there, the surveys can be separated by section easily by using regular expressions (regex) methods. For the first and last section, all answers can also be easily found by using regex to search for the question.

For the four table sections, extracting the information was not as straightforward. As with the first and last section, the sections were found using regex to pull the information between section headers. After pulling the information, it was necessary to determine how many rows in the table were actually filled in. To do this, it was determined that if there was no capital letter followed by at least one lower case letter (accented letters were included for both searches since everything was recorded in Creole), then the row must be blank. This was determined optimal because if the name columns had been filled in then there should be at least one lower case letter, whereas if they were left blank then the row would only read 'MF' and 'WN' repeatedly. Any rows that were not filled in were automatically ignored from the rest of the process. From there each row became a string in a list. Each string was constantly trimmed and regex was used to help extract the responses to each question. The trimming was necessary because there was a collection of yes (W) no (N) questions that made it difficult to pull the answers from normally if any of them had been skipped or filled in incorrectly.

There were numerous problems when trying to extract the information for analysis that stemmed from the format of the survey. Since Microsoft Word had been used, the tables that were created gave unusual spacing between columns. Sometimes answers in columns next to each other would appear right next to each other and other times there would be as many as twenty-five spaces between the answers. This was caused from the way the answers were recorded. For the yes/no and gender questions, instead of filling in one answer per question, these columns were further broken down into two

more columns each contained one possible answer pre-filled in. Respondents needed to thus erase/delete the answer that they did not select. This meant that if the respondent skipped the question then both answers would remain filled in which made determining which question was not filled in properly difficult. Eventually it was discovered that if a question was not properly filled in, then there would be eight or less spaces between the two possible answers whereas if two columns were properly filled in then their answers would have anywhere from eleven to twenty-five spaces between them. Furthermore, the third table section included an excess two columns in the middle of the table which just created even more spaces between answers.

Any question that allowed respondents to write in their own answers also had numerous problems. These answers are particularly difficult to analyze because there's a multitude of ways each person wrote their responses. For example, some people would write 'Doule' others wrote 'doule' and others wrote 'douel' when asked what other diseases family members had. This is made even more difficult since sometimes people would add on to their answers further. While some people would give one worded answers, others would give phrases, but most just did not answer these questions. While these answers could be coerced into categories, it requires an intimate knowledge of both written Creole and Creole shorthand to do so. Shorthand can be seen in the school level and age questions. When answering the ages of family members, ages under one years old were given in shorthand to reflect either months (written with an m next to a number), days (written with a j next to a number), or even hours (written with an e next to a number). The shorthand for days and months was similar to French and could be easily verified but uncovering that 'e' meant hours was not nearly so easy to confirm since hours is not a common way to report age. When asked for the highest level of education, every answer was given in shorthand. While some could be deciphered as being in primary school or high school (answers ending in 'eAf') others were impossible to determine even after extensive research into Haiti's school system (like answers ending in 'es' or just 'e').

Other problems with the data revolved around question clarity. For example, the first three table sections all asked if the person had

any other diseases yet most people left this question blank. It is unknown whether the blanks mean that they had no other diseases or if they did not want to answer the question. This is further confounded by the fact that sometimes people would write 'Anyen' meaning 'nothing', thus it is unclear whether the survey creators intended for people to write down 'Anyen' if there were no other diseases to report or not. Another clarity issue comes from not clearly marking follow up questions in the survey. The last section asks whether or not there are any pregnant women in the household and then has two follow up questions asking for their ages and whether they would go to a hospital for the delivery. These were not clearly marked as follow-up questions though as evidenced since most people that said there were no pregnant women in their household still answered if they would go to a hospital.

Further confusion arose on the side of data analysis when trying to determine what each question was asking. Each table section had only key words written as column headers for the questions which required translating and some detective work to clarify. There was one question that no amount of translating or detective work could solve though: 'Li gen Larouli'. It translates directly to 'It(he/she) has Larouli' which would imply that Larouli is some type of disease, but in reality it does not exist at all. What the survey creators probably meant was 'lawouli' which means dizziness. Since Creole is originally a spoken language and not a written one, 'larouli' could very well be how most people pronounce the word, but without speaking to the survey creators it is unclear if the mistake is a spelling error, if the creators wrote down how one would pronounce the word to make it more obvious, or if "Lawouli" means something else entirely different.

Chapter 4

Methodology

Data analysis will be broken into two parts: baseline and augmented. The baseline data refers to all the predictors in 4.1 and the cholera response whereas the augmented data refers to all the predictors in 4.1 and 4.2 with cholera as the response. There are a few reasons to use two different versions of the data. For the sake of the analysis, starting with a data set that is made up of variables of the same type make analysis more straightforward since they can all be treated the same way. Furthermore, by comparing a baseline data set that consists of questions from only the first table to an augmented data set that consists of questions from each section, a comparison can be made that determines if more variables is necessarily better.

While logistic regression is possible for this problem type, it is not included in the scope of this thesis because there were sampling errors that occurred when using sampling techniques and ensembles. Resampling sometimes yielded training sets where only one level of a multi level predictor appeared which made the logistic regression model inestimable. Although there was a lot of time dedicated to fixing this problem, with the limited time available logistic regression was not considered. There had however been results from logistic regression with no sampling technique done that mirrored results given by CART.

* While the survey claims this predictor is 'Larouli', further research into the translation provides that the word could be misspelled and instead might be 'Lawouli' which means dizziness.

Before trying to classify the data, five sampling techniques will be used: over-sampling, under-sampling, SMOTE, borderline SMOTE, and ADASYN. Another pre-processing technique that is employed as well is imputation which cleans up the missingness in the data so

TABLE 4.1: Predictors in Baseline Data

Predictor	Variable Type	Defined
Coffee	Binary	Do they drink coffee?
HardWater	Binary	Do they use hard water?
Alcohol	Binary	Do they drink alcohol?
Smoke	Binary	Do they smoke?
ChewTobacco	Binary	Do they chew tobacco?
Operation	Binary	Have they had an operation?
Larouli*	Binary	Do they get nauseous?
Migraines	Binary	Do they get migraines?
HyperTension	Binary	Do they have hyper tension?
Hospital	Binary	Have they been hospitalized?

TABLE 4.2: Additional Predictors in Augmented Data

Predictor	Variable Type	Defined
Age	Numeric	Age of person
Gender	Binary	Gender of person
Schooling	Categorical	Highest level of schooling attained
HaveFamilyElsewhere InHaitiWithCholera	Binary	Do they have family members elsewhere in Haiti with cholera?
HaveFamilyAbroad WithCholera	Binary	Do they have family members abroad with cholera?
HaveFamilyThat DiedofCholera	Binary	Do they have family members that died of cholera?
No.HomesOwned	Numeric	How many homes do they own?
No.PeopleInHouse	Numeric	How many people live in their house?
TreatWater	Binary	Do they treat/clean their water?
WaterTreatmentUsed	Categorical	What sort of treatment do they use?
Bathroom	Binary	Do they have a bathroom in their house?
BathroomMaterials	Categorical	What type of bathroom is it?
LandUse	Categorical	What is their land used for?
Firewood	Binary	Do they use firewood?
Garden	Binary	Do they have a garden?

that SVM can be used. To keep things simple, imputation on any binary predictors just uses the median. Cross validation is used to help train and optimize models. All models were tuned with 50 repetitions of cross validation. The following models are then created with training sets of the data: trees, SVM, and novelty detection. Each of these methods will also be modified in different ways. Trees will be

run as trees, bagged trees, boosted trees, and random subspace learning trees. SVM and novelty detection is run as SVM/novelty detection, bagged SVM/novelty detection, and random subspace learning SVM/novelty detection. SVM and novelty detection will also explore two different kernels: the Gaussian Radial Basis Function kernel and the Laplace kernel. Ensemble methods to be used include random forest and Bayesian additive regression trees. Random forest is also modified to show balanced random forest, weighted random forest, and balanced and weighted random forest. Each method will be compared using the F-measure, the G-mean, AUC, the sensitivity, and the specificity.

4.0.1 Cross-Validation

Cross validation is used when creating a model to ensure that a model is created that is accurate without over-fitting. To begin with, a data set will be defined as:

$$\begin{aligned}
\mathcal{D} &= \{(x_i, y_i) \overset{\text{iid}}{\sim} P_{XY}(x, y), i = 1, \ldots, n\} \\
&\equiv \text{Given data set} \\
&\equiv \mathcal{L} \cup \mathcal{V} \cup \mathcal{T} \\
&\equiv \mathcal{T}^c \cup \mathcal{T} \\
\mathcal{L} &\equiv \text{Training set} \\
&\quad \text{used for learning (fitting) a model} \\
\mathcal{V} &\equiv \text{Validation set} \\
&\quad \text{used for tuning hyperparameters withing training set} \\
\mathcal{T} &\equiv \text{Test set} \\
&\quad \text{used for proxy to prediction/generalization}
\end{aligned}$$

TABLE 4.3: Cross Validation visual

$\mathcal{T}^c$					$\mathcal{T}$
v_1	v_2	v_3	...	v_m	
v_1	v_2	v_3	...	v_m	
v_1	**v_2**	v_3	...	v_m	
			⋮		
v_1	v_2	v_3	...	**v_m**	

You can visualize cross validation as cutting up the data set into your training and test set by making them two separate sections of the data. To start cross validation, you split up the training set into m different chunks. decide on m, the number of chunks you want to split the training set into. This should split it up evenly so that each chunk has the same number of observations. Starting at the first chunk, v_1, and continuing to the last chunk,v_m, you would disregard one block at a time and train the model with the rest of the chunks in the training set. For each model created, you would then calculate the CV error of each model. The optimal model is then chosen based on the smallest cross validation error. Before going into details, here are some need-to-know calculations necessary for cross validation.

For $\ell_1, \ell_2, \ldots, \ell_{|\mathcal{L}|} \in \mathcal{L}$

Training Error

$$err(f; \mathcal{L}) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} \ell\big(Y_{\ell_i}, f(x_{\ell_i})\big)$$

Validation Error

$$err(f; \mathcal{V}) = \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \ell\big(Y_{v_i}, f(x_{v_i})\big)$$

Test Error

$$err(f; \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \ell\big(Y_{t_i}, f(x_{t_i})\big)$$

Step-By-Step Cross Validation
for (r in 1:m)

1. leave out rth chunk (v_r)
2. train the function f on $\mathcal{T}^\rfloor$ without v_r and form $\hat{f}(\cdot, \mathcal{T}^c \backslash v_r, \alpha) \equiv \hat{f}^{(-r)}(\cdot) \equiv$ estimator built without rth chunk
3. $err(\hat{f}^{(-r)}, v_r) = \frac{1}{|v_r|} \sum\limits_{(x_i, y_i) \in v_r} \ell(y_i, \hat{f}^{(-r)}(x_i))$
4. $cv(\alpha) = \frac{1}{m} \sum_{r=1}^{m} err(\hat{f}^{(-r)}, v_r, m)$

When you need to tune for hyperparamteters, you would simply put the above loop inside another loop that would run through a sequence for the hyperparameter. You would then choose the setting of the hyperparameter such that

$$\alpha^{opt} \equiv \underset{\alpha \in \text{set}}{\text{argmin}} \{cv(\alpha)\}$$

4.1 Sampling Techniques

4.1.1 Over-Sampling

Randomly over-sampling the minority class consists of just two steps:

1. Determine the difference in size between the majority class, S_{maj} and the minority class, S_{min}

$$d = |S_{maj}| - |S_{min}|$$

2. Bootstrap a sample of size d, with replacement, from the minority class. This sample is then added to the minority class to increase the size of the class to be equal to the size of the majority class.

4.1.2 Under-Sampling

Randomly under-sampling the majority class is just as straightforward as over-sampling.

1. Determine the difference in size between the majority class, S_{maj} and the minority class, S_{min}

$$d = |S_{maj}| - |S_{min}|$$

2. Randomly select d observations in the majority class and remove them. This way the size of the majority class is reduced to be equal to the size of the minority class.

4.1.3 SMOTE

Synthetic minority oversampling technique (SMOTE) increases the size of the minority class by creating synthetic samples. For non-continuous variables, this is done by taking the majority vote of the observation vector being considered and its k nearest neighbors. This majority vote is the value given to the new synthetic sample created for the minority class. For continuous variables the difference between the observation being considered and its k nearest neighbor is found and then randomly multiplied by a number between 0 and 1. This product is then added to the original observation's variable amount to produce the new synthetic amount.

$$x_{new} = x_i + (x_i^{KNN} - x_i) \times \delta, \delta \in [0, 1]$$

One of the major advantages to using SMOTE is that the synthetic minority observations created cover a region that the data does not normally reflect. By creating these new minority observations, classifiers are better able to predict when observations belong in the minority class (Chawla et al. 2003)(He and Garcia 2009).

KNN

In KNN, classes are assigned to points based on what their nearest neighbors classes are. KNN is flexible in that you can choose what distance measure to use and different distance measures can produce different predictive results. KNN can be modified with weights when dealing with class imbalance. In this way, observations from the minority class are weighted more heavily than those from the majority class in an attempt to more correctly predict when an observation is from the minority class. The general algorithm is:

- given data $\mathcal{D} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$ where $x_i \in \mathcal{X}^p, Y_i \in \{1, \ldots, c\}$
- First decide on a distance measure to use and k, the number of neighbors to consider
- For point x^*, compute the distance between it and each point
- Rank all the distances in increasing order
- Determine x^*'s class via the most frequent label in x^*s k nearest neighbors

4.1.4 Borderline SMOTE

Borderline SMOTE is an extension of SMOTE that focuses on creating synthetic points that are considered to be in the 'danger' set. For the most part, the algorithm follows SMOTE.

1. Determine the set of k nearest neighbors, S_{NN}, for each observation in the minority class, $x_i \in S_{min}$
2. For each x_i considered, determine the number of nearest neighbors that belong to the majority class, $|S_{NN} \cap S_{maj}|$

3. Select observations such that:

$$\frac{m}{2} \leq |S_{NN} \cap S_{maj}| < m$$

These observations are considered to be in the danger set since the observation, x_i, is in the minority class but has more nearest neighbors in the majority class than minority class neighbors and are thus more likely to be misclassified. However if all of the nearest neighbors are from the majority class, then the observation is considered noise and not to create synthetic observations.

4. These observations are then put through the normal SMOTE algorithm to create synthetic points that exist only along the border of the minority and majority classes.

(He and Garcia 2009)

4.1.5 ADASYN

ADASYN (adaptive synthetic sampling) takes a similar approach to borderline SMOTE in that it attempts to determine which observations need help in being predicted properly.

1. First ADASYN determines how many observations need to synthetically created by multiplying some number from 0 to 1 (this is a parameter that can be adjusted in cross validation) by the difference between the size of the majority class and the minority class.

$$G = (|S_{maj}| - |S_{min}|) \times \beta, \beta \in [0, 1]$$

2. Next, the k nearest neighbors of every single observation(x_i) in the minority class are found and the ratio τ_i is calculated by dividing the number of k nearest neighbors(δ_i) that belong to the majority class by the number of neighbors considered (K) and then by diving that quotient by Z which is a normalizing constant. The normalizing constant ensures that $\sum \tau_i = 1$ and is therefore a distribution function.

$$\tau_i = \frac{\delta_i / K}{Z}, i = 1, \ldots, |S_{min}|$$

3. Now the number of synthetic samples that need to be created for each minority class observation is determined by finding the product of G and τ_i

$$g_i = G \times \tau_i$$

4. Finally, g_i synthetic observations are created via regular SMOTE for each $x_i \in S_{min}$.

(He and Garcia 2009)

4.2 Models

4.2.1 Trees

Decision trees partition the input space into q disjoint regions such that:

$$\mathcal{X} = R_i \cup R_2 \cup \cdots \cup R_q \tag{4.1}$$

$$= \bigcup_{l=1}^{q} R_l \tag{4.2}$$

The prediction in each region in R_l consists of a single constant, $f(x)$, which makes trees a piecewise constant function estimator.

This partitioning is created using the response variable and a criterion. At each split, variables are scanned and one is chosen to create a rule. A threshold is created with this variable and the splitting continues until reaching a stopping condition which determines a node to be a leaf. At each node, the label is determined by majority rule.

For binary classification, the most common criterion are:

$$\text{Misclassification} : 1 - \max(p, 1-p)$$

$$\text{Gini} : 2p(1-p)$$

$$\text{Entropy/Likelihood} : -p\log(p) - (1-p)\log(1-p)$$

4.2.2 SVM

SVM uses symmetrical margins around a decision boundary to determine which points are closer to the decision boundary. In general,

the margin should be as large as possible so that the separation is as clean as possible. One disadvantage of using SVM is that there can be no missingness in the data, thus imputation is sometimes necessary. The decision boundary is determined as follows:

SVM

$$h(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b\right)$$

We want to maximize the margin such that:

$$\text{Margin} = \frac{2}{||w||}$$

where w is the p-dimensional vector of the coefficients of the variables.

Thus we want to maximize the margin subject the the w that allows most all observations to be classified correctly. Using the kernel trick, SVM can make non-linear boundaries by projecting into a higher dimension. Kernel SVM also allows us to avoid directly computing the inner product $x^T x$ by instead having us compute $\phi(x)^T \phi(x)$ as a kernel (similarity measure). In these cases, SVM is solved via quadratic programming (Seung-Seok, Sung-Hyuk, and Tappert 2010).

Kernel SVM

$$h(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T \phi(x) + b\right)$$

$$h(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

There are numerous kernels that can be used for classification. Two kernels that will be used are the Gaussian Radial Basis Function (RBF or RBFdot) and the Laplace kernel.

RBF

$K(x_i, x_j) = exp(-\delta||x_i - x_j||_2^2)$

$\delta \equiv$ bandwidth

Laplace

$K(x_i, x_j) = exp(-\delta||x_i - x_j||_1)$

$\delta \equiv$ (usually) inverse of the number of predictors

4.2.3 Novelty Detection

In novelty detection, the objective is to properly recognize when some data is different from the rest of the data. Typically, this method is used when one class has overwhelmingly more observations than another class, thus making it the majority class. The training set data only comes from the reference class, which should just be the majority class. The test data is then a mixture of both the minority and majority class. We assume that the minority class is in low density regions and since everything in the minority class is considered an outlier, both $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)^T$(the vector of sample means of each predictor) and $s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$(the vector of sample standard deviations of each predictor) are heavily influenced. Part of novelty detection involves density estimation of the one class which is normally difficult to do because covariance must be estimated and if the covariance is non-homogeneous then this estimation becomes very complicated. The basic steps of density estimation are:

- for $x_1, x_2, \ldots, x_n \overset{iid}{\sim} f_x(x, \theta)$
- Estimate the density function $f_x(x; \theta)$ using $\hat{f}_x(x; \theta)$
- Get the contours,$\hat{h}(x)$, of $f_x(x; \theta)$
- Set τ, the threshold of wrong predictions of the rare class
- Define reference, boundary, and novelty observations as the following
  - reference= $C(\tau) = \{x \in \mathcal{X}, \hat{h}(x) < \tau\}$
  - boundary= $B(\tau) = \{x \in \mathcal{X}, \hat{h}(x) = \tau\}$

- novelty= $N(\tau) = \{x \in \mathcal{X}, \hat{h}(x) > \tau\}$

By using $\phi(x_i)$ instead (where $\phi : \mathcal{X} \to \mathcal{F}$ brings values of $\mathcal{X}$ into some higher dimension $\mathcal{F}$) in the density, more complicated setups where the boundary is not spherical can be handled. Instead of estimating a covariance, kernels could be used instead which would make the algorithm essentially like one-class SVM. In one-class SVM, the model is trained only on the minority class and then tested on both the minority and majority class. By training only on the minority class, the algorithm aims to learn only the minority class so that when it is given both the majority and minority classes from the test set it can more accurately find minority class observations(Pimentel et al. 2014).

4.3 Ensemble Methods

4.3.1 Random Forest

In random forests, numerous trees are created and then aggregates across all the trees created to determine the classification of each observation.

$$\hat{f}_{RF}^{(B)}(x^{\text{new}}) \overset{\text{argmax}}{y \in Y} \left\{ \frac{1}{B} \{ I(\hat{f}^{(b)}(x^{\text{new}}) = y) \} \right\}$$

For random forest, the general algorithm consists of creating each tree and storing them for aggregation when predicting:

for ($b = 1$ to B)

- Draw a bootstrap sample from the data
- Build the b^{th} tree, $\hat{f}^{(b)}(\cdot)$
  - at each node, only q variables are considered (where there are a total of p variables and $q \lll p$ and q was selected a priori)
  - Select the best split $(x_{jk}, \tau_k), k = 1, \ldots, q$ out of the q selected variables
  - Do not prune the tree
  - Store $\hat{f}^{(b)}(\cdot)$

- Aggregate predictions of the ensemble to make the final prediction of each observation in the test set

Balanced Random Forest

To accomplish balanced random forest, the class priors of the training set are st to be equal either through over-sampling or under-sampling. Since under-sampling does better with trees, that is how random forest will be run to achieve equal class class priors. Balanced random forest runs similarly to random forest but with one twist at the beginning of each iteration:

for ($b = 1$ to B)

- Draw a bootstrap sample from the minority class with replacement and randomly draw the same amount from the majority class
- Build the b^{th} tree, $\hat{f}^{(b)}(\cdot)$
  - at each node, only q variables are considered (where there are a total of p variables and $q \lll p$ and q was selected a priori)
  - Select the best split $(x_{jk}, \tau_k), k = 1, \ldots, q$ out of the q selected variables
  - Do not prune the tree
  - Store $\hat{f}^{(b)}(\cdot)$
- Aggregate predictions of the ensemble to make the final prediction of each observation in the test set

(Chen, Liaw, and Breiman 2001)

Weighted Random Forest

In weighted random forest, the goal is to place weights on the class priors of the response in the training set such that these class weights penalize misclassifying the minority class more than the majority class. This is accomplished by placing class weights in two places in the random forest algorithm: at each terminal node and at the final class prediction. Thus at each terminal node, predictions are made by weighted majority vote. The final class prediction is then

made by aggregating these weighted votes in each individual tree. The individual trees are weighted by the average weight of all their terminal nodes(Chen, Liaw, and Breiman 2001).

4.3.2 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) takes the idea of Bayesian model averaging and applies it directly to the CART algorithm to produce a sum-of-trees models. Similar to random forest, BART sums up the predictions of each tree it creates to help determine test predictions. The only difference is that instead of averaging the sum of the predictions like random forest, BART imposes a prior on each tree to make the predictions small. This way, each prediction is constrained to be a small piece of the total prediction (similar to how weighted random forest gets its final prediction) such that when all of these pieces are added up the sum represents the final prediction.

$$\hat{f}_{BART}(x^{\text{new}}) = \sum_{b=1}^{B} \hat{f}^{(b)}(x^{\text{new}})$$

Similar to gradient boosting, BART creates priors based on the strength of predictions from each tree. This is accomplished using iterative MCMC (1200 loops are used for this Haiti data set) where the residuals from each model tell the strength of the tree and thus determine the weight of the tree(Chipman, George, and McCulloch 2012).

4.3.3 Boosting

Boosting essentially searches and obtains different learners for different portions of the data with each new learner leading to an aggregate performing better than the previous. Adaboost is a version of boosting which is fairly robust to overfitting. Adaboosting's algorithm is as follows:

- Set T, the total number of base learners
- Choose a base learner (ie tree or SVM)

- Given the data $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, n, y \in \{-1, +1\}, x_i \in \mathcal{X}\}$, initialize w at $w^{(1)} = w_1^{(1)}, w_2^{(1)}, \ldots, w_n^{(1)}$ where $w_i^{(1)} = \frac{1}{n} \equiv$weight of observation i

- For t=1 to T

  - train base learner, h_t
  - calculate the error,ϵ_t, of h_t by determining the AUC of h_t
  - determine if h_t passes pre-specified criteria to be a good enough base learner
  - compute strength of h_t; this determines the weight of model h_t

$$\alpha_t = \frac{1}{2} log(\frac{1 - \epsilon_t}{\epsilon_t})$$

  - update the weights of all the observations $i = 1, \ldots, n$
    - $\tilde{w}_i^{t+1} = w_i^{(t)} \text{exp}\{-\alpha_t y_i h_t(x_i)\}$
    - $w_i^{t+1} = \frac{\tilde{w}_i^{t+1}}{\sum_{t=1}^{T} \tilde{w}_i^{t+1}}$

Thus, $\hat{f}_{boost}(x) = \text{sign}\Big(\sum_{t=1}^{T} \alpha_t h_t(x)\Big)$ (Buhlmann and Hothorn 2005)

4.3.4 Bagging

Bagging is an ensemble method that is actually bootstrap aggregating. Bootstrapping refers to the method of resampling the data with replacement. Bagging is implemented to help reduce the variability of models that may be unstable(like classification trees). However if the model is stable, then bagging has been shown to slightly degrade the performance of the model(Breiman 1996). Lee also states that the jth element in a bagged predictor is defined as:

$$\frac{1}{M} \sum_{m=1}^{M} G(X, w^{(m)})_j$$

where $w(m)$ represents the weight of bootstrap m and $G()$ is the function that gives the predictions(Lee and Clyde 2004).

In general, bagging consists of three basic steps:

1.) Draw a bootstrap sample from the data

2.) Train a model based on the bootstrap sample and store in the bag of models

3.) After created all the models wanted, average the predicted values of all bootstrap models in bag of models on the test set

Random Subspace Learning

Random subspace learning is just like bagging, with one exception: instead of using all of the predictors, random subspace learning uses only some. One heuristic to determine the number of predictors used is to round $\sqrt{p}$, where p is the number of predictors, to the nearest integer.

Similar to bagging, random subspace learning follows a few basic steps: In general, bagging consists of three basic steps:

1.) Randomly select $d \approx \sqrt{p}$ variables to use

2.) Draw a bootstrap sample of the d variables from the data

3.) Train a model based on the bootstrap sample and store in the bag of models

4.) After created all the models wanted, average the predicted values of all bootstrap models in bag of models on the test set

(Skurichina and Duin 2002)

4.4 Performance Metrics

All models created will be compared using a few different performance metrics since there is no one metric that is necessarily better than the rest. All of the metrics used will be derived from the confusion matrix found in 4.4.

TABLE 4.4: Confusion Matrix

		Predicted Values	
		Negative	Positive
True Values	Negative	TN	FP
	Positive	FN	TP

$TN =$ True Negative, observations that were properly predicted as negative

$TP =$ True Positive, observations that were properly predicted as positive

$FN =$ False Negative, observations that were improperly predicted as negative

$FP =$ False Positive, observations that were improperly predicted as positive

$$TPR = \frac{TP}{TP + FN} \equiv \text{True Negative Rate}$$

$$FPR = \frac{FP}{FP + TN} \equiv \text{False Positive Rate}$$

4.4.1 AUC

AUC refers to the Area Under the ROC Curve. The ROC curve is formed by plotting the true positive rate versus the false positive rate (as the y and x values respectively) and is often used to determine how well a model does at prediction compared to random guessing. The AUC is thus the area underneath this curve which can be found by calculating the integral of the ROC curve. However, as He and Garcia 2009 mentions, optimizing AUC does not mean that sensitivity and precision will also be optimized. In this case, optimizing sensitivity is a priority so AUC will only be used as an additional metric for comparison between models. AUC determines the probability that the model will rank a cholera positive instance higher than a cholera negative instance.

4.4.2 Sensitivity

Also known as recall, this refers to a model's ability to properly determine when observations are positive for a disease (in this case, when they do have cholera).

$$= \frac{TP}{TP + FN}$$

4.4.3 Specificity

This refers to a models ability to properly determine when observations are negative for a disease (in this case, when they do not have

cholera).

$$= \frac{TN}{TN + FP}$$

4.4.4 F-Measure

The F-measure gives the harmonic mean of sensitivity and precision to show how well the model is doing at accurately predicting cholera positive cases.

$$\begin{aligned} F - measure &= \frac{2 \times \frac{TP}{TP+FN} \times \frac{TP}{TP+FP}}{\frac{TP}{TP+FN} + \frac{TP}{TP+FP}} \\ &= \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \end{aligned}$$

4.4.5 Geometric Mean

The geometric mean (gmean) gives a geometric average of sensitivity and specificity to show how well the model is doing at predicting both classes overall.

$$\begin{aligned} G - mean &= \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \\ &= \sqrt{Sensitivity \times Specificity} \end{aligned}$$

Chapter 5

Analysis

Since there was so many models created, to make things easier the chapter will first be split into two separate sections: the first deals with the baseline data and the second deals with the augmented data. Within these sections, models will be compared based on their base method (trees, SVM, novelty detection, model averaging, and random forest) and a single "best" model will be chosen from each of these sections to then be compared to each of the other method's "best" model in the conclusion. To determine the "best" model for each of the following sections, the F-measure, G-mean, and AUC will be used to pick winners. If there is not clear cut winner from these three methods then sensitivity will be used to determine the winner. There are instances for both the F-measure and the G-mean where 'NA' is reported, however this does not mean that models were not created. Instead 'NA' is reported whenever the measure could not be calculated (for example, the F-measure cannot be calculated when sensitivity is 0). When comparing tables, it should be noted that there are occasions when the mean of sensitivity is 0, yet the mean of the F-measure and G-mean does not reflect it. This is because NAs are ignored when creating these overall measures (since NA cannot be treated as a number) and thus the final average and standard deviation only reflect when they could be calculated.

5.1 Baseline Data

5.1.1 Trees

When creating trees, four different methods were explored: regular trees, bagged trees, boosted trees, and random sub space learning trees. When the data is left alone and no sampling techniques are

applied, both the F-measure and G-mean of all but the boosted trees are incalculable. This is because the sensitivity is always 0. Since no sampling technique is applied, the severe imbalance prevents the model from being able to learn anything about the minority class. This means that every single prediction comes out as being in the majority class, which gives these models 100% specificity, but 0% sensitivity. This is, however, flipped for boosted trees where the sensitivity is 96% but the specificity is only 2%. Sticking strictly to the best F-measure, G-mean, and AUC, over sampled bagged trees and SMOTE bagged trees are tied for top spot. Based on sensitivity, over sampled bagged trees appears to be marginally better than the SMOTE bagged trees. Although this method is chosen as the best of the other tree models, it is actually a poor model since the F-measure and G-mean are so close to 0. It appears that since SMOTE and over sampling are nearly indistinguishable that the SMOTE algorithm is optimized in this case when the sizes of the minority and majority classed are equal. This also tells us that since none of the sampling techniques with synthetic examples are significantly better that the synthetic examples created either could not help the minority class or they were created poorly. In particular, borderline-SMOTE tends to do poorly in comparison to SMOTE and ADASYN with both low averages and fairly high variances.

TABLE 5.1: Under-Sampled Trees on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.17	.23	.61	.54	.68
SD	.01	.01	.02	.04	.01

5.1.2 SVM

SVM was run in two different conditions, one used the Gaussian kernel and the other used the Laplace kernel, and three different model types, SVM, bagged SVM, and RSSL SVM. Boosted SVM was not run because continued iterations had difficulty bootstrapping samples that consisted of at least one of each class in the predictors. This meant that for each new model created, the training set was randomly pulled from the data and would more often than not give

predictors that did not represent all possible values of the predictors. Numerous methods were employed to try to combat this problem, but with the large computing time SVM called for, only a certain number of attempts could be tried before time constraints determined the algorithm too difficult to fix.

Although the kernels are different, the results are nearly indistinguishable. The Laplace kernel appears to do marginally better than the Gaussian, but the benefit is so marginal that it does not appear to matter. Similar to trees, any model run with the baseline data with no sampling techniques does very poorly. The average F-measure and G-mean of these models either could not be calculated since every model had a sensitivity of 0 or was .02 which is so close to 0 that the ensemble techniques do not seem to help much. Again, this is because the data has such a severe imbalance problem that it is near impossible to run a model that ever predicts anything in the minority class correctly. Besides the no-sampling-technique (original/regular) data, borderline-SMOTE once again does very poorly in comparison to the other sampling techniques used (fairly small average and fairly large standard deviation). Both under-sampled RSSL SVM and ADASYN RSSL appear equal for the Gaussian kernel whereas under-sampled bagged SVM is the best for the Laplace kernel, but again the results are all so close that the methods shown to be "the best" are only marginally so. While both under-sampled RSSL SVM and ADASYN RSSL have sensitivities with fairly large standard deviations, it is the former that has both a higher average and lower variance.

TABLE 5.2: Under-Sampled RSSL SVM Gaussian Kernel on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.17	.22	.60	.48	.73
SD	.01	.02	.02	.07	.05

TABLE 5.3: Under-Sampled Bagged SVM Laplace Kernel on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.17	.23	.61	.53	.68
SD	.01	.02	.02	.07	.04

5.1.3 Novelty Detection

The novelty detection method used was one-class SVM, thus it suffered the same boosting problems as SVM and could not be used. However, the two different kernels, Gaussian and Laplace, were still used in conjunction with novelty detection, bagged novelty detection, and RSSL novelty detection. Since one-class SVM is only trained on the minority class, sampling techniques cannot be used to alleviate the imbalanced classes. To clarify this further, only the positive class is taken and then is split up into training and test sets. For sampling techniques to work, they can only be done on the training set and thus in this case would only be able to be used on a training set that has one response level.

The two kernels produce similar F-measure, G-mean, and AUC results however the Gaussian kernel has clearly higher sensitivity values. For one-class SVM, bagged SVM actually does the poorest, with an average sensitivity of 0. Overall novelty detection and RSSL novelty detection tend to be on par with each other regardless of the kernel. The non-ensemble version just barely surpasses the RSSL version in terms of sensitivity though. Both results show some of the overall highest sensitivity scores of all methods done on the baseline data set (the only model to surpass these two in terms of sensitivity is boosted trees on the baseline data set).

TABLE 5.4: ND Gaussian Kernel on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.03	.11	.54	.74	.19
SD	0	.01	.02	.05	.02

TABLE 5.5: ND Laplace Kernel on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.04	.11	.54	.69	.25
SD	0	.01	.02	.09	.13

5.1.4 Model Averaging

Unsurprisingly for model averaging, the Bayesian method (BART) does (marginally) better than it's non-Bayesian twin random forest.

Under-sampling just barely surpasses the other sampling techniques. Again the outliers here are the original data and the borderline-SMOTE models. The original data does so poorly that sensitivity is always 0 and specificity is always 1 (the models always predicts data as belonging to the cholera negative class) whereas the borderline-SMOTE models have have much lower average sensitivity scores (.35) and highly variable sensitivity scores (standard deviation .12).

TABLE 5.6: Under-Sampled BART on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.17	.24	.61	.55	.70
SD	.01	.02	.02	.05	.02

5.1.5 Random Forest

By far the most successful models come from the modified random forest set. Here the lowest F-measure is .67 and the lowest G-mean is .70 (both belong to weighted and balanced random forest on the original data). All of the other models tend to be equally good. The top two models based on F-measure, G-mean and AUC are balanced random forest on the original data and balanced random forest with ADASYN. Again the synthetic examples created for SMOTE, borderline-SMOTE, and ADAYN seem to make no discernible difference. While SMOTE and the original data are very close for the first three measures, sensitivity scores show the original balanced random forest as the better method with a sensitivity score .03 better than SMOTE balanced random forest.

Interestingly, while weighted and balanced random forest on the original data has the worst F-measure, G-mean, and AUC, it also has the highest sensitivity (.96). However, there is a trade-off with the specificity (where the average is .06) that shows the model is simply predicting all the observations as belonging to the minority class which is simply the opposite of our original problem. Synthetic sampling techniques also had a problem with sensitivity; while they have the highest averages of the group (.51 to .59), they also have very large standard deviations (.13 to.18) which renders the high average useless since models can be so variable.

TABLE 5.7: Balanced Random Forest on Baseline Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.84	.84	.62	.55	.68
SD	0	0	.02	.04	.01

5.2 Augmented Data

5.2.1 Trees

As with the baseline data set, the models created with the original data had 0% sensitivity for every iteration which made the F-measure and G-mean incalculable. Boosted trees have significantly lower average F-measure scores than bagged, RSSL, and regular trees, yet their G-mean scores are in line with the other models. While the synthetic methods give slightly larger F-measures, over-sampled RSSL trees actually beats out all of the methods with a slightly larger AUC and F-measure.

TABLE 5.8: Over-Sampled RSSL Trees on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.18	.23	.61	.47	.75
SD	.02	.02	.02	.06	.04

5.2.2 SVM

Again the kernels seem to make nearly no difference to how SVM does with the data set. The only noticeable difference this time is that the Gaussian kernel tends to have a slightly higher F-measure than the Laplace kernel. For both kernels, most of the models were very close in the performance metrics. RSSL seemed to do the best regardless of the kernel though, with over-sampling just barely beating SMOTE for the Gaussian kernel and under-sampling barely beating both ADASYN and over-sampling for the Laplace kernel.

TABLE 5.9: Over-Sampled RSSL SVM Gaussian Kernel on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.18	.22	.61	.46	.75
SD	.01	.02	.02	.05	.04

TABLE 5.10: Under-Sampled RSSL SVM Laplace Kernel on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.17	.23	.61	.52	.70
SD	.01	.02	.02	.06	.05

5.2.3 Novelty Detection

Not only do the kernels seem to have no effect on the models, but the ensembles seem to also not matter for the augmented data set. Both kernels have nearly identical results with only slight changes to provide a "winner".

TABLE 5.11: ND Gaussian Kernel on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.04	.11	.52	.76	.24
SD	0	.01	.01	.05	.02

TABLE 5.12: RSSL ND Laplace Kernel on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.04	.12	.52	.64	.36
SD	0	.01	.01	.09	.08

5.2.4 Model Averaging

Overall, BART tends to do better than non-Bayesian random forest for each of the three performance metrics. For everything recorded except sensitivity for BART, all of the synthetic methods have the exact same results (the exception being that SMOTE has a standard deviation .01 less than the others in sensitivity). It appears that either

all synthetic observations created are nearly identical or the synthetic observations created do absolutely no better regardless of how they are created.

TABLE 5.13: Under-Sampled BART on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.16	.23	.61	.58	.64
SD	.01	.01	.02	.04	.02

5.2.5 Random Forest

Although under-sampling had been doing very well in the other five sections, it does the worst as far as F-measure, G-mean, and AUC are concerned. It's important to remember though that the worst here means leagues better than the best of any of model type tested (ie for the F-measure the worst here is an average of .79 with a standard deviation of .01). Under-sampling also has the highest sensitivity scores in this section (average .66). In comparison, balanced random forest on the original data has an F-measure of .97 but a sensitivity of .02. Over-sampling, borderline-SMOTE, and weighted random forest for the original data are all guilty of having high F-measure and G-mean scores and abysmal sensitivity scores. Since the goal is to correctly identify cholera positive cases, these models cannot be selected since they clearly fail to achieve this. In fact, none of the synthetic sampling techniques could be selected since the highest sensitivity rating they got was .14. In this case, we can be pickier with the models since the performance metrics are much better overall. Instead of choosing based on the best F-measure and G-mean, we'll pick the "best" model based on the highest sensitivity. In this case it happens to be under-sampled weighted random forest.

TABLE 5.14: Under-Sampled Weighted Random Forest on Augmented Data

	F-Measure	G-Mean	AUC	Sensitivity	Specificity
Mean	.79	.85	.60	.66	.54
SD	.02	.02	.02	.02	.07

Chapter 6

Suggestions for Survey Design

6.1 Structure

The survey has a huge fundamental flaw with how it was distributed: one family member (presumably the head of the household) was given the survey to fill out on behalf of their family. This meant that one person had to answer questions about their entire household, all family members elsewhere in Haiti, all family members abroad, and all family members that recently died. These questions ranged from fairly straightforward and simple (gender, age, etc.) to some very personal and potentially embarrassing questions (do they have cholera, what medical problems they have, etc.). There's a number of reasons that respondents would be inclined to not respond to questions asked in the survey since they are all fairly personal, but with this method of distributing the survey non-response bias could be introduced simply because people do not know these intimate details about others' lives. To avoid this problem, surveys could be distributed to every household and encompass only those currently living there (and those that recently died that lived in the household), however this would cost a lot more time and money than what is currently spent. There is a clear tradeoff between the accuracy of the data and the time and money spent on the survey process.

Another problem this method of survey distribution introduces is that people could potentially be double counted. For example, the head of a household in Chapel could answer questions for a family member in Sent Mari who is also claimed by another family member in Marikongo. However since the survey was designed for this distribution method, the only way to be sure to avoid double counting individuals would be to check that every name only shows up once

and that those that do show up more than once are not the same person. This would be very costly in terms of time and money though since creating a program to check would take a long time and a computer scientist would have to be paid. Overall, this problem does not affect the analysis done in this thesis, but could affect future analyses if one is interested in estimating the number or percent of people given a certain attribute.

Another structural problem lies in how the questions are presented in the survey. The first few questions ask how many people are living in the household with the respondent, but then instantly start asking about sensitive information (a lot of medical questions). These questions could make the respondent want to stop the survey, lie, or skip the question. The response bias formed from these questions include: social desirability (from not wanting to break the social norm, ie saying they do not chew tobacco or smoke), prestige (from wanting to look good to the interviewer, ie saying they do not drink or have any diseases), threat (from being anxious about answering questions, ie becoming anxious when stating medical problems or how family members died), and hostility (from being angry at being asked certain questions, ie when asked about education level and employment). These questions are all very sensitive, but necessary to ask for the purpose of the survey, however they should be asked later in the survey so that the respondent has time to build a rapport with the interviewer and thus feels more comfortable answering such questions. To fix this problem, the last section which asks questions about the respondents living space could be placed at the beginning of the survey instead.

6.2 Formatting

While Microsoft Word is a powerful tool for writing, it should not be the tool of choice for creating surveys nor for recording results. Instead, the surveys should be created with a dedicated survey formatter so that the form is as attractive, clean, and organized as possible. Furthermore, the results for this survey were not stored in any spreadsheet. Instead, the completed surveys of the same development were all saved in the same Microsoft Word file and concatenated one after another. This presents multiple problems: first, it

becomes much more difficult to keep track of all developments that have completed the survey since they are not all saved together; second, saving the survey results as filled in surveys makes it much more difficult to extract the responses; third, converting the Word documents into any other form to extract the responses ruins the formatting; and fourth, it is not easy to add in more rows for the tables that the respondents had to fill out.

One of the major problems that was presented when attempting to extract the data is that by using tables within a Word document, the spacing between columns is variable. Take for example the first table in the survey. There are numerous columns side by side that are dichotomous questions whose responses are recoded with 'W' or 'N' in the appropriate column. When extracting responses, these responses were separated by eight to twenty-five spaces. The spacial length between columns relied on if the response was recorded with a space (ie 'W ' instead of just 'W'), if columns were left blank, if columns were left all filled in, etc. The space between columns was so variable and dependent on respondents only filling in one choice that the extraction progress was hindered by having to create functions to properly extract and store data.

The survey has numerous formatting problems that greatly hindered the analysis process since more time had to be dedicated to extracting all the information from the survey properly. The first issue that is most notable is how the dichotomous questions had their responses recorded. While all other questions were limited to one column per response, these questions split up their one column in two (one for each choice). Thus, instead of recording answers simply as 'W' (Wi, Yes) or 'N' (Non, No), responses have to be have an 'W' or 'N' in the proper columns. This is made further convoluted because all of the columns are pre-filled in, thus a blank survey comes with all 'W' and 'N' columns filled (or in the case of the question on gender, 'M' and 'F'). These questions could be made easier to understand by including written directions for how to handle them. Other solutions include combining the separated columns into one such that respondents write in the answer and starting the columns as blank (nothing pre-filled).

Another issue with the formatting was that the third table in the survey included an excess two columns. These columns did not have

any questions attached to them, yet they were left in the final version in the middle of the table. This made extracting the data harder because it added a variable amount of space between the columns that were filled in. This problem is easily fixed by just getting rid of the blank columns which should have been done before the survey was given out.

Since the survey took place in Haiti, the survey was written in Haitian Creole so that all residents could understand what was being asked of them. This presents a problem when translating the survey if one is not fluent in Haitian Creole. While close to French, there are numerous differences between the two languages that can cause confusion which is not easily clarified. Microsoft Word, for example, does not recognized Haitian Creole as a language and thus when viewing the surveys in Word, French is as the recognized language which prevents someone from telling if a word is actually misspelled or not. An exact translation of the survey would be very useful to determine exactly what questions are asking and how they should be answered without spending hours translating them. It would also be very useful because the survey questions tend to be written with very little to no spaces used between words. While French can be used to make educated guesses on where words should end, it can only help so much when both grammar and spelling are misused.

One of the largest problems this led to was misunderstanding what one of the questions was asking. The question asked 'Li gen Larouli', which when translated roughly gives 'he/she has Larouli'. Countless hours were spent scouring the internet for Creole dictionaries to translate 'Larouli', but nothing turned up. Originally it had been assumed to be a specific disease that was rare or had another name. It wasn't until a few months after starting to work with the data that a written response in one of the surveys gave a hint to what 'Larouli' could be. A respondent had filled in the question on any other diseases/illnesses by saying that they had/got 'lawouli', which when translated meant that they had dizziness. Whether the survey creators meant 'lawouli' or 'Larouli' is unknown, but it is more likely that 'lawouli' was intended since it can be translated and makes sense with the other questions being asked.

The lack of written instructions makes analyzing the data harder as well since it is unknown how certain questions were intended to

be answered or treated. The question asking respondents to list other illnesses/diseases brings the biggest issue since it is an unstructured (fill in the blank style) question. Respondents were free to answer the question however they like. This meant that some answered in single words, some in phrases, and the majority answered by not filling in the question at all. Ignoring how difficult it is to analyze such questions for the time being, one of the problems this question presented was that it was unclear what the nonresponse meant. Some respondents had even filled in answers stating that they were fine or that they had no diseases/illnesses. Were respondents told to leave the question blank if they had nothing to add or did they not want to answer because they felt embarrassed/threatened? It is unclear since there are no directions to refer back to, and with a nonresponse rate of over 50% the data extracted for the question is nearly useless. Another example of directions being needed for answering questions comes in when the survey asks if there were any pregnant women living with the respondent and then gives two follow up questions asking for their ages and if they would be brought to a hospital. There was a clear lack of understanding that the questions were follow up questions because several people that answered yes did not fill in the ages of the pregnant women and most people that answered no still answered if the pregnant women would be brought to the hospital. This problem could be avoided by having directions and by clearly marking questions as follow ups.

Overall, the next iteration of this survey should include the following changes:

1. Clear written directions on how to answer every question
2. Clearly marked follow up questions
3. Blank surveys should be entirely blank, no answers pre-filled in
4. One column per question, dichotomous questions are answered by writing either response possible

6.3 Questions

For the most part, the survey asks structured (closed-ended) questions that are fairly straightforward and easy to analyze. There are a few unstructured questions asked throughout the survey however. Most of these questions were left out of the analysis because of the difficulty they presented when trying to coerce them into a categorical form. Since there were thousands of observations recorded, it becomes very difficult to coerce these type of answers into something easy to analyze because there are so many different answers. While text mining can be used to make this easier, an intimate knowledge of Haitian Creole is needed since many people write the same thing numerous different ways (ie 'Douel', 'Doule', 'douel', 'Doule do', 'Doule nan', 'Dole nan', etc. are all answers given when asked about other illnesses and diseases that probably should mean the same thing). These numerous different ways of writing similar answers makes text mining even harder to accomplish since it is difficult to determine exactly what each person means and how similar their answer is to someone else's.

Another problem with the unstructured questions is the use of shorthand. Both questions on how old the people are and how much education they have are answered in shorthand. The education level answers are all answered in one of the following patterns (note that \\d means some digit from 0 to 9):

- Pres
- Seg
- Filo
- Reto
- Iniv
- \\deAf
- \\deS
- \\de

These are all shorthand for different grade levels in different schools (primary, secondary, university) as well as different education titles.

However there is nothing to indicate what each answer stands for and thus without a native speaker it is impossible to be able to differentiate when someone says '3rd grade primary school' and '3rd year of high school'.

The age question had a similar problem. While most people just wrote numbers, implying years old, some would write numbers with a single letter next to them. This signified months ('m'), weeks ('s'), and days ('d'). This was easy enough to figure out since French uses a similar scheme for time, but when someone had written a number with an e next to it, more research was required to know how to treat the answer. After hours looking, it was finally discovered that the 'e' meant hours; someone had recently given birth and then took the survey. While confusing, these questions were still easier to deal with than the disease question which had to be left out due to the large non-response and varied answers. These questions could be made easier by explicitly stating the scale or coding one should use. For example, education could be written in years total in school or an inclusive list of mutually exclusive items could be given in which people select one answer.

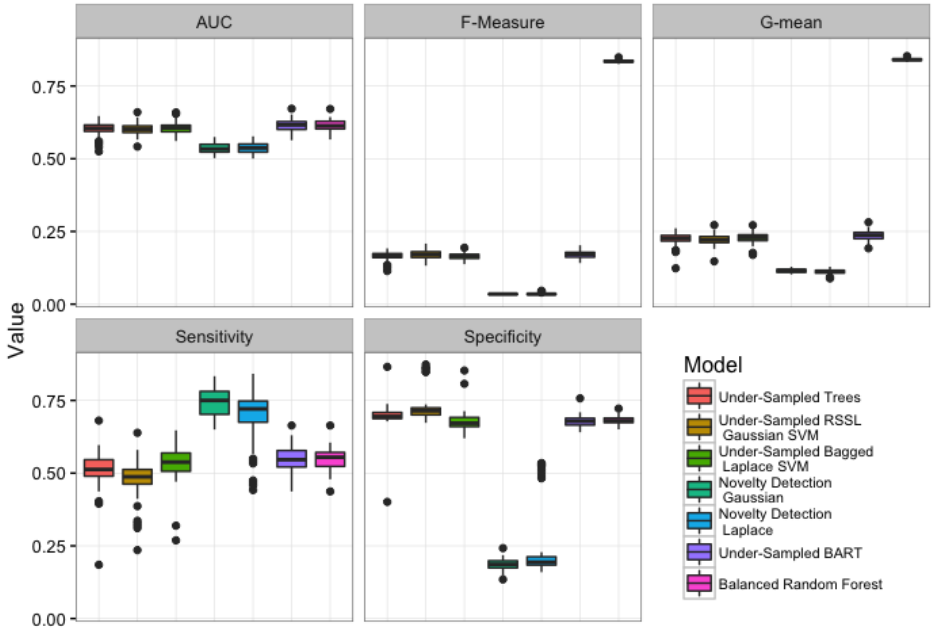
Chapter 7

Conclusion

Overall,the modified random forests outperform other algorithms for both versions of the data. As seen in 7.1 and 7.2 the model chosen from the augmented random forest algorithms has clearly superior G-mean and F-measure values. However, there does not seem to be much gain in AUC for these models, and furthermore, they do not have the largest sensitivity measures. The Gaussian kernel novelty detection models give the highest sensitivity values. Surprisingly this and specificity are where the kernels appear to have different strengths. The Gaussian kernel gives a higher sensitivity, but a lower specificity. While it's important to be able to correctly predict minority cases, all of the novelty detection methods seem to suffer from over predicting observations as belonging to the minority class, which results in extremely low specificities. Between the two data sets though there appears to be a difference in how precise the models can get. The baseline data set tends to produce models that normally have outliers (this can be seen for each of the performance metrics), yet the augmented data set produces models with much less outliers (ie the sensitivity metric shows no outliers). At the same time, the augmented data shows that models have slightly larger standard deviations. This illustrates that while the extra data added fills in some information, it also introduces more variability.

A better comparison of the augmented and baseline data set can be seen in 7.3 which showcases the difference between balanced random forest for the baseline data set (chosen as the "best" model for the baseline models) and under-sampled weighted random forest for the augmented data set (chosen as the "best" model for the augmented models). The increase in variance becomes clear here as the augmented data produces larger boxplots for each performance metric. With this increase in variance though comes a decrease in bias for

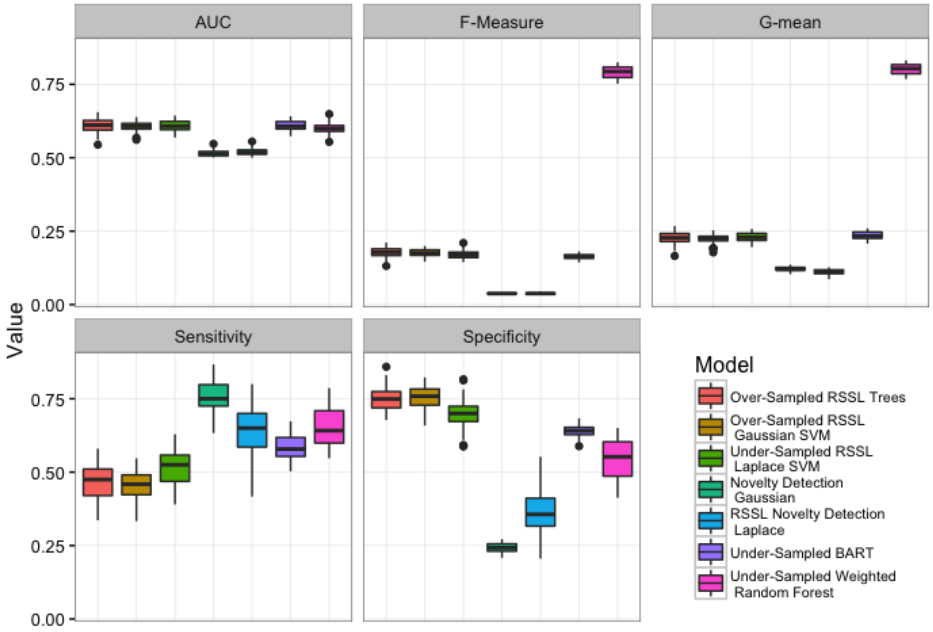
FIGURE 7.1: Comparison of all selected models for the baseline data



determining the minority class correctly as shown in the sensitivity boxplots. Here the augmented data is superior in correctly predicting the minority class. Both models are clearly the best of their respective data sets, yet there is not a clear winner between them. If predicting the minority class is more important than anything else, then the augmented data would have to be used. However if misclassifying cholera negative cases is also important then an argument for the baseline data can be made.

The difficulty in picking a clear "best" model extends from the worst models types to the best model types. Each model has it's strengths and weaknesses, but no one outshines the others. The modified random forest is a clear winner in that it gives much higher F-measure and G-means. As compared to the random forest in the model averaging section that do not balance classes or give classes different weights, it is obvious that these modifications are exactly what produced the better results. Using sampling techniques to balance the class priors before models was not enough to combat the imbalance problem since trees and random forest are creating with subsets of the data at each node which is not guaranteed to maintain

FIGURE 7.2: Comparison of all selected models for the augmented data

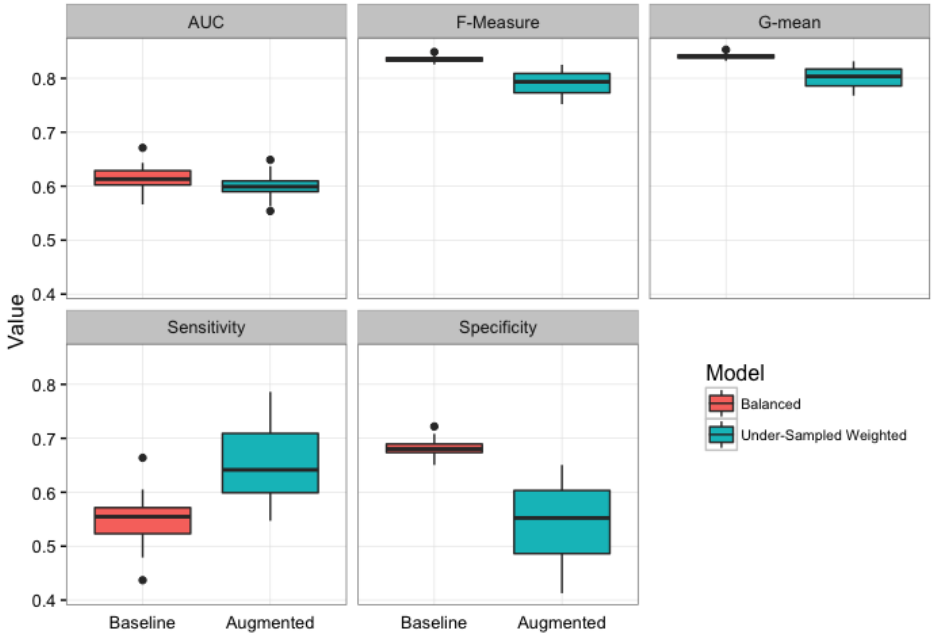


the equal class priors.

Furthermore, SVM seemed to fail the data in both kernel types which tells us that either the data requires another kernel, not yet investigated, or a boundary between classes for this data does not exist. To explore this idea further, other kernels could be tested in the future to see if any significant improvements can be made in the predictions. Properly learning the minority class is rare due to cholera positive cases being truly rare thus making the imbalance due to absolute rarity (once you have it, you have less than a day to get it treated before you die) but could be made even more complicated if subconcepts do exist in the minority class which would also introduce a within-class imbalance. This combination makes learning the minority highly difficult.

An alternative reason for why models fair so poorly could be simply because the data itself is of poor measurement quality. It is possible that predictors could be cleaned up better to give more quality data. The education levels could be grouped such that all levels in primary school are in one category, all levels in secondary school are together, graduate school, etc. The disease question could also be

FIGURE 7.3: Comparison of selected random forests models



added in and coerced to be categorical using text mining methods such as grabbing key words to create categories (ie back pain, head pain, fatigue, etc.). Data collection could also be improved so that answers are directly related to immediate family or one person at a time instead of a family and their extended family which they may not know intimately. More questions revolving around how people access, use, and clean water as well as how they deal with their wastes could also help to provide clues on who is more inclined towards getting sick.

More research into different model types could also be done in the future to try a wider selection. This could include methods like SMOTE-boost and modified KNN methods that require much longer computing time. Since priors and Bayesian attributes seem to do better in learning and predicting the data, it is important that future methods have a solid Bayesian foundation. Bayesian networks, which is a probabilistic model that tries to determine what predictors are conditional on each other and which are independent, is worth investigating.

Appendix A

Survey Materials

Dat: Kòd:

Seksyon: Zòn: Abitasyon: Lakou: _____ KB kay KB mennaj? KB gason? KB fiy?

KB moun? KB granmoun? konbyen timoun? Lòt enfòmasyon si

genyen:_____

1-Di non yo, yonn apwè lòt pou mwen.

00	Ki Laj li	Ki Seks li	Ki Klas li fè	Ki sa lap fè?	Li te gen Kolera	Eske li bwekase	bwe dite	bwèta fya	Li rimen	Li pwan ablak	Li fè Opwasyon	Li gen Lanoul	Li gen tèt fè mal	Li fè fantasyon	Ki lòt maladi li konn abitye genyen pi souvan	Li ale lòpital
01		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
02		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
03		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
04		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
05		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
06		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
07		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
08		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
09		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
10		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
11		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
12		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
13		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N
14		M F			W N	W N	W N	W N	W N	W N	W N	W N	W N	W N		W N

2-Konbyen moun nan kay la kap viv nan yon lòt komin nan peyia ?

00	Ki laj li	Ki Sèks li	Ki Klasifikasyon fè	Ki kote li ye	Ki sa li apfè	Li te genKòlera	Li bwèla fè	Li bwèta fyà	Li fimen	Li Pwan Tabak	Li feÒpwe-syon	Li genTèt fè mal	Li fèTan-syon	Ki lòtmaladi li konnabityegenyen pi souvan	Li ale lopital
01		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
02		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
03		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
04		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
05		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
06		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
07		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
08		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
09		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
10		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
11		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
12		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
13		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
14		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
15		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
16		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
17		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
18		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
19		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N
20		M F				W N	W N	W N	W N	W N	W N	W N	W N		W N

2-Konbyen moun nan kay la kap viv nan yon lòt komin nan peyia ?

00	Ki laj li	Ki Sèks li		Ki Klas li fè	Ki kote li ye	Ki sa li apfè	Li te genKolera		Li bwèka fe		Li bwèta fya		Li fimen		Li Pwan Tabak		Li féOpwe-syon		Li genTètè mal		Li féTan-syon		Ki lòtmaladi li konnabityegenyen pi souvan	Li ale lopital	
		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
01		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
02		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
03		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
04		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
05		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
06		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
07		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
08		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
09		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
10		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
11		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
12		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
13		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
14		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
15		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
16		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
17		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
18		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
19		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N
20		M	F				W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N		W	N

5-Koman kò kay nap viv la fèt: an Blòk___/an klis___/an pwela___/an bwa___/an tòl___/ 6-Kouvèti kay la: an beton___/an tòl___/an pay___/ an pwela___/an chengèl___/ 7-Atè an dedan kay la: an beton___/an tè___/An mozayik___/an planch 8-Eske nou trete dlo nap sèvi ?Wi___/Non___/ 9-Kòman nou fè sa ? Bouyi___/Bokit sistèm___/Avèk solèy____/ Klòw___/ Sistèm UV___/Lòt mwayen : _____ 10-Eske nou gen latrin nan kay la ?Wi___/ Non___/ 11-Kòman li fèt ? Yon twou___/An mi oublòk___/ Nan jouk___/Konstwi sou dlo___/Twalèt ijenik___/Lòt mwayen : _____ 12-Eske moun nan kay la abitye koupe pyebwa pou fè chabon ou byen planch? Wi___/ Non___/ 13-ki kote nou elve zannimo yo ?_____ eske nou fè jaden? Wi____/ Non____/ pou kiyès tè a ye? Fèmye____/ Tèprive____/Tèleta_____/ 14-Ki elvaj moun nan kay la pwatike plis ?Cheval___/bourik___/milèt___/ Kabrit___/bèf___/kochon___/chyen___/chat___/poul___/kana___/kòdenn___/pentad___/Pwason___/Kribich___/ 15-Eskegen fanm ansent nan kay la ? Wi___/ Non___/ 16- Ki laj li ? ____/____/____/ ____/ ____/ ____/ 17-Eske li ale lopital ? Wi____/ Non____/ 18-Lè yon moun nan fanmiy lan malad ki sa nou fè pou ede li ? Nou fè remèd lakay la___/ Nale kay chalatan___/ Nale Kay bòkò___/ Nale Lopital___/ Nou pa fè anyen menm___/ 19-Di nou 3 bagay ou ta renmen ki fèt nan zòn nan, nan zafè devlòpman? 1- mache kominote _____ 2- sant sante _____ 3- latrin _____ 20-Ki konsèy ou ta renmen bay A.S.B. ?

_____

_____

21-Remak ankèt la ?_____

_____

_____
Anketè a

Appendix B

Data Tables

B.1 Baseline Data

TABLE B.1: F-measure

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	NA	NA	.17	.01	.17	.01
Bagged Trees	NA	NA	.17	.01	.17	.01
Boosted Trees	.11	0	.08	.01	.07	.01
RSSL Trees	NA	NA	.17	.01	.18	.02
SVM-RBF	.02	0	.16	.01	.16	.01
Bagged SVM-RBF	.02	0	.16	.01	.16	.01
RSSL SVM-RBF	NA	NA	.17	.02	.17	.01
SVM-Laplace	.02	0	.17	.01	.16	.01
Bagged SVM-Laplace	.02	0	.17	.01	.17	.01
RSSL SVM-Laplace	NA	NA	.17	.01	.17	.01
ND-RBF	.03	0				
Bagged ND-RBF	.03	0				
RSSL ND-RBF	.03	.01				
ND-Laplace	.04	0				
Bagged ND-Laplace	.03	0				
RSSL ND-Laplace	.03	.01				
BART	NA	NA	.17	.01	.17	.01
RF	NA	NA	.17	.01	.17	.01
Balanced RF	.84	0	.84	.01	.84	.01
Weighted RF	.84	.02	.84	.01	.84	.02
W&B RF	.67	.02	.84	.01	.83	.01

TABLE B.2: F-measure Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.16	.01	.12	.06	.16	.01
Bagged Trees	.17	.01	.11	.04	.17	.01
Boosted Trees	.08	.01	.10	.01	.08	.01
RSSL Trees	.17	.01	.08	.04	.18	.01
SVM-RBF	.16	.01	.11	.03	.16	.01
Bagged SVM-RBF	.16	.01	.11	.03	.16	.01
RSSL SVM-RBF	.17	.02	.04	.04	.17	.01
SVM-Laplace	.17	.01	.10	.04	.16	.01
Bagged SVM-Laplace	.17	.01	.10	.04	.16	.01
RSSL SVM-Laplace	.18	.01	.05	.05	.18	.01
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.17	.01	.12	.03	.17	.01
RF	.17	.01	.12	.02	.17	.01
Balanced RF	.84	.01	.84	.05	.84	.01
Weighted RF	.84	.03	.80	.06	.84	.02
W&B RF	.84	.02	.80	.06	.84	.01

TABLE B.3: G-mean

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	NA	NA	.22	.01	.22	.02
Bagged Trees	NA	NA	.23	.01	.23	.02
Boosted Trees	.24	0	.15	.01	.12	.04
RSSL Trees	NA	NA	.22	.02	.22	.02
SVM-RBF	.04	.03	.22	.02	.22	.02
Bagged SVM-RBF	.04	.04	.22	.02	.22	.02
RSSL SVM-RBF	NA	NA	.22	.02	.22	.02
SVM-Laplace	.01	.03	.22	.02	.23	.02
Bagged SVM-Laplace	.02	.03	.22	.02	.23	.02
RSSL SVM-Laplace	NA	NA	.22	.02	.22	.02
ND-RBF	.11	.01				
Bagged ND-RBF	.11	.01				
RSSL ND-RBF	.10	.02				
ND-Laplace	.11	.01				
Bagged ND-Laplace	.11	.01				
RSSL ND-Laplace	.08	.03				
BART	NA	NA	.23	.02	.24	.02
RF	NA	NA	.23	.02	.23	.02
Balanced RF	.84	0	.84	.01	.84	.01
Weighted RF	.84	.01	.84	.01	.84	.02
W&B RF	.70	.02	.84	.01	.84	.01

TABLE B.4: G-mean Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.22	.02	.12	.06	.22	.02
Bagged Trees	.23	.01	.11	.05	.23	.01
Boosted Trees	.15	.01	.21	.02	.14	.01
RSSL Trees	.22	.02	.09	.04	.22	.02
SVM-RBF	.21	.02	.11	.03	.21	.02
Bagged SVM-RBF	.21	.02	.11	.03	.21	.02
RSSL SVM-RBF	.22	.03	.08	.04	.22	.02
SVM-Laplace	.22	.02	.10	.03	.22	.02
Bagged SVM-Laplace	.22	.02	.10	.03	.22	.02
RSSL SVM-Laplace	.23	.02	.08	.05	.22	.02
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.23	.02	.16	.03	.23	.02
RF	.22	.02	.16	.03	.23	.02
Balanced RF	.85	0	.85	.05	.85	.01
Weighted RF	.85	.02	.81	.06	.85	.02
W&B RF	.85	.02	.81	.06	.84	.01

TABLE B.5: AUC

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	.50	0	.60	.02	.60	.02
Bagged Trees	.50	0	.61	.02	.61	.02
Boosted Trees	.51	.01	.61	.02	.60	.03
RSSL Trees	.50	0	.60	.02	.61	.02
SVM-RBF	.50	0	.59	.02	.60	.02
Bagged SVM-RBF	.50	0	.59	.02	.60	.02
RSSL SVM-RBF	.50	0	.60	.02	.60	.02
SVM-Laplace	.50	0	.60	.02	.61	.02
Bagged SVM-Laplace	.50	0	.60	.02	.61	.02
RSSL SVM-Laplace	.50	0	.61	.02	.60	.02
ND-RBF	.54	.02				
Bagged ND-RBF	.54	.02				
RSSL ND-RBF	.53	.03				
ND-Laplace	.54	.02				
Bagged ND-Laplace	.54	.02				
RSSL ND-Laplace	.54	.03				
BART	.50	0	.61	.02	.61	.02
RF	.50	0	.61	.02	.61	.02
Balanced RF	.62	.02	.61	.02	.61	.02
Weighted RF	.61	.02	.61	.02	.61	.02
W&B RF	.51	.01	.61	.02	.61	.02

TABLE B.6: AUC Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.60	.02	.53	.03	.60	.02
Bagged Trees	.61	.02	.52	.02	.61	.02
Boosted Trees	.60	.02	.55	.03	.61	.02
RSSL Trees	.60	.02	.51	.01	.61	.02
SVM-RBF	.59	.02	.53	.02	.59	.02
Bagged SVM-RBF	.59	.02	.53	.02	.59	.02
RSSL SVM-RBF	.60	.03	.50	.01	.60	.02
SVM-Laplace	.60	.02	.52	.02	.60	.02
Bagged SVM-Laplace	.60	.02	.52	.02	.60	.02
RSSL SVM-Laplace	.61	.02	.51	.01	.61	.02
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.60	.02	.54	.03	.60	.02
RF	.60	.02	.54	.03	.60	.02
Balanced RF	.61	.02	.54	.03	.61	.02
Weighted RF	.60	.02	.58	.02	.61	.02
W&B RF	.61	.02	.58	.02	.61	.02

TABLE B.7: Sensitivity

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	0	0	.5	.04	.51	.07
Bagged Trees	0	0	.54	.04	.54	.05
Boosted Trees	.96	0.02	.48	.04	.38	.14
RSSL Trees	0	0	.47	.09	.47	.09
SVM-RBF	0	0	.49	.05	.53	.07
Bagged SVM-RBF	0	0	.49	.05	.52	.06
RSSL SVM-RBF	0	0	.47	.09	.48	.07
SVM-Laplace	0	0	.50	.04	.54	.06
Bagged SVM-Laplace	0	0	.50	.04	.53	.07
RSSL SVM-Laplace	0	0	.48	.09	.46	.11
ND-RBF	.74	.05				
Bagged ND-RBF	0	0				
RSSL ND-RBF	.62	.23				
ND-Laplace	.69	.09				
Bagged ND-Laplace	0	0				
RSSL ND-Laplace	.46	.29				
BART	0	0	.53	.04	.55	.05
RF	0	0	.52	.05	.54	.05
Balanced RF	.55	.04	.53	.05	.54	.05
Weighted RF	.53	.06	.53	.06	.54	.08
W&B RF	.96	.07	.53	.04	.55	.05

TABLE B.8: Sensitivity Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.50	.04	.12	.10	.50	.06
Bagged Trees	.53	.04	.09	.08	.53	.04
Boosted Trees	.48	.04	.81	.09	.48	.04
RSSL Trees	.45	.09	.02	.03	.46	.09
SVM-RBF	.48	.05	.11	.05	.48	.05
Bagged SVM-RBF	.48	.05	.11	.05	.48	.05
RSSL SVM-RBF	.46	.10	.01	.03	.47	.09
SVM-Laplace	.49	.04	.10	.05	.48	.06
Bagged SVM-Laplace	.49	.04	.10	.05	.49	.06
RSSL SVM-Laplace	.47	.09	.01	.03	.46	.09
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.52	.05	.35	.12	.52	.05
RF	.51	.05	.35	.12	.51	.05
Balanced RF	.52	.05	.37	.08	.52	.06
Weighted RF	.51	.13	.59	.18	.51	.18
W&B RF	.51	.05	.57	.06	.52	.05

TABLE B.9: Specificity

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	1.00	0	.70	.02	.69	.05
Bagged Trees	1.00	0	.68	.01	.68	.02
Boosted Trees	.02	.01	.31	.02	.42	.19
RSSL Trees	1.00	0	.74	.07	.74	.07
SVM-RBF	1.00	0	.70	.02	.68	.04
Bagged SVM-RBF	1.00	0	.70	.02	.67	.04
RSSL SVM-RBF	1.00	0	.73	.06	.73	.05
SVM-Laplace	1.00	0	.71	.02	.67	.04
Bagged SVM-Laplace	1.00	0	.70	.02	.68	.04
RSSL SVM-Laplace	1.00	0	.74	.06	.74	.08
ND-RBF	.19	.02				
Bagged ND-RBF	.19	.02				
RSSL ND-RBF	.32	.21				
ND-Laplace	.25	.13				
Bagged ND-Laplace	.22	.05				
RSSL ND-Laplace	.47	.30				
BART	1.00	0	.69	.02	.70	.02
RF	1.00	0	.69	.02	.70	.02
Balanced RF	.68	.01	.69	.01	.68	.03
Weighted RF	.68	.05	.69	.03	.68	.07
W&B RF	.06	.09	.56	.02	.69	.05

TABLE B.10: Specificity Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.70	.02	.94	.06	.70	.04
Bagged Trees	.69	.01	.95	.05	.69	.01
Boosted Trees	.31	.02	.09	.05	.31	.02
RSSL Trees	.75	.07	.99	.01	.75	.07
SVM-RBF	.71	.02	.95	.03	.70	.02
Bagged SVM-RBF	.71	.02	.95	.03	.70	.02
RSSL SVM-RBF	.74	.06	1.00	.01	.74	.06
SVM-Laplace	.71	.02	.95	.02	.71	.03
Bagged SVM-Laplace	.71	.02	.95	.02	.71	.04
RSSL SVM-Laplace	.74	.06	1.00	.01	.75	.06
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.69	.02	.70	.16	.69	.02
RF	.70	.02	.70	.16	.70	.02
Balanced RF	.70	.01	.69	.15	.70	.03
Weighted RF	.70	.07	.54	.20	.70	.06
W&B RF	.70	.05	.56	.19	.69	.03

B.2 Augmented Data

TABLE B.11: F-measure

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	NA	NA	.15	.01	.14	.02
Bagged Trees	NA	NA	.15	.01	.15	.01
Boosted Trees	.11	0	.08	.01	.07	.01
RSSL Trees	NA	NA	.18	.02	.17	.02
SVM-RBF	NA	NA	.15	.01	.16	.01
Bagged SVM-RBF	NA	NA	.15	.01	.16	.01
RSSL SVM-RBF	NA	NA	.18	.01	.17	.01
SVM-Laplace	NA	NA	.17	.02	.16	.01
Bagged SVM-Laplace	NA	NA	.17	.02	.16	.01
RSSL SVM-Laplace	NA	NA	.18	.01	.17	.01
ND-RBF	.04	0				
Bagged ND-RBF	.04	0				
RSSL ND-RBF	.04	0				
ND-Laplace	.04	0				
Bagged ND-Laplace	.04	0				
RSSL ND-Laplace	.04	0				
BART	NA	NA	.17	.01	.16	.01
RF	.02	.01	.08	.02	.15	.01
Balanced RF	.86	.01	.96	0	.79	.01
Weighted RF	.97	0	.95	0	.79	.02
W&B RF	.83	.01	.95	0	.79	.02

TABLE B.12: F-measure Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.16	.02	.15	.02	.15	.02
Bagged Trees	.16	.02	.15	.02	.17	.02
Boosted Trees	.09	.01	.09	.01	.09	.01
RSSL Trees	.18	.03	.18	.02	.17	.03
SVM-RBF	.16	.02	.14	.02	.16	.02
Bagged SVM-RBF	.16	.02	.14	.02	.16	.02
RSSL SVM-RBF	.18	.02	.17	.02	.17	.02
SVM-Laplace	.17	.02	.14	.03	.17	.02
Bagged SVM-Laplace	.17	.02	.14	.03	.17	.02
RSSL SVM-Laplace	.17	.02	.16	.02	.18	.02
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.15	.02	.15	.02	.15	.02
RF	.10	.02	.05	.02	.10	.02
Balanced RF	.95	0	.96	0	.95	0
Weighted RF	.94	0	.96	0	.94	0
W&B RF	.94	.01	.96	0	.94	.01

TABLE B.13: G-mean

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	NA	NA	.21	.02	.21	.02
Bagged Trees	NA	NA	.21	.01	.22	.02
Boosted Trees	.22	.01	.14	.03	.05	.03
RSSL Trees	NA	NA	.23	.02	.23	.02
SVM-RBF	NA	NA	.18	.01	.23	.01
Bagged SVM-RBF	NA	NA	.18	.01	.23	.01
RSSL SVM-RBF	NA	NA	.22	.02	.23	.02
SVM-Laplace	NA	NA	.18	.02	.23	.02
Bagged SVM-Laplace	NA	NA	.18	.02	.23	.02
RSSL SVM-Laplace	NA	NA	.22	.02	.23	.02
ND-RBF	.11	.01				
Bagged ND-RBF	.11	.01				
RSSL ND-RBF	.10	.02				
ND-Laplace	.12	.01				
Bagged ND-Laplace	.12	.01				
RSSL ND-Laplace	.12	.01				
BART	NA	NA	.20	.02	.23	.01
RF	.05	.03	.09	.02	.23	.01
Balanced RF	.85	0	.85	.05	.85	.01
Weighted RF	.85	.02	.81	.06	.85	.02
W&B RF	.85	.02	.81	.06	.84	.01

TABLE B.14: G-mean Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.19	.02	.18	.02	.19	.03
Bagged Trees	.19	.02	.18	.02	.19	.02
Boosted Trees	.17	.02	.17	.02	.18	.02
RSSL Trees	.19	.03	.20	.03	.19	.03
SVM-RBF	.18	.02	.14	.02	.18	.02
Bagged SVM-RBF	.18	.02	.14	.02	.18	.02
RSSL SVM-RBF	.20	.02	.19	.02	.19	.02
SVM-Laplace	.18	.02	.14	.03	.18	.02
Bagged SVM-Laplace	.18	.02	.15	.03	.18	.02
RSSL SVM-Laplace	.18	.02	.17	.02	.19	.02
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.15	.02	.15	.02	.15	.02
RF	.11	.02	.07	.03	.11	.02
Balanced RF	.95	0	.96	0	.95	0
Weighted RF	.94	0	.96	0	.94	0
W&B RF	.94	.01	.96	0	.94	.01

TABLE B.15: AUC

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	.50	0	.57	.02	.57	.03
Bagged Trees	.50	0	.59	.02	.59	.02
Boosted Trees	.54	.02	.57	.02	.54	.02
RSSL Trees	.50	0	.61	.02	.61	.03
SVM-RBF	.50	0	.57	.01	.60	.02
Bagged SVM-RBF	.50	0	.57	.01	.60	.02
RSSL SVM-RBF	.50	0	.61	.02	.60	.02
SVM-Laplace	.50	0	.57	.01	.60	.02
Bagged SVM-Laplace	.50	0	.57	.02	.60	.02
RSSL SVM-Laplace	.50	0	.60	.02	.61	.02
ND-RBF	.52	.01				
Bagged ND-RBF	.51	.01				
RSSL ND-RBF	.52	.01				
ND-Laplace	.51	.01				
Bagged ND-Laplace	.51	.01				
RSSL ND-Laplace	.52	.01				
BART	.50	0	.59	.02	.61	.02
RF	.50	0	.52	.01	.59	.02
Balanced RF	.60	.01	.52	.01	.59	.02
Weighted RF	.51	0	.52	.01	.60	.02
W&B RF	.60	.02	.52	.01	.60	.02

TABLE B.16: AUC Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.57	.02	.57	.02	.57	.02
Bagged Trees	.58	.02	.57	.02	.58	.02
Boosted Trees	.58	.02	.57	.02	.57	.02
RSSL Trees	.58	.02	.58	.02	.57	.02
SVM-RBF	.57	.02	.54	.01	.57	.02
Bagged SVM-RBF	.57	.02	.54	.01	.57	.02
RSSL SVM-RBF	.58	.02	.58	.02	.58	.02
SVM-Laplace	.57	.02	.54	.01	.57	.02
Bagged SVM-Laplace	.57	.02	.54	.01	.57	.02
RSSL SVM-Laplace	.57	.02	.56	.02	.57	.02
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.55	.01	.55	.01	.55	.01
RF	.52	.01	.51	.01	.52	.01
Balanced RF	.53	.01	.51	.01	.53	.01
Weighted RF	.54	.01	.52	.01	.54	.01
W&B RF	.54	.01	.52	.01	.54	.01

TABLE B.17: Sensitivity

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	0	0	.51	.13	.57	.13
Bagged Trees	0	0	.50	.08	.58	.06
Boosted Trees	.86	.04	.45	.14	.08	.06
RSSL Trees	0	0	.47	.06	.51	.07
SVM-RBF	0	0	.35	.03	.56	.05
Bagged SVM-RBF	0	0	.35	.03	.57	.05
RSSL SVM-RBF	0	0	.46	.05	.51	.07
SVM-Laplace	0	0	.26	.03	.56	.06
Bagged SVM-Laplace	0	0	.26	.03	.57	.05
RSSL SVM-Laplace	0	0	.43	.04	.52	.06
ND-RBF	.76	.05				
Bagged ND-RBF	0	0				
RSSL ND-RBF	.76	.06				
ND-Laplace	.76	.05				
Bagged ND-Laplace	0	0				
RSSL ND-Laplace	.64	.09				
BART	0	0	.38	.04	.58	.04
RF	.01	0	.05	.02	.64	.05
Balanced RF	.45	.03	.06	.01	.64	.04
Weighted RF	.02	.02	.08	.02	.66	.02
W&B RF	.54	.04	.08	.07	.66	.08

TABLE B.18: Sensitivity Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.34	.08	.35	.09	.36	.10
Bagged Trees	.35	.05	.34	.06	.33	.04
Boosted Trees	.61	.09	.59	.08	.64	.08
RSSL Trees	.27	.05	.30	.05	.27	.06
SVM-RBF	.26	.04	.16	.03	.26	.04
Bagged SVM-RBF	.26	.04	.17	.03	.26	.04
RSSL SVM-RBF	.30	.05	.30	.05	.30	.06
SVM-Laplace	.25	.04	.14	.03	.25	.04
Bagged SVM-Laplace	.25	.04	.14	.03	.25	.04
RSSL SVM-Laplace	.23	.05	.21	.04	.25	.05
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.16	.03	.16	.02	.15	.02
RF	.08	.02	.03	.01	.08	.02
Balanced RF	.10	.02	.04	.02	.10	.02
Weighted RF	.14	.03	.06	.02	.13	.02
W&B RF	.14	.04	.06	.02	.14	.03

TABLE B.19: Specificity

	Original		Over		Under	
	Mean	SD	Mean	SD	Mean	SD
Trees	1.00	0	.63	.14	.57	.15
Bagged Trees	1.00	0	.67	.08	.60	.04
Boosted Trees	.07	.01	.41	.15	.84	.07
RSSL Trees	1.00	0	.75	.04	.70	.05
SVM-RBF	1.00	0	.78	.01	.63	.04
Bagged SVM-RBF	1.00	0	.78	.01	.62	.03
RSSL SVM-RBF	1.00	0	.75	.04	.70	.05
SVM-Laplace	1.00	0	.89	.01	.64	.04
Bagged SVM-Laplace	1.00	0	.89	.01	.64	.03
RSSL SVM-Laplace	1.00	0	.77	.03	.70	.05
ND-RBF	.24	.02				
Bagged ND-RBF	.28	.01				
RSSL ND-RBF	.22	.05				
ND-Laplace	.25	.02				
Bagged ND-Laplace	.29	.02				
RSSL ND-Laplace	.36	.08				
BART	1.00	0	.79	.01	.64	.02
RF	1.00	0	.98	0	.54	.03
Balanced RF	.75	.01	.98	0	.54	.02
Weighted RF	.99	0	.96	.01	.54	.07
W&B RF	.66	.03	.96	.01	.54	.07

TABLE B.20: Specificity Cont.

	SMOTE		BL-SMOTE		ADASYN	
	Mean	SD	Mean	SD	Mean	SD
Trees	.80	.06	.78	.07	.79	.07
Bagged Trees	.81	.03	.80	.04	.82	.03
Boosted Trees	.23	.08	.26	.07	.21	.06
RSSL Trees	.89	.02	.87	.03	.88	.03
SVM-RBF	.87	.01	.93	.01	.87	.01
Bagged SVM-RBF	.87	.01	.93	.01	.87	.01
RSSL SVM-RBF	.86	.02	.86	.03	.86	.04
SVM-Laplace	.89	.01	.95	.01	.88	.01
Bagged SVM-Laplace	.89	.01	.95	.01	.89	.01
RSSL SVM-Laplace	.90	.03	.90	.02	.89	.03
ND-RBF						
Bagged ND-RBF						
RSSL ND-RBF						
ND-Laplace						
Bagged ND-Laplace						
RSSL ND-Laplace						
BART	.94	.01	.94	.01	.94	.01
RF	.97	0	.99	0	.97	0
Balanced RF	.96	0	.99	.01	.96	.01
Weighted RF	.94	0	.98	.01	.94	.01
W&B RF	.94	0	.98	.01	.93	.01

Bibliography

Amini, Shahram M., and Christopher F. Parmeter. 2011. "Bayesian model averaging in R". *Computational Statistics & Data Analysis* 56 (6): 1–35. ISSN: 01679473. doi:10.1080/02664760902899774. https://core.ac.uk/download/pdf/6494889.pdf.

Breiman, Leo. 1996. "Bagging Predictors". *Machine Learning* 24 (421): 123–140. ISSN: 0885-6125. doi:10.1007/BF00058655.

Buhlmann, Peter, and Torsten Hothorn. 2005. "Boosting Algorithms: Regularization, Prediction and Model Fitting". *Statistical Science*: 1–52.

Chawla, Nitesh V, et al. 2003. "SMOTEBoost: improving prediction of the minority class in boosting". *Principles of Knowledge Discovery in Databases, PKDD-2003*: 107–119. ISSN: 03029743. doi:10.1007/b13634. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.80.1499.

Chen, Chao, Andy Liaw, and Leo Breiman. 2001. "Using random forest to learn imbalanced data". *Machine Learning* 45 (1999): 5–32. ISSN: 0016-1152. doi:ley.edu/sites/default/files/tech-reports/666.pdf.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2012. "BART: Bayesian additive regression trees". *Annals of Applied Statistics* 6 (1): 266–298. ISSN: 19326157. doi:10.1214/09-AOAS285. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1273584455.

Clyde, Merlise. 2003. "Model Averaging". Chap. 13 in *Subjective and Objective Bayesian Statistics*, ed. by James Press, 1–25. New York: John Wiley & Sons. https://stat.duke.edu/courses/Spring05/sta244/Handouts/press.pdf.

Grzyma\la-Busse, J W, J Stefanowski, and Sz. Wilk. 2005. "A Comparison of Two Approaches to Data Mining from Imbalanced Data". *Journal of Intelligent Manufacturing* 16 (6): 565–573.

Han, Hui, Wen-yuan Wang, and Bing-huan Mao. 2005. "Borderline-SMOTE : A New Over-Sampling Method in": 878–887.

Hansen, B. 2007. "Lecture notes Model Averaging": 138–144.

He, Haibo, and Edwardo Garcia. 2009. "Learning from Imbalanced Data Sets." *IEEE transactions on Knowledge and Data Engineering* 21 (9): 1263–1284. ISSN: 1041-4347. doi:10.1109/TKDE.2008.239. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5128907.

He, Haibo, et al. 2008. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". *Proceedings of the International Joint Conference on Neural Networks*, no. 3: 1322–1328. ISSN: 1098-7576. doi:10.1109/IJCNN.2008.4633969.

Hernandez, Julio, Jesús Ariel Carrasco-Ochoa, and José Francisco Martínez-Trinidad. 2013. "An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets". In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I*, ed. by José Ruiz-Shulcloper and Gabriella di Baja, 262–269. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-41822-8. doi:10.1007/978-3-642-41822-8{_}33. http://dx.doi.org/10.1007/978-3-642-41822-8_33.

Hoeting, Jennifer a, et al. 1999. "Bayesian model averaging: a tutorial". *Statistical Science* 14 (4): 382–417. ISSN: 08834237. doi:10.2307/2676803. http://www.jstor.org/stable/2676803.

Lee, HKH, and MA Clyde. 2004. "Lossless online Bayesian bagging". *The Journal of Machine Learning Research* 5:143–151. ISSN: 15337928. http://dl.acm.org/citation.cfm?id=1005337.

Lusa, Lara, and Rok Blagus. 2013. "SMOTE for high-dimensional class-imbalanced data". *BMC bioinformatics* 14 (1): 106. ISSN: 1471-2105. doi:10.1186/1471-2105-14-106.

Pimentel, Marco A F, et al. 2014. "A review of novelty detection". *Signal Processing* 99:215–249. ISSN: 01651684. doi:10.1016/j.sigpro.2013.12.026. http://dx.doi.org/10.1016/j.sigpro.2013.12.026.

Raj, Des. 1971. "Planning of Surveys: Initial Steps". In *The design of sample surveys*, 111–125. New York: McGraw-Hill. ISBN: 9780070511552.

Rea, Louis M, and Richard A Parker. 1992. "Developing Survey Questions". In *Designing and conducting survey research: a comprehensive guide*, 56–79. San Francisco: Jossey-Bass Publishers.

Rubin, Donald B. 1987. "Introduction". In *Multiple imputation for nonresponse in surveys*. New York: Wiley. ISBN: 047108705X.

Seung-Seok, Choi, Cha Sung-Hyuk, and Charles C Tappert. 2010. "A Survey of Binary Similarity and Distance Measures." *Journal of Systemics, Cybernetics & Informatics* 8 (1): 43–48. ISSN: 16904524. doi:10.1.1.352.6123. http://www.iiisci.org/journal/CV$/sci/pdfs/GS315JG.pdf.

Skurichina, Marina, and Robert P W Duin. 2002. "Bagging, boosting and the random subspace method for linear classifiers". *Pattern Analysis and Applications* 5 (2): 121–135. ISSN: 14337541. doi:10.1007/s100440200011.

Sun, Yanmin, et al. 2007. "Cost-sensitive boosting for classification of imbalanced data". *Pattern Recognition* 40 (12): 3358–3378. ISSN: 00313203. doi:10.1016/j.patcog.2007.04.009.

Wang, Benjamin X., and Nathalie Japkowicz. 2010. "Boosting support vector machines for imbalanced data sets". *Knowledge and Information Systems* 25 (1): 1–20. ISSN: 02191377. doi:10.1007/s10115-009-0198-y.

Wasserman, L. 2000. "Bayesian Model Selection and Model Averaging." *Journal of mathematical psychology* 44 (1): 92–107. ISSN: 0022-2496. doi:10.1006/jmps.1999.1278. http://www.ncbi.nlm.nih.gov/pubmed/10733859.