Volume 2, Issue 11, November 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Prediction of Higher Education Admissibility using Classification Algorithms

Dr. A. Padmapriya*

Department of Computer Science & Engineering,
Alagappa University, Karaikudi, Tamil Nadu, INDIA

Abstract* — This paper proposes to apply data mining techniques to predict higher education admissibility. This paper presents the results from data mining research, performed at one of the famous and prestigious Government Arts and Science Colleges in Tamil Nadu, with the main goal to predict the higher education admissibility of women students. In this research, real data about 690 under-graduate students from Government arts college (W), Pudukkottai were used. The research is focused on the development of data mining models for predicting the students likely to go for higher studies, based on their personal, precollege and graduate-performance characteristics. The dataset used for the research purpose includes data about students admitted to the college in three consecutive years. Several well known data mining classification algorithms, including a decision tree classifier and Naive Bayesian classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.

***Keywords*— Educational Data Mining, Classification, Prediction, Decision Tree, Naive Bayesian Classifier**

I. INTRODUCTION

Recent years have shown a growing interest and concern in many countries about the problem of post graduation admissibility and the determination of its main contributing factors. This problem is known as the "the one hundred factors problem" and a great deal of research has been done on identifying the factors that affect the enrolment of post graduate students. Data mining [19], generally defined as the process of discovering meaningful patterns in large quantities of data, offers a great variety of techniques, methods and tools for thorough analysis of available data in various fields. The implementation of data mining in the educational sector, recently defined as "educational data mining" (EDM) [1], is a new stream in the data mining research field.

The educational data mining research community is constantly growing, starting by organizing workshops since 2004, then conducting an annual International Conference on EDM beginning since 2008, and now already having a Journal on EDM (the first issue being published in October 2009). There are already a large number of research papers discussing various problems within the higher education sector and providing examples for successful solutions reached by using data mining. Extensive literature reviews of the EDM research field are provided by Romero and Ventura in 2007 [1], covering the research efforts in the area between 1995 and 2005, and by Baker and Yacef in 2009 [2], for the period after 2005. The problems that are most often attracting the attention of researchers and becoming the reasons for initiating data mining projects at higher education institutions are focused mainly on retention of students (by better knowing their peculiarities and needs, and by providing proper support in advance), more effective targeted marketing, improving institutional efficiency, and alumni management.

The performed research work, presented in this paper, focuses on the development of data mining models for predicting student higher education admissibility by using two data mining algorithms [18] for classification – a Decision tree algorithm and a Naive Bayesian Classifier. Classification is a technique of data mining in which a model or classifier is constructed to predict the class or categorical labels of given data. Data classification is two step process. In first step classifier is created by analyzing training data. This step is known as learning step of classification. Then in second step using classifier made in first step, prediction of given data tuples are done.

The rest of the paper is organized in four sections. The research motivation and the state-of-the-art are presented in Section II. The adopted methodological approach, the experimentation data selection and pre-processing are described in Section III. The obtained results from the application of the selected data mining algorithms are presented in Section IV. The paper concludes with a summary of the achievements and discussion of further research.

II. RESEARCH MOTIVATION

Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers [3]. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes) [4]. The aim of this paper is to determine the factors estimating the higher education of women students undergoing their education in rural area.

The literature review reveals that these problems have been of interest for various researchers during the last few years. The development of data mining models for predicting student performance at various levels, and comparison of those models, are discussed in a number of research papers. In 2000 the results of a study are described [4] aimed at finding weak students and involving them in additional courses for advanced support by extracting association rules from data. The retention of students is a problem discussed also by Luan, who implemented clustering, neural network and decisions tree methods to predict the students in risk of failure [5], [6]. Data mining methods are implemented for modeling online student grades [7], using three classification approaches used (binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). Kotsiantis et al. [8] also deal with predicting student performance, recognizing dropout-prone students based on demographic characteristics (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment). Pardos et al. [9] use data from an online tutoring system for teaching Math and implement a regression approach for predicting the math test score based on individual skills.

Superby et al. [10] predict students at risk of drop-out, determining factors influencing the achievement of the first-year university students, classifying students into three classes – low-risk, medium-risk and high-risk, using Decision trees, Random forest method, Neural networks and Linear discriminant analysis. Vandamme et al. [11] also deal with early identification of three categories of students: low, medium and high-risk students using Decision trees, Neural networks and Linear discriminant analysis. Cortez and Silva in [12] attempt to predict student failure by applying and comparing four data mining algorithms, Decision Tree, Random Forest, Neural Network and Support Vector Machine. The implementation of predictive modelling for maximizing student recruitment and retention is presented in the study of Noel-Levitz [13]. The development of enrolment prediction models based on student admissions data by applying different data mining methods (Decision trees, Rule induction, Feature subset selection) is the research focus of Nandeshwar [14]. Dekker et al. [15] focus on predicting students drop out. Kovacic in [16] uses data mining techniques (feature selection and classification trees) to explore the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block) that may influence persistence or dropout of students, identifying the most important factors for student success and developing a profile of the typical successful and unsuccessful students. Ramaswami et al. in [17] focus on developing predictive data mining model to identify the slow learners and study the influence of the dominant factors on their academic performance, using the popular CHAID decision tree algorithm.

In this research work, the samples are drawn from the Government Arts College for Women, Pudukkottai. A model is built describing a predetermined set of data classes or concepts. This model is called as classification model and is then used to classify future data for which the class it belong to is unknown. Here the knowledge acquired in classification model can be used to predict whether a student will go for higher studies or not. Moreover the accuracy, size and speed of the two classification techniques are compared.

III. CLASSIFICATION ALGORITHMS AND DATA SETS

Classification [4] is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of “training data” – data objects whose class label is known. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. The learned model is represented in the form of classification rules, decision tree or mathematical formulae.

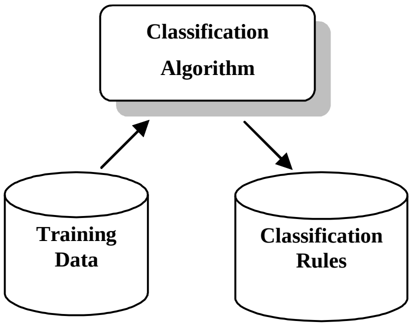

Figure. 1. The data classification process – Learning : Training data are analysed by a classification algorithm

In the second step, the model is used for classification. First the predictive accuracy of the model is estimated. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known. The basic techniques for data classification are decision tree induction and Bayesian classification.

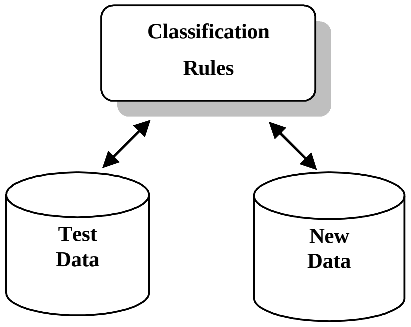

Figure. 2. The data classification process – Classification: Test data are used to estimate the accuracy of the classification rules

A. Classification by Decision Tree Induction

A decision tree is a flow-chart-like structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf nodes represent classes or class distribution [19]. The top-most node in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees [18] can easily be converted to classification rules.

The basic strategy is as follows

- The tree starts as a single node representing the training samples.
- If the samples are all of the same class, then the mode becomes a leaf and is labelled with the class.
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the "test" or "decision" attribute at the node. In this version of the algorithm, all attributes are categorical, that is, discrete-valued. Continuous-valued attributes must be discretized.
- A branch is created for each known value of the test attribute, and the samples are partitioned accordingly.
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents.
- The recursive partitioning stops only when any one of the following conditions is true:

  i. All samples for a given node belong to the same class, or

  ii. There are no remaining attributes on which the samples may be further partitioned. In this case majority voting is employed. This involves converting the given node into a leaf and labelling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored.

  iii. There are no samples for the branch test-attribute = a_i. In this case, a leaf is created with the majority class in samples.

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. Such an information-theoretic approach minimizes the expected number of tests needs to classify an object and guarantees that a simple tree is found.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i=1, ….., m). Let s_i be the number of samples in S in class C_i. The expected information needed to classify a given sample is given by

$$I\ (s_1,s_2,\ldots..,s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i / s.

Let attribute A have v distinct values, $\{a_1,a_2,\ldots,a_n\}$. Attribute A can be used to partition S into v subsets, $\{S_1,S_2,\ldots,S_v\}$, where S_j contains those samples in S that have value a_i of A. If A were selected as the test attribute then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{ij} + \ldots.. + s_{mj}}{s} I(s_{ij} + \ldots. + s_{mj})$$

The term (s_{ij} + ... + s_{mj}) / s acts as the weight of the j^{th} subset and is the number of samples in the subset divided by the total number of samples in S. The smaller the entropy value, the greater will be the purity of the subset partitions. Note that for a given subset S_j,

$$I(s_1,s_2,\ldots,s_m) = \sum_{i=1}^{m} p_{ij} \quad \mathrm{Log}_2(p_{ij})$$

where p_{ij} = s_{ij} / S_j and is the probability that a sample is S_j belongs to class C_j.

The encoding information that would be gained by branching on A is

$$\mathbf{Gain(A) = I\ (s_1,s_2,\ldots..,s_m) – E(A)}$$

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labelled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

B. Naive Bayesian Classifier

Bayesian classifiers are statistical classifiers [10]. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayes Theorem. Naive Bayesian Classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and in this sense is considered naive. The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1) Each data sample is represented by an n-dimensional feature vector, X = ($x_1,x_2,\ldots,x_n$), dependency n measurements made on the sample from n attributes, respectively, $A_1,A_2,\ldots,A_n$.
2) Suppose that there are m classes, $C_1,C_2,\ldots.,C_m$. Given an unknown data sample X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$\mathbf{P(C_i|X) > P(C_j|X)} \text{ for } 1 <= j <= m, \quad j \text{ not equal to } i.$$

Thus we maximize $\mathbf{P(C_i|X)}$. The class C_i for which is $\mathbf{P(C_i|X)}$ maximized is called the maximum posteriori hypothesis. By Bayes theorem

$$\mathbf{P(C_i\,|X) = \frac{P(X|C_i)P(C_i)}{P(X)}}$$

(1) As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class priori probabilities are not known, then it is commonly assumed that the classes are equally likely, that is $P(C_1) = P(C_2) = \ldots. = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class priori probabilities may be estimated by $P(C_i) = s_i / s$, where s_i is the number of training samples of class C_i, and s is the total number of training samples.

(2) Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, that is, there are no dependence relationships among the attributes.

Thus,

$$\mathbf{P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)}$$

The probabilities $P(x_1|C_i)$, $P(x_2|C_i), \ldots,$ $P(x_n|C_i)$ can be estimated from the training samples, where

$$\mathbf{P(x_k|C_i) = s_{ik} / s_i}$$

Where s_{ik} is the number of training samples of class C_i having the value x_k for A_k, and s_i is the number of training samples belong to C_i. (A_k is categorical)

(3) In order to classify an unknown sample X, $P(X|C_i)P(C_i)$ is evaluated for each of the class C_i. Sample X is then assigned to the class C_i if and only if

$\mathbf{P(C_i|X) > P(C_j|X)}$ for 1 <= j <= m, j not equal to i.

C. Selection of samples and their characteristics

The samples for the classification are drawn from Government Arts College for Women, Pudukkottai. In general the data is randomly partitioned into two independent sets, a training set and a test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set.

Here the training samples that are analysed to build the model collectively with the details of the students currently are doing their post graduation in the college. The classification rules learned from the analysis of data from existing post graduates can be used to predict whether a student will join post graduate (PG) course or not. The students doing under graduation form the test samples. For efficient and useful classification the test samples are drawn particularly from final year under graduate (UG) students. The open source software WEKA, offering a wide range of classification methods for data mining [19], is used as a data mining tool for the research implementation.

IV. EXPERIMENTAL RESULTS

The research work has carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting higher education admissibility of students studying in Government arts and science colleges. The data set used for the research work is pre-processed in order to transform them into a suitable format to be used by the prediction tool. First, a new summary table (see Table 1) is created, which integrates the most important information for the objective. Actually the data set has 27 attributes before pre-processing the real data set. The influencing attributes alone extracted after pre-processing.

TABLE I
ATTRIBUTES USED FOR EACH STUDENT

Type of Data	**Attribute Name**
Personal Data	Annual Income Parents educational qualification Age No. of Children in the Family Native place
Pre college data	Place of School Education Profile School Education Score of School Education
Under Graduation Data	Branch Name Percentage of marks Interest for higher studies

To do all the experiments, two classification algorithms have been used that are available in the well-known Weka DM software (Witten et al., 2011): Decision tree algorithm and Naive Bayesian Classifier algorithm. These algorithms have been selected because they are considered as “white box” classification model, that is, they provide an explanation for the classification result and can be used directly for decision making.

Classification and prediction methods can be compared and evaluated according to the following criteria.

- Classification Accuracy – It is the percentage of total prediction that was correct.
- Misclassification Rate – It is the other side of accuracy – it simply measures the percentage of predictions that were wrong.
- Speed – This refers to the computation costs involved in generated and using the model.
- Size – It refers to the size of the code and the database.

The Classification accuracy and the misclassification rate of the two classification algorithms are compared and the comparison chart is as shown in figure 3.

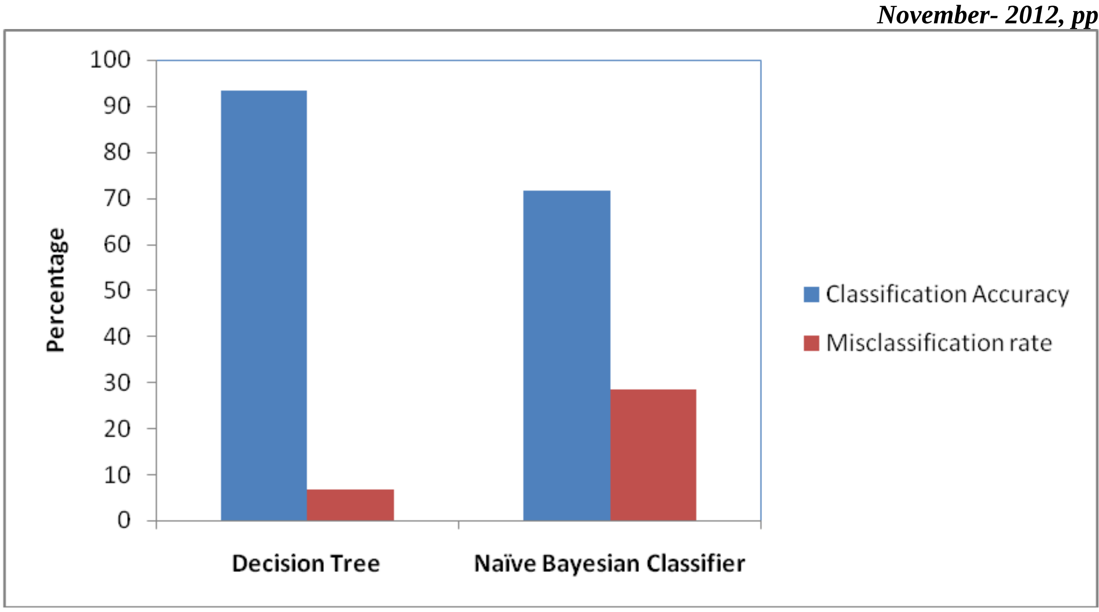

Figure. 3. Comparison of the Classification algorithms

Factors to be noted while comparing the classification methods are as follows.

- Classification Accuracy – The accuracy of Decision Tree Induction is best when compared to that Naive Bayesian Classifier. The accuracy of Decision Tree Induction is 93.33 % and Naive Bayesian Classifier is 71.67 %.
- Misclassification Rate – The error rate of the Decision Tree Induction is low when compared to Naive Bayesian Classifier. The error rate of Decision Tree Induction is 6.67 % and Naive Bayesian Classifier is 28.33 %.
- Speed - Naive Bayesian Classifier is faster than decision tree induction. The response time of Naive Bayesian Classifier algorithm ranges from sub second to 1 second where as the Decision Tree Induction is 1 – 2 seconds.
- Size - Naive Bayesian Classifier occupies less space than decision tree induction. The size of Naive Bayesian Classifier algorithm is 22 KB where as the size of Decision Tree Induction is 28 KB.

V. Conclusions

The classification models, generated by applying the selected two data mining algorithms – Decision tree algorithm and Naive Bayesian Classifier algorithm, on the available and carefully pre-processed student data, reveal classification accuracy between 93.33 % and 71.67 %. The highest accuracy is achieved for the Decision Tree model (93.33%). The Decision Tree model predicts with higher accuracy the “Strong” class, while the other three models perform better for the “Weak” class. The data attributes related to the students’ personal data and under-graduation data are among the factors influencing most the classification process. Finally, as the next step in the research, is the development of a new classification algorithm using grammar-based genetic programming and cost sensitive classification for comparison versus other classification algorithms.

References

[1] Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications 33, 2007, pp.135-146.

[2] Baker, R., Yacef, K. (2009). The State of Educational Data mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, Vol.1, Issue 1, Oct. 2009, pp.3-17.

[3] Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Predicting Performance. Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.

[4] Ma, Y., Liu, B., Wong, C. K., Yu, P. S., Lee, S. M. (2000). Targeting the right students using data mining. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, pp 457-464.

[5] Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, Special Issue titled Knowledge Management: Building a Competitive Advantage in Higher Education, Vol. 2002, Iss.113, pp.17–36.

[6] Luan, J. (2004). Data Mining Applications in Higher Education. SPSS Executive Report, SPSS Inc.

[7] Minaeli-Bidgoli, B., Kashy, D., Kortemeyer, G., Punch, W. (2003). Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA. 33rd ASEE/IEEE Frontiers in Education Conference, 5-8 Nov 2003, Boulder, CO.

[8] Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2004). Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

[9] Pardos Z., Heffernan N., Anderson B., and Heffernan C. (2006). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. In Proceedings of the Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems (ITS2006), June 26, 2006, Taiwan.

[10] Superby, J. Vandamme, J., Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp37-44.

[11] Vandamme, J., Meskens, N., Superby, J. (2007). Predicting Academic Performance by Data Mining Methods. Education Economics, 15(4), pp405-419.

[12] Cortez, P., Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[13] Noel-Levitz White Paper (2008). Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling. Noel-Levitz, Inc., 2008.

[14] Nandeshwar, A., Chaudhari, S. (2009). Enrollment Prediction Models Using Data Mining. Available at: http://nandeshwar.info/wpcontent/uploads/2008/11/DMWVU_Project.pdf

[15] Dekker, G., Pechenizkiy, M., Vleeshouwers, J. (2009). Predicting Students Drop Out: A Case Study. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp.41-50.

[16] Kovacic, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of Informing Science & IT Education Conference (InSITE) 2010, pp.647-665.

[17] Ramaswami, M., Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No.1, January 2010, pp.10-18.

[18] Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. 2000 SPSS Inc. CRISPWP-0800. Available at: http://www.crisp-dm.org/CRISPWP-0800.pdf

[19] Witten, I., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, Elsevier Inc. 2005.