Innovation in Medicine and Healthcare 2014
M. Graña et al. (Eds.)
IOS Press, 2014

doi:10.3233/978-1-61499-474-9-92

Handling Varying Amounts of Missing Data when Classifying Mental-Health Risk Levels

Sherine Nagy SALEH [a,1] and Christopher D. BUCKINGHAM [b,2]
[a] *Arab Academy for Science and Technology, Alexandria, Egypt*
[b] *Computer Science, Aston University, Birmingham, UK*

Abstract. One of the main challenges of classifying clinical data is determining how to handle missing features. Most research favours imputing of missing values or neglecting records that include missing data, both of which can degrade accuracy when missing values exceed a certain level. In this research we propose a methodology to handle data sets with a large percentage of missing values and with high variability in which particular data are missing. Feature selection is effected by picking variables sequentially in order of maximum correlation with the dependent variable and minimum correlation with variables already selected. Classification models are generated individually for each test case based on its particular feature set and the matching data values available in the training population.

The method was applied to real patients' anonymous mental-health data where the task was to predict the suicide risk judgement clinicians would give for each patient's data, with eleven possible outcome classes: zero to ten, representing no risk to maximum risk. The results compare favourably with alternative methods and have the advantage of ensuring explanations of risk are based only on the data given, not imputed data. This is important for clinical decision support systems using human expertise for modelling and explaining predictions.

Keywords. missing data, feature selection, risk prediction, mental health, correlation, partial correlation

Introduction

Data mining is a general term for applying methods from statistics and machine learning that find patterns and relationships within data sets [1]. In the health domain, data mining can be used to aid in the process of illness diagnosis, treatment options and prediction of health problems [2,3]. In order to get the best outcome, a number of issues need to be addressed, one being the nature of the data. Medical data, especially those collected from general patient assessments, often have a large number of variables, unbalanced number of samples per class, and missing values [3]. The most effective methods of analysing

[1]Corresponding Author:Computer Engineering, Arab Academy for Science and Technology, Alexandria, Egypt; E-mail: sherine_nagi@aast.edu.
[2]This work was part supported by Grant SRG-0-060-11 awarded to C.D.Buckingham from the American Foundation for Suicide Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Foundation for Suicide Prevention.

them need to be robust with respect to these qualities. They are particularly prevalent in mental health data used to predict risks such as suicide, which is the focus of this paper's research.

The next section will introduce the mental-health domain and the nature of risk data. The paper will then discuss types of missing data and how they influence the most appropriate classification algorithms. Different methods for selecting features from the patient sample are introduced, followed by some classification algorithms using the selected features. Conclusions lead to the new algorithm proposed in this paper. Its results are compared with other common approaches and the paper ends with a discussion of the method's utility for GRiST and other clinical decision support systems.

1. Background

Mental health risk data have high dimensionality and large variability in missing values because they encompass all areas of a person's life, including social, emotional, mental, and physical. Physical health problems tend to be more constrained. Common methods of dealing with missing data are either to impute them, or ignore any records with missing values. Imputation usually involves filling the missing data with values, which could be the mean, mode or a random value from a sample of the data set that has similar features. However, imputation is not always better than models built only on the data that are present in the sample, especially when the data are not missing completely at random [3,4].

A method for reducing the scale of missing data is by preprocessing using feature selection. The aim is to obtain a subset of features that best represent the complete data set. It reduces the dimensionality of the sample by ignoring the features that are redundant or irrelevant, which not only improves classification performance but also diminishes the number of variables that can have missing values. Selection of features is carried out by a subset generation unit, with each subset evaluated and enhanced until a predefined evaluation criterion is met [5,6].

There are two main methodologies for determining useful information from the selected subset of features: supervised and unsupervised learning. In unsupervised learning such as clustering, data instances are grouped together based on similarities without prior knowledge of their natural grouping. In contrast, supervised learning uses the known grouping of data into classes and concentrates on how to assign new, unclassified objects into the most appropriate class [1,7]. This paper uses supervised learning because the risk data has been assigned by clinicians to one of eleven classes.

1.1. The GRiST Mental Health Data Set

The data set on which all the experimentation was done is from the GRiST project [8] for assessing risks associated with mental-health problems, such as suicide, self-harm, harm to others, vulnerability, and self-neglect. GRiST is a clinical decision support system based on a psychological model of classification [9] for representing mental-health expertise. The expertise was elicited and implemented from mental-health practitioners by interviews and focus groups, and then refined through feedback from using the model in practice [10,11,12]. GRiST is used by many mental-health organisations within the

UK and currently (May, 2014) has 500,000 completed individual risk assessments from about 60,000 patients collected by almost 3,000 clinicians. The amount of data within a particular assessment varies depending on where in the care pathway the assessment is being conducted (initial screening, full assessment, repeat assessment, etc), the amount of time a clinician may have, and the perceived level of risk attributed to the patient, which affects whether or not the assessor collects additional data. All these factors create a heterogeneous patient population that is difficult for mathematical classifiers to handle consistently. Furthermore, the representation of expertise within GRiST is a deliberate design choice to ensure advice can be explained using knowledge and reasoning that resonates with the mental-health clinician's own cognition. This means it cannot be based on imputed data that the clinicians did not actually provide. Neither can it use techniques for dimensionality reduction that extract new features [13] not recognised as part of their mental-health knowledge structures.

This paper proposes an algorithm to select the most relevant data for each patient separately, as opposed to finding a singe subset to be applied across the whole data set, as is commonly done by feature selection algorithms. The optimal feature subset for a patient is extracted and then the classification rule for this subset is learned by only using samples in the database where all matching features have values. Hence, neither the training data nor the patient vector have any missing data, which means the classification prediction can be explained only in terms of information the assessor actually provided.

1.2. Missing Data

Medical data sets have large proportions of missing values in their records. The cause may vary from data entry mistakes to information that patients neglect to share or assessors do not request. The reasons for the missing data are important because they impact on the degree of bias caused by ignoring them. Data Missing At Random (MAR) are not linked by relationships or patterns between items in the data set. Instead, they could be MAR due to external influences, which might only affect certain variables but not in any systematic way. Data that are Not Missing At Random (NMAR) have a relationship between the missing items and the sample set, such as assessors skipping a certain question for a particular group of patients because of the consequences of the answers. Data Missing Completely at Random (MCAR) are those where there is no relationship between the data items or the output classes [14,15,16,17].

One of the most common ways to deal with missing data is by discarding the records that include them. This concept may be applicable if a small percentage of records has missing data and their removal from the data set would not affect the diversity of information in the remaining sample, especially if the data are MCAR or MAR. On the other hand this methodology could lead to a bias in the resulting classification if the data set is of small size. An alternative method is imputation where missing data is added by, for example, the hot deck imputation method which involves randomly choosing a value of the missing feature from the set of samples which are closest to the sample with the missing value [1,2,4,18].

More complicated methods include multiple imputations, resulting in several data sets with alternative imputed values for the missing variables. Each set is passed to the classifier and the results are merged to produce a consensual classification. Another example is the K nearest neighbor algorithm (KNN) where a subset of the data set is cho-

sen according to a distance measure to the sample with missing data. The replacement value is provided according to a predefined criterion, sometimes simply the mode for discrete values and mean for continuous, and other times by applying weights according to their distances. These methods are computationally more expensive and their effect is dependent on the problem structure[14].

Choosing the treatment methodology for missing data is dependent on the problem and how this problem is affected by the advantages and disadvantages of the selected methodology. Some data sets may work well with the deletion of records at the price of losing information or by using imputation at the price of increasing the computational cost [4,14].

The mental health research addressed here is built on a data set that has a large percentage of missing values in varying amounts per sample. Applying list-wise deletion on the whole data set would have been impossible since there are no complete records available. Imputation would likewise be impractical because of the computational cost and the degradation of accuracy in imputing such a large percentage of missing records. This paper takes the opposite approach by working only with data that are present in the sample to be classified and are also matched by samples in the training set. Although some records would again be ignored, there are fewer of them and the chosen ones are similar to the patient because they match on the same group of questions, which should help reduce the bias.

1.3. General Feature Selection Approaches

Reducing the number of features used in classification necessarily reduces the amount of missing data. Feature selection is an important preprocessing task because it needs to produce a subset that best represents the whole data set. Reducing the features by removing redundant and irrelevant ones also helps the performance of classifiers. A practical advantage of importance to GRiST, where the learning and classification is effected in real time, is that the models can be built more quickly. The following subsections provide a brief description of how feature selection can be performed.

1.3.1. Subset Generation

A subset of features is chosen according to two criteria. First, whether the feature selection should start with an empty set and have features gradually added to it (forward selection) or whether they should start with all the features and have ones gradually removed (backward selection). A combined approach could be used where it starts with either method and features are added or removed according to some bidirectional measure.

Secondly, a search strategy for choosing the best sample set must be implemented, which can be either through complete or sequential or random search. A complete search aims at producing the optimal subset. A sequential search adds or removes features according to an evaluation criterion until a stopping condition is fulfilled. Finally a random search starts with a random subset of features and continues in the same manner as the sequential search or keeps producing other random subsets until the best subset that fulfills the stopping condition is met [5,6].

1.3.2. Subset Evaluation

The evaluation of subsets is done depending on whether a wrapper or filter model is used. The difference is that the wrapper uses dependent criteria, where it evaluates the model based on the performance of a classifier to see whether the selected features suit the chosen classification algorithm or not. The filter model uses independent criteria that evaluate the subset without knowing which classifier will be used for classification. Independent criteria include correlation measures between features and classes or information measures to calculate the information gained from the addition (or removal) of a feature. A hybrid model that combines both the filter and the wrapper methods can be used with a mining algorithm and independent measures to evaluate the subsets [5,19]. Filter methods are most commonly used, are simple, do not rely on a prediction algorithm and allow the classification to be performed on a real-time basis. They are most suited to the mental-health data and its application for the GRiST decision support system so the work presented here will be based on it.

1.3.3. Stopping Criteria and Result Validation

Subset evaluation is followed by specifying the stopping condition defining acceptable performance. It could be that all the subsets are evaluated, or the evaluation could stop when addition or deletion of new subsets fails to make any improvement. Or there could be an acceptable error threshold and the first subset that meets it is chosen. Finally, the selected features could be compare to ones already determined by experts or by comparing results produced by the mining algorithm on the selected features and on the whole set of features [5].

1.4. Correlation-Based Feature Selection

Correlation-based Feature Selection (CFS) is a filter-based method that uses correlation to select the subset of features. The CFS applies a forward selection method and the subset evaluation applies the correlation measure.

First, two matrices are created, feature-feature correlation and feature-class correlation, from the training data set. The idea is to use these matrices to choose features that maximise correlation with the class vector and minimise correlation between each other [20,21]. Assuming X and Y are two continuous features, in order to get the correlation between them, Equation 1 is applied where n is the number of samples and σ_X and σ_Y are the standard deviation of features X and Y respectively.

$$r_{XY} = \frac{\sum xy}{n\sigma_X \sigma_Y}, \tag{1}$$

In this paper, the proposed feature selection methodology is an upgrade to CFS by using the partial correlation equation. Partial correlation is a calculation that shows how much a certain variable X correlates with another variable Y after the removal of a set of other influencing variables, Z [22]. Equation 2 shows the calculation of the partial correlation where it can be seen that the correlation of X and Y are calculated and the influence of Z is subtracted.

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2} - \sqrt{1 - r_{YZ}^2}}, \tag{2}$$

The use of correlation guarantees choosing features that are highly related to the output vector of classes and partial correlation would form the constraint that any feature added to this subset would still be highly correlated with the output vector but least correlated with the features already chosen. This would reduce the redundancy between the selected features and thus increase the value of the chosen subset of features as a representation of the dataset. This Dynamic Feature Selection and Classification (DFSC) algorithm is introduced in the next section followed by the experimental results.

2. Proposed Dynamic Feature Selection and Classification (DFSC) Algorithm

In this section, the novel DFSC methodology for handling missing features is proposed that employs list-wise deletion but with a new concept. First a matrix is created with only those features that are present in the vector to be classified and then a subset is chosen from this matrix for the classification algorithm. Some traditional methods of handling missing data add new information to the sample but this may lead to increased bias and also redundant information [4]. The idea behind our methodology is to address a patient record by considering only those questions the assessor did answer and not try to predict or assume any questions that were not recorded.

Part of the motivation is that the method will be used to explain risk advice within the context of a psychological model of classification. The reasons why a patient has been predicted to be high risk, for example, must be explained using concepts that resonate with the assessor and, most importantly, only use data that were provided by the assessor. It would be a step too far for assessors to trust the advice if they could see it used "fabricated" data, irrespective of the reasons for adding it to the classification algorithm or the degree of data authenticity.

Given the variety of data subsets collected for patients, it would be difficult to find one that was common to them all. For example, the suicide risk data has 138 separate potential variables but only about half are answered on average and rarely if ever the same half. Hence, it was decided to treat patients independently, such that each record would be classified according to a different set of features that are selected only from the questions associated with that one patient. This maximises the choice of variables for a patient but requires the construction of individual classification models for each one. If this dynamic feature selection and classification model learning is to be achieved in real time, the algorithm needs to be tractable and fast.

The basic idea presented in Algorithm 1 is to start with a training set and calculate the correlation matrix for all patient features with the output class vector. This will be used for test patients to select their features. The method of selecting the subset of features is an essential step in the algorithm. For each patient, it uses forward selection from the available features, starting with the feature that has the highest correlation with the class vector. Features are then added to this set by choosing the one that is next in order of correlation with the class and has least correlation with those already selected in S. The correlation between feature S and the class vector is calculated only using the rows where both values exist. In order to add more features to the subset, the partial correlation coefficient is used to ensure that each time a new feature is added, it attempts to maximise correlation with the class vector and minimise correlation with the chosen features S. The objective is to add features that continue to explain the dependent variable's vari-

Algorithm 1 Dynamic Feature Selection and Classification (DFSC) without involving missing data

```
Given a training set, T, of all patient vectors and their associated classes
Calculate the feature-class correlation matrix
for i = 0 to TestPatients.count do
    Starting with empty set, S, of selected patient features
    Starting with complete training set, T, of all features
    Choose the feature in i with the highest value in the feature-class matrix
    Add chosen feature to S
    for j = 0 to RequiredFeat.count do
        Choose feature correlating most with class vector and least with S
        Add chosen feature to S
    end for
    Remove all vectors from T that have missing values for features in S.
    Learn classification model using S and T
    Classify the test patient record
end for
```

ance but without unnecessarily adding redundancy to the sample set of features. Reducing redundancy is important for the performance of the chosen logistic regression [23] classification method as well as speeding up model construction by keeping the vectors as lean as possible.

After the feature set is selected, a logistic regression classifier is created for the patient using only those members of the training set that have answers for all features in the subset, *S*. Logistic regression is used partly because of its simplicity and effectiveness for this domain but mostly because the results can be output in a format commensurate with the GRiST model of expertise. The regression weights, values, and predictions can be presented within the classification model so that the assessor can understand the rationale for a certain risk class prediction.

3. Experimental Results

In the presented experiments, the classifiers and feature selection methods used for comparing models are implemented via the WEKA software package [24]. The implementation of the proposed DFSC algorithm was done in MATLAB but with WEKA also used to compare the results with alternative classification algorithms.

3.1. Data Set

The new algorithm is implemented on GRiST data for patients with suicide risk evaluations [8]. The data were collected by clinicians using the GRiST clinical decision support system as part of their normal practice. All risk data are automatically anonymous as part of the collection and data storage implementations and ethics approval was obtained for analysing the results.

The clinicians collect patient data and provide their own judgement of the suicide risk, giving a value between 0 and 10, i.e one of 11 classes. This sample GRiST data

set is very challenging because the percentage of missing data is 59% and none of the patient records are complete. Neither are the risk classes equally distributed, with Table 1 showing the number of patient records available per class (GRiST measures risk between 0 and 1, which is why the judgement classes have been divided by 10). The number of features in the data set are 138, including the risk class variable, and the total number of patient records are 31,942. The data were collected from almost 3000 different physicians working in various organizations, with different training regimes and patient population characteristics. This heterogeneity is testimony to the difficulty in building a single model that can accurately represent classification behaviour across all patients and clinicians.

Risk	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Risk Count	3050	7221	7873	6087	2813	2189	972	922	557	199	59

Table 1. Number of Patients in Each Risk Class

3.2. Results Comparison

First the results produced by classifying the whole data set with no feature selection will be discussed, followed by the results when selecting features based on the correlation feature selection algorithm explained in Section 1.4. Finally the results of the new DFSC algorithm are presented. All experiments were implemented using 10-fold cross validation and classified using logistic regression.

The results are measured as the accuracy percentage calculated on two levels. One is the 100% correct classification, meaning the patient's risk is classified into the same risk category as that given by the clinician. The other is a 1-class shift tolerance, which allows the classification to be considered correct if it is only 1 class away from the clinical judgement, meaning that a 0.1 risk patient could be classified as 0 or 0.1 or 0.2 risk. This tolerance is clinically legitimate because a granularity of 11 classes is higher than the semantic categories used to distinguish patient risk, which are not more than five (none, low, medium, high, max, for example). Many risk tools only have three and sometimes just a binary low/high.

In order to compare approaches, missing data must first be addressed. List-wise deletion would not be suitable since all the patient records include missing data. A more practical solution is imputation. In this experiment the missing data were replaced by the mean of the values from the data set and then applied to the logistic regression classifiers. The results produced are shown in Table 2. It provides the baseline for comparing no feature selection and data imputation for all missing data with feature selection and no missing data imputation.

	Logistic Regression
100% Correct	32.44%
1 class shift	76.16%

Table 2. Results of applying logistic regression on the whole dataset

Table 3 shows the results for feature selection without missing data imputation using the CFS algorithm that was discussed in Section1.4 and the new DFSC algorithm introduced in this paper. As for the full data results in Table 2, logistic regression was used for

classification. Comparing the results in Table 2 with the CFS column in Table 3, it can be seen that feature selection did improve the results by around 3% in both the 100% correct and the 1-shift mode. The CFS algorithm produced its best results when using about nine or ten features. As one might expect, the results of the DFSC algorithm produced highest accuracy for fewer features, with between four and seven being optimal. Furthermore, these fewer features produced nearly a 3% improvement in the 100% correct column and almost 5% improvement in the 1 class shift.

These two results support the rationale used for the proposed method because reducing redundancy helps logistic regression performance as well as requiring fewer variables to explain the same amount of variance in the dependent variable. Even with only two or three features, the proposed algorithm has shown better results than the CFS algorithm. Clearly the best two or three features for a particular patient are more representative of that patient than the best two or three across all the patients taken together.

	CFS		DFSC	
Feat Count	100% Correct	1 Class Shift	100% Correct	1 Class Shift
2	28.53%	70.26%	36.49%	82.15%
3	30.40%	72.54%	36.98%	83.01%
4	33.56%	77.18%	37.51%	84.04%
5	34.68%	78.67%	38.07%	83.33%
6	34.81%	78.90%	37.77%	82.41%
7	34.98%	79.31%	36.72%	81.01%
8	35.00%	79.11%	35.63%	79.38%
9	35.07%	79.33%	34.80%	77.68%
10	35.18%	79.17%	33.99%	76.42%

Table 3. Results of proposed DFSC methodology

4. Conclusion

In this research, classification of mental-health risk data was conducted without imputing any missing data and by minimising the list-wise deletion for training the classification model. The proposed DFSC algorithm achieved this by calculating separate classification models for each patient based on the particular set of features they possessed. Features were selected by maximising correlation with the class and minimising redundancy. The results compared well with more standard approaches and with fewer variables in the classification model.

The obvious caveat is that better imputation and more sophisticated classification algorithms would have improved the competition. Nevertheless, this algorithm demonstrates the feasibility of generating individual logistic regression models tailored specifically for each patient's context. For the GRiST decision support system, predictions need to be generated in real time, which means calculating a patient's regression model as quickly as possible. The proposed DFSC facilitates this by reducing the number of variables required for optimal performance. Overall, the approach has important implications for providing decision support in problem domains with high dimensionality but sparsely populated data sets.

References

[1] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, 2005.

[2] R. Bellazzi and B. Zupan, "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines," *International journal of medical informatics*, vol. 77, pp. 81–97, Feb. 2008.

[3] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature," *Journal of medical systems*, vol. 36, pp. 2431–48, Aug. 2012.

[4] I. Myrtveit, E. Stensrud, and U. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999–1013, 2001.

[5] H. Liu and L. Yu, "Towards Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[6] K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data Mining: A Knowledge Discovery Approach*. Springer, 2007.

[7] E. Berner, *Clinical Decision Support Systems: Theory and Practice*. Springer, second edi ed., 2006.

[8] GRiST, "Galatean risk and safety tool." www.egrist.org, 2014. accessed April 4th.

[9] C. D. Buckingham, "Psychological cue use and implications for a clinical decision support system," *Medical Informatics and the Internet in Medicine*, vol. 27, no. 4, pp. 237–251, 2002.

[10] C. D. Buckingham, A. E. Adams, and C. Mace, "Cues and knowledge structures used by mental-health professionals when making risk assessments," *Journal of Mental Health*, vol. 17, no. 3, pp. 299–314, 2008.

[11] C. D. Buckingham, A. Ahmed, and A. E. Adams, "Using XML and XSLT for flexible elicitation of mental-health risk knowledge," *Medical Informatics and the Internet in Medicine*, vol. 32, no. 1, pp. 65–81, 2007.

[12] C. D. Buckingham, A. Ahmed, and A. Adams, "Designing multiple user perspectives and functionality for clinical decision support systems," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS)*, (Krakow), FedCSIS, 2013.

[13] Y. Saeys, I. Inza, and P. Larrnaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[14] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern Classification with Missing Data: A Review," *Neural Computing and Applications*, vol. 19, pp. 263–282, Sept. 2009.

[15] Y. Ding, "An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data," *The Journal of Machine Learning Research*, vol. 11, pp. 131–170, 2010.

[16] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, 2 ed., Sept. 2002.

[17] R. Little, "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association*, vol. Vol. 83, no. No. 404, pp. 1198–1202, 1988.

[18] M. Ghannad-Rezaie, H. Soltanian-Zadeh, and H. Ying, "Selection-fusion Approach for Classification of Datasets with Missing Values," *Pattern Recognition*, vol. 43, pp. 2340–2350, 2010.

[19] G. Doquire and M. Verleysen, "Feature Selection with Missing Data using Mutual Information Estimators," *Neurocomputing*, vol. 90, pp. 3–11, Aug. 2012.

[20] A. Mark, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," in *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, (San Francisco), pp. 359–366, Morgan Kaufmann Publishers Inc., 2000.

[21] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," in *23rd International Symposium on Computer and Information Sciences*, pp. 1–4, Oct. 2008.

[22] B. H. Cohen, *Explaining Psychological Statistics*. New Jersey: John Wiley & Sons, 2008.

[23] D. W. Hosmer, S. Lemshow, and R. X. Strudivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.

[24] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington: Morgan Kaufmann, 3rd ed., 2011.