Ensemble of naïve Bayesian approaches for the study of biofilm development in water distribution systems

E. Ramos-Martínez [a]*,* M. Herrera[b], J. Izquierdo[a], R. Pérez-García[a]

[a]*FluIng-IMM, Universitat Politècnica de València,*
C. de Vera s/n, Edif 5C, 46022 Valencia, Spain

[b]*BATir - Université libre de Bruxelles,Av F. Roosevelt, 50 CP 194/2*
B-1050 Brussels, Belgium

Abstract

The survival and regrowth of microorganisms in drinking water distribution systems (DWDSs) can be affected not only by biological aspects but also by the interaction of various other factors. Some of these factors have been found to be clearly related to biofilm development in DWDSs. However, the complexity of the microenvironment under study and the biofilm growth characteristics have so far led the various methodologies applied to produce ambiguous or not easily comparable, and thus not very useful, results. In this study we compile the information currently available on biofilm ecology in DWDSs and apply various machine learning algorithms based on naïve Bayesian networks. In addition, as a step forward, we also use ensemble methods. These methods have been widely adopted to improve the robustness and the overall prediction accuracy of single models through their accumulative experience on the performance of multiple applications in learning systems. We claim that they also reduce the high uncertainty associated with the process of biofilm development in DWDSs. Specifically, we propose alternatives for the base naïve Bayesian model to outperform its individual results while maintaining its simplicity. These alternatives include augmentation of the arcs in the graph and bagging initial approaches. Then, both ideas are combined by a hybrid algorithm that produces explanatory decision trees. As a result, it is possible to achieve a deeper understanding of the consequences that the interaction of the relevant hydraulic and physical factors of DWDSs has for biofilm development.

Keywords: drinking water distribution systems; biofilm; naïve Bayesian; ensemble methods.

1 Introduction

Biofilms develop in drinking water distribution systems (DWDSs) as layers of microorganisms bound by a matrix of organic polymers and attached to pipe walls. These

[**] *Corresponding author. Email: evarama@upv.es

communities of organisms form spontaneously in DWDSs due to the presence of moisture, bind strongly against the initial repulsion at the inner pipe wall, and modify the pipe as they capture more nutrients and new bacteria [18]. A developed biofilm is very resistant [16] and may pose a significant problem when a clean and disinfected environment is needed. Thus, taking into account that nowadays one of the main challenges of drinking water utilities is to ensure microbial high quality supply, this approach to biofilms represents a real paradigm in urban water supply management [21, 30].

Besides the health risk that biofilms involve, due to their role as a pathogen shelter [18], a number of additional problems associated with biofilm development in DWDSs can be identified. Among others, aesthetic deterioration of water [7], proliferation of higher organisms [2], biocorrosion [29] and disinfectant decay [5] are universally recognized.

Survival and regrowth of microorganisms in DWDSs it is affected not only by biological aspects but also by the interaction of various factors [32]. Numerous studies have been approached in relation to the influence that a number of characteristics of the DWDSs have in biofilm development [24, 27, 28, 33]. Nevertheless, their joint influence, apart from few exceptions [26], has been scarcely studied, due to the complexity of the community and the environment under study [21]. This work aims to approach this problem studying the effect that the interaction of the relevant hydraulic and physical characteristics of the DWDSs has in biofilm development. As a consequence, it achieves deeper understanding of the cause-effect relations involved in biofilm assessment. To address the difficulties usually found in data relative to this work environment, we propose focussing our analysis on a naïve Bayes approach [22]. It supposes a simplification of unrestricted Bayesian networks. However, it often achieves similar accuracy even comparable on performance with decision trees or neural networks classifiers [8].

This paper tries to eendow naïve Bayes classifiers with more predictive power, without renouncing to the simplicity of the initial approach. Three different alternatives are proposed: a tree augmented naïve Bayes classifier (TAN) [6, 11], a bagging combination of naïve Bayes approaches (BNB), and an ensemble of these approaches by a modified version of Bayesian network tree [12] (NBT), where a bagging process is applied in their leaf nodes. All of these alternatives improve the results of any straightforward application of the naïve Bayes algorithm (B-NBT), can work with small sized data bases in a suitable way, and are easy to implement and computationally efficient. All these processes have been mainly implemented in R Language [20], by their package 'e1071' [17] together with the R interface to Weka, 'RWeka' [9, 31].

The present study has the following structure. Section 2 introduces suitable combinations of bayesian-based algorithms in the study of biofilm development in DWDSs. Section 3 presents the biofilm data base to be analyzed. Section 4 focuses on the outcomes obtained by the implementation of the above techniques. Finally, Section 5 reviews the implications of the different proposed procedures in the biofilm assessment and introduce further applications.

2 Naïve Bayesian approaches

This paper focuses on naïve bayesian methods and a number of variants in order to assess the biofilm development degree in DWDS. A naïve Bayesian network classifier, which is sometimes called naïve Bayes classifier (NBC for short), has a very simple structure while its classification performance in practice is surprisingly high. The structure assumes that all the attributes are mutually independent given the class. This simplify the way in which the process works.

Let T be a training set of samples, each with their class labels. There are k classes, $C_1, \ldots, C_k$. Each sample is represented by an $n-$dimensional vector, $\mathbf{X} = \{x_1, \ldots, x_n\}$, depicting n measured values of the n attributes. Then, the classifier will predict that $\mathbf{X}$ belongs to the class having the highest *a posteriori* probability, conditioned on $\mathbf{X}$ (see Equation 1).

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \text{ for } 1 \leq j \leq n, j \neq i. \tag{1}$$

The probabilities involved in this model can be calculated following Equation 2.

$$P(C_h|\mathbf{X}) \propto P(C_h) \prod_{i=1}^{n} P(X_i|C_h), \tag{2}$$

where $P(C_h)$ represents the *a priori* information respect to the classification of the variable of interest in the class h.

In order to predict the corresponding class of $\mathbf{X}$, the expression $P(C_i)P(\mathbf{X}|C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of $\mathbf{X}$ is C_i if and only if it is the class that maximizes $P(C_i)P(\mathbf{X}|C_i)$. Thus, a final classifier is obtained by Equation 3.

$$\arg\max_{c} P(C) \prod_{i=1}^{n} P(X_i = x_i|C = c). \tag{3}$$

Despite the fact that the far-reaching independence assumptions are often inaccurate, an NBC has several properties that make it exceptionally useful in practice. In particular, the decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This, for example, helps alleviate problems stemming from the curse of dimensionality and also allows working with missing and scarce data.

2.1 Augmented Bayesian Classifiers

The tree augmented naïve (TAN) classifier [6] is obtained by allowing each attribute to have at most one other attribute as a parent, in addition to the class. Therefore a maximum of $n - 1$ number of edges can be added to an NBC to obtain a TAN classifier. Then, this algorithm outperforms the accuracy of the naïve Bayes algorithm by relaxing the conditional independence assumption [11].

In order to apply, in a computationally efficient way, the algorithm, Keogh & Pazzani [11] proposes the following approach for each TAN classifier to be built. In the first step, the results of equation 2 are stored in a $J \times I$ matrix, (J is the number of instances in the training set, I is the number of distinct classes) where each element is the probability that example j belongs to class C_i. When testing a new classifier has an arc from node X_b to node X_a, we adjust the matrix by multiplying element (i, j) by

$$\frac{P(X_a = x_{a_j} | C_i, X_b = x_{b_j})}{P(X_a = x_{a_j} | C_i)}. \tag{4}$$

This approach means that the time taken to evaluate one instance of a TAN classifier will be independent of the number of attributes. So, the speed-up achieved by this optimization is approximately of order n, the number of nodes.

2.2 A combined approach: bagging naïve bayes

Bootstrap aggregating, *bagging*, predictors are used to generate multiple versions of a predictor which are then used to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets [1]. Bagging votes classifiers generated by different bootstrap samples: $S_1, \ldots, S_B$. From each sample S_i a classifier is induced by the same learning algorithm (NBC in this case). Classifiers obtained in this manner are then combined by majority voting respect to the B classifiers (see Figure 1). This aggregation process helps to mitigate the impact of random variation and provides stability to the classifier method [13].

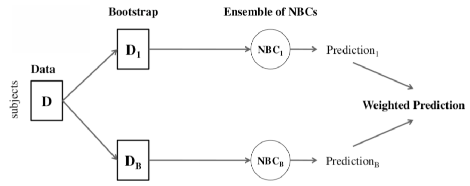

Figure 1: Bagging naïve Bayes process

The procedure, iterated for B bootstrap samples, results in an ensemble of B NBCs, each one with a possibly different set of features. Unseen subjects are then classified by making each NBC estimate output class probabilities, and by averaging the probabilities across all B NBCs. Such an approach increases the robustness of the predictions [1].

2.3 A hybrid approach: Bagging leafs of naïve Bayesian network trees

A decision tree is a decision support tool that uses a schematic tree-shaped diagram graph which model decisions and their possible consequences. Each branch of the decision tree represents a possible decision or occurrence. The tree structure shows how one choice leads to the next, and the use of branches indicates that each option is mutually exclusive. Decision trees are learned in a top-down fashion, with an algorithm known as Top-Down Induction of Decision Trees (TDIDT), recursive partitioning, or divide and conquer learning. The algorithm selects the best attribute for the root of the tree, splits the set of examples into disjoint sets, and then adds corresponding nodes and branches to the tree [23].

A naïve Bayesian network tree applies different NBCs to different regions of the input space inducing a hybrid decision tree classifiers: the decision tree nodes contain univariate splits as regular decision trees, but their leafs contains NBCs [12]. In this way, the main part of this approach is by classical recursive partitioning schemes as in usual decision trees (such as the above-mentioned TDIDT). However, the corresponding leaf nodes created are NBCs instead of nodes predicting a single class.

Besides the NBT approach, this paper also proposes a new strategie on leaf nodes. It consists on bootstraping elements in the leaf nodes, followed by a bagging process based on NBCs. This approach tries to take advantage of the tree structure of the data, which obtains a suitable starting point to apply a re-sampling method. Thus, it represents a first step where the process diminishes variability and prevent bias in the creation of the bootstrap process; which redound in an optimized bagging classifier. By the nature of this ensemble learning method proposed, the overall process still being simple and computationally eficient.

3 Case-study

To carry out this work, information about the hydraulic and physical characteristics of DWDSs that affect biofilm development have been gathered from experts. The result of the aggregation of this information consists of a database including the following variables.

(i) Water age (*wage*). This node represents the residence time of water in the DWDS under study. It is important to include this variable because the older water influences on the decay of disinfectant residual, sediment deposition and temperature increase [10]. All of them are factors that favor the biofilm development. In this case, the water age is measured by an index between 0 and 1, which is increasing with the age of water.

(ii) Flow velocity (*velocity*). The formation of biofilms increases with the flow velocity of water due to the increase of nutrient mass transfer [15]. Nevertheless, specific velocities between 3-4 m/s may favor its release [4].

(iii) Hydraulic regime (*flow*). It can be turbulent or laminar. It has been demonstrated that some biofilms tend to be more active in turbulent flow, having more mass per cm^2, increased cell density and distinct morphology than biofilms in laminar flow [25].

(v) Pipe material (*material*). The pipe materials can be classified into metallic, plastic, or cement (see Table 1). In general, metallic pipes tend to develop more biofilm than cement pipes and these more than the plastic ones [19]. This is mainly explained by the roughness of the materials, since rough surfaces provide more surface to biofilm growth and protect them from water shear forces. Pipes with a rough surface have greater potential for biofilms growth [2].

(vi) Pipe age (*page*). The age of the pipes of the DWDS under study are divided into *young*, *medium*, and *old* (see Table 1). The accumulation of corrosion products and dissolved substances in the older pipes can increase their roughness [3], thus favoring the biofilm development. In addition, older deposits may have greater biomass and bacteria content [2].

(vii) Biofilm (*biofilm*). Heterotrophic plate count (HPC) was chosen as biofilm quantification method. Although there are another methods, this is the most commonly used, having more data available. Based on the observed biofilm data distribution and on the experts criteria these data was divided in *normal* and *high* biofilm development (see Table 1).

Table 1: Variables and categories of the database

P.MATERIAL	P.AGE (years)
metallic	high [$\geq$ 31]
cement	medium [11-30]
plastic	reduced [0-10]
HIDRAULIC REGIME	BIOFILM (HPC/cm^2)
laminar	high [$\geq$ 10^7]
turbulent	normal [0-10^6]
WATER AGE [0-1]	FLOW VELOCITY (m/s)

4 Results and discussion

4.1 Learning augmented Bayesian classifiers in practice

From the obtained results (Figure 2) we have a clear relationship between the pipe material, and the pipe age and the flow velocity. It may be explained by the fact that, normally, in the DWDSs, the older pipes are metallic while the newest ones are plastic due to the historical evolution of the pipelines and the distribution systems design. In the same way, the flow velocity in the pipes could be related to the diameter of the pipes, and this with the pipe material.

In relation to the results obtained after the stratified cross-validation, we found that the Kappa statistic value is just 0.199. This statistic value determines to what extent the observed agreement exceeds the expected results obtained by chance. According to the margins that [14] proposed to assess the degree of agreement according to this Kappa index it means that, in this case, the degree of agreement is trivial. In the same way, the root mean squared error (RMSE) is 0.3813. These results show us that this algorithm is not strong enough to develop a good model in our case study. This is an example of the obstacles that are found when studying biofilm from an applied point of view and explains, the fact, that till now there are almost no studies of the effect that the joint interaction of the different characteristics of the DWDSs has in biofilm development.

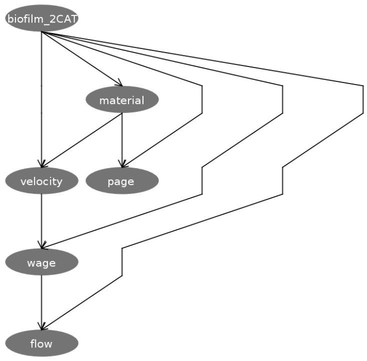

Figure 2: Learning Augmented Bayesian Classifiers

4.2 Bagging naïve Bayes classifiers in practice

According to the BNB results there is a probability of 0.78 of having a normal biofilm development in a pipe in a DWDSs. It is worth to notice that in the cases where the biofilm development is high the velocity tends to be a little bit lower than in the cases with normal biofilm development (see Table 2). The same happens with the water age, in the cases with high biofilm development, the water age tends to be lower than in the cases with normal biofilm development (Table 2). These results do not correspond with those expected from the bibliography. In the case of the velocity, it could be thought that with the increase in velocity biofilm could suffer detachment and, because of that get more biofilm development when the flow velocity is lower. However, the mean velocity obtained in the case of normal biofilm development is too low to suppose that detachment phenomena are involved. Anyway, it must be highlighted that the obtained standard deviations are quite big in all cases.

When we focus on the stratified cross-validation results we see that these are better than the results obtained previously with the TAN but they are not good enough yet. The RMSE is very similar to the previous one, while the Kappa has increased. The Kappa index has a value of 0.319 and the RMSE of 0.3854.

4.3 Bagging leafs of naïve Bayesian network trees in practice

It is worth noting that, in the obtained tree (Figure 3), the pipe material appears as the first classification node. According to this, pipe material seems to be the main factor that affects biofilm development. This variable is known to be important since it is assumed that metallic pipes tend to develop more biofilm than cement pipes and these more than plastic pipes [19]. In BNT shows that in the case of cement pipes, the NBC is obtained directly, while, in the other cases, there is a second division node.

When the pipes are metallic the second node is the velocity, discriminating between velocities under and above 1.015 m/s (Figure 3). The flow velocity is considered low when it is not bigger than around 0.8 m/s. This value is near the one obtained in the NBT, so it can be said that the node seems to discriminate between the cases with low

Table 2: Bagging naïve Bayes results

Attribute	H	N
FLOW		
T	46	156
L	1	11
VELOCITY		
mean	0.787	1.035
SD	0.599	1.067
WATER AGE		
mean	0.036	0.093
SD	0.037	0.103
PIPE MATERIAL		
C	46	6
M	9	41
P	33	121
PIPE AGE		
M	8	3
O	11	14
Y	29	151

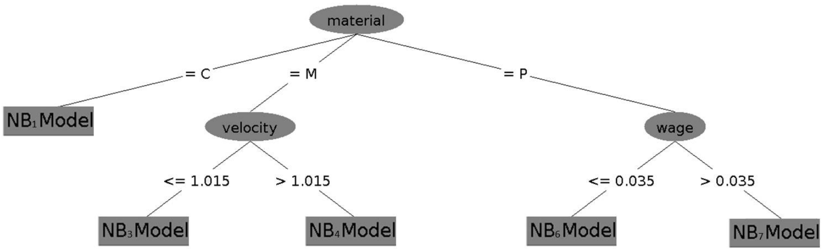

Figure 3: Naïve Bayes Tree

flow velocity and the rest of the cases. The effect of the velocity could be relevant in the cases of metallic pipes because the corrosion products of the metals can be used as nutrients by the biofilm favoring its development [2] and the velocity is related to the increase of the nutrient mass transfer [15].

In the branch of plastic pipes the second division node is the water age. The threshold value (0.035) attracts attention since the water age index goes from 0 to 1. According to the results it seems to be distinguished between the pipes located in the chlorination points and the rest of the pipes of the DWDSs.

Table 3: Naïve Bayes Tree Probabilities

NB	H	N
1	0.67	0.33
3	0.02	0.98
4	0.58	0.42
6	0.52	0.48
7	0.04	0.96

The probablilities of the NB classifier obtained in each branch of the tree when applying B-NBT are shown in the Table 3. According to them, in the cement pipes the probability of suffering high biofilm development is higher than the probability of having normal biofilm development. In the cases of metallic pipes, if the flow velocity is low (below 1.015 m/s) normal biofilm development is expected, while when the velocity is higher than 1.015 m/s having high biofilm development is more probable (Table 3). These results may be in relation to the fact that with the flow velocity increase, the mass transfer of nutrients also increases [15]. When the pipes are plastic the cases with a water age bigger than 0.035 have more probabilities of having a normal biofilm development (N) than those with a water age under 0.035 (Table 3).Having fewer probability of high biofilm development when the water age is lower (Table 3) is not what it was expected according to the bibliography [10]. However, it may be explained by the fact that this variable appears to be relevant just in the case of plastic pipes and it may be due to interaction with other factors under-considered in other biofilm works.

According to the results the pipes with highest tendency to support high biofilm development would be those made of cement. In the same way, the metallic pipes with a flow velocity higher than 1.015 m/s and the plastic pipes with a water age under 0.035 are also prone to develop a high development. The pipes with less risk of having a high biofilm development would be the metallic pipes with a velocity under 1.015 m/s and the plastic pipes with a water age bigger than 0.035 (Table 3).

When applying just the NBT, the results of the stratified cross-validation show a Kappa statistic value of 0.66 and a root mean squared error of 0.2891. These results are much better that the ones obtained with the TAN algorithm and the BNB. However, applying B-NBT these results are improved. The index Kappa increased to 0.708 while the RMSE is reduced to 0.210 (Figure 4). The model, in this case, achieves a good degree of agreement [14], classifying correctly 88.57% of the cases.

4.4 Summary

The complexity of the community and the environment under study is the reason why there is a lack of works that study the influence that the whole set of characteristics of the DWDSs has into biofilm development. We have approached this problem through the naïve Bayes algorithm showing that the intricacy of the problem under study is a big handicap to get the final aim.

It has been demonstrated that ensemble techniques are more useful in this complex case, obtaining better results than the simpler methods because the iterations increased the robustness of the process. However, this has not been enough to get a good model. Hybrid ensemble techniques have been necessary to achieve good results (Figure 4). The accumulative experience on the performance of multiple applications of different learning systems is the suitable way to achieve our aim, reducing the uncertainty and improving the overall prediction accuracy of the model. Furthermore, the approach proposed in this paper, has demonstrated to be the best way to achive a good model in this case. It has shown to be able to exploit the advantages of the different techniques used. Avoiding bias and decreasing the uncertainty with the classification trees, improving the efficiency through the naïve bayes classifier and, finally, gaining accuracy applying bagging.

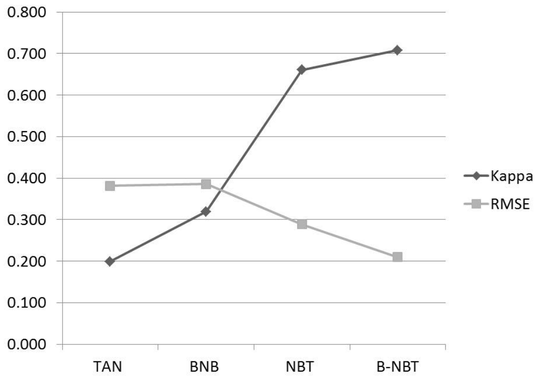

Figure 4: Kappa statistic value and RMSE for TAN, BNB, NBT and B-NBT

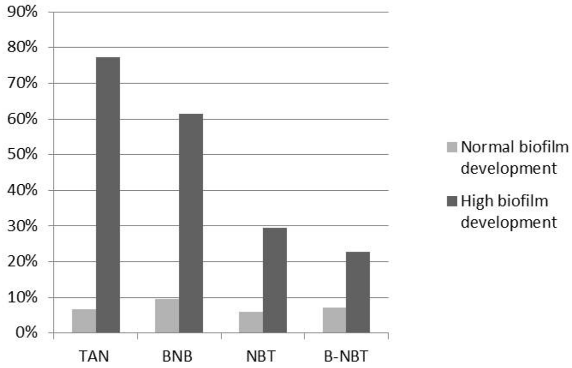

Figure 5: Error percentages of the confusion matrix

The improvement of the output is not shown only in the goodness indexes, but also in the results (Figure 5). Although, in the cases with normal biofilm development, the error percentage of the B-NBT method is a little bit bigger than the obtained with the NBT, the error rate of the cases with high biofilm development, in which we are interested due to their implication in noumerous DWDSs problems, is greatly reduced. So it, once more, the methodology that we have developed in this paper seems to be the best way to approach this problem.

5 Conclusions

This work is characterized by offering an innovative perspective in the study of biofilms development in DWDSs. On one hand, techniques of intelligent data analysis are introduced in this field. On the other, the effect that the interaction among the hydraulic and physical characteristics of the DWDSs, relevant in biofilm development, has been introduced in this proposal.

Until now, the effect that the different physical and hydraulic characteristics of the DWDSs have on biofilms development were studied individually in a majority of cases. This is due to the complexity of the community and the environment under study, together with the scarcity on data. These are the main reasons to propose simple algorithms to approach biofilm assessment in DWDSs. To gain robustness and accuracy different combinations of simple processes, which produce good performance, have been introduced. Thus, by an ensemble algorithm we have achieved deeper understanding of the consequences that the interaction of the relevant hydraulic and physical factors of the DWDSs has in biofilm development. But we have gone further, increasing even more the accuracy of the obtained model by B-NBT, reaching better results while the procces still being simple and computationally eficient. According to the results obtained in this work, there are some pipes with a greater tendency to have high biofilm development. As a general rule, the water utility managers, in order not to suffer high biofilm development in the DWDSs, should consider to avoid the presence of cement pipes and favor medium or high flow velocities in the metallic pipes and water ages above 0.035 in the plastic pipes.

This paper represents the base of a more complex tool of decision support system in DWDSs. The problems related to biofilm development in these systems could be solved or mitigated thanks to it. The present work is an advance in the study of biofilms development in DWDSs as it allows deeper understanding of the ecology of these communities and facilitates better understanding of the processes and interactions that occur in DWDSs related to the development of these communities.

6 Acknowledgements

This work has been performed under the support of the project IDAWAS, DPI2009-11591 of the Dirección General de Investigacin del Ministerio de Ciencia e Innovacin (Spain) and ACOMP/2011/188 of the Conselleria de Educació of the Generalitat Valenciana. We want to express our gratitude to the research grant (FPI), Ministerio de Ciencia e Innovacin (ref.: BES-2010-039145).

References

[1] L. Breiman, *Bagging predictors*, Machine Learning 24 (1996), pp. 123–140.

[2] S. Chowdhury, *Heterotrophic bacteria in drinking water distribution system: a review*, Environmental Monitoring and Assesment (2011), pp. 2407–2415.

[3] R. Christensen, *Age effects on iron-based pipes in water distribution systems*, Master's thesis (2009).

[4] T. Cloete, D. Westard, and S. van Vuuren, *Dynamic response of biofilm to pipe surface and fluid velocity*, Water Science and Tecnology 45 (2003), pp. 57–59.

[5] D. de Beer, R. Srinivansa, and P. Stewart, *Direct measurement of chlorine penetration into biofilms during disinfection*, Applied Environmental Microbiology 60 (1994), pp. 4339–4344.

[6] N. Friedman, D. Geiger, and M. Goldszmidt, *Bayesian network classifiers*, Machine Learning 29 (1997), pp. 131–163.

[7] M.F. Gelves, *Deterioro de la calidad del agua por el posible desprendimiento de las biopelículas en las redes de distribución de agua potable*, Universidad de los Andes (2005).

[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier (2006).

[9] K. Hornik, C. Buchta, and A. Zeileis, *Open-source machine learning: R meets Weka*, Computational Statistics 24 (2009), pp. 225–232.

[10] IWA, *Effects of water age on distribution system water quality*, Tech. rep., 2002.

[11] E.J. Keogh and M.J. Pazzani, *Learning the structure of augmented bayesian classifiers*, Artificial Intelligence Tools 11(4) (2002), pp. 587–601.

[12] R. Kohavi, *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*, in *2nd International Conference on Knowkedge Discovery and Data Mining*, AAAI Press, 1996, pp. 202–207.

[13] S. Kotsiantis, *Combining bagging, boosting, rotation forest and random subspace methods*, Artificial Intelligence Review 35(3) (2011), pp. 223–240.

[14] J. Landis and G. Koch, *The measurement of observed agreement for categorical data*, Biometrics 33 (1977), pp. 159–174.

[15] M. Lehtola, M. Laxandera, I. Miettinena, A. Hirvonec, T. Vartiainenb, and P. Martikainenc, *The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system...*, Water Research 40 (2006), pp. 2151–2160.

[16] C. Mains, *Biofilm control in distribution systems*, Biofilm control in distribution systems 8 (2008).

[17] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien* (2012), URL `http://CRAN.R-project.org/package=e1071`, r package version 1.6-1.

[18] M. Momba, R. Kfir, S. Venter, and T. Cloete, *An overview of biofilm formation in distribution systems and its impact on the deterioration of water quality*, Water Research Comission (2000).

[19] P.M. Niquette, P. Servais, and R. Savoir, *Impacts of pipe materials on densities of fixed bacterial biomass in a drinking water distribution system*, Water Resources 34 (2000), pp. 1952–1956.

[20] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2011), URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

[21] E. Ramos-Martínez, *Evaluación del desarrollo de biofilms en los sistemas de distribución de agua potable mediante la extracción de conocimiento a través de los datos (knowledge discovery data kdd)*, Master's thesis (2012).

[22] B. Ripley, Cambridge University Press (1996).

[23] G.I. Sammut, Claude; Webb (ed.), *Encyclopedia of Machine Learning*, Springer (2010).

[24] C.C. Silhan, J. and A. H.J., *Effect of temperature and pipe material on biofilm formation and survival of escherichia coli in used drinking water pipes: a laboratory-based study*, JWater Science and Technology 54 (2006), pp. 49–56.

[25] M. Simoes, M. Pereira, and M. Vieira, *The role of hydrodynamic stress on the phenotypic characteristics of single and binary biofilms of pseudomonas fluorescens*, Water Science and Tecnology 55 (2007), pp. 437–445.

[26] M. Simoes, L. Simoes, I. Machado, M. Pereira, and M. Vieira, *Control of flow-generated biofilms with surfactants evidence of resistance and recovery*, Food and Bioproducts Processing 84 (2006), pp. 338–345.

[27] Y. Tsai, *Impact of flow velocity on the dynamic behaviour of biofilm bacteria*, Biofouling 21 (2005), pp. 267–277.

[28] Z. Tsvetanova, *Study of biofilm formation on different pipe materials in a model of drinking water distribution system and its impact on microbiological water quality*, Chemicals as Intentional and Accidental Global Environmental Threats (2006), pp. 463–468.

[29] H. Videla and L. Herrera, *Microbiologically influenced corrosion: looking to the future*, International Microbiology 8 (2005), pp. 169–180.

[30] J. Wingender and F. Hans-Curt, *Biofilms in drinking water and their role as reservoir for pathogens*, Hygiene and Environmental Health 214 (2011), pp. 417–423.

[31] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco (2005).

[32] K.D. Yu, J. and T. Lee, *Microbial diversity in biofilms on water distribution pipes of different materials*, Water Science and Technology 61.

[33] Z.Y. Zhou, L.L. and G. Li, *Effect of pipe material and low level disinfectants on biofilm development in a simulated drinking water distribution system*, Journal of Zhejiang University 10 (2009), pp. 725–731.