Web Prediction Framework for College Selection Based on the Hybrid Case Based Reasoning Model and Expert's Knowledge

Bruno Trstenjak[a][1], Dzenana Donko[b]

[a]Dept. of Computer Engineering, Medimurje University of Applied Sciences Cakovec, Croatia
[b]Dept. of Computer Science, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina

Abstract

Higher education today represents the basis of any successful society. Every day we are witnessing an increase in the number of HEI, an increase in the number of students but also an increase in the number of dropouts. This paper presents a new concept of the prediction framework which enables the selection of future college students based on their socio-demographic characteristics. The framework enables college autonomy in creating their own predictive models based on the characteristics of its students. In the prediction process, the framework has the ability of dynamic adjustment according to specific characteristics of each college. The framework is object-oriented and enables the performance of an online prediction process. The proposed framework uses a hybrid Case Based Reasoning (CBR) model and expert's knowledge. The hybrid CBR model has integrated several methods of machine learning: Information Gain, K-means and Case-based reasoning. The study used datasets collected from several colleges, a part of the Croatian Information System for Higher Education (ISVU). The case study demonstrates that our proposed web prediction framework is efficient and capable of providing very good results in the process of prediction. The achieved results provide guidelines for the future development of the prediction framework.

Keywords: hybrid Case Based Reasoning, expert knowledge, prediction, web framework.

1. Introduction

Higher education is one of the foundations of the development of today's society. Higher education worldwide records development and progress. Every day the number of people who decide to study is increasing. Increasing the number of students has become a problem for the society, because there is still a large number of dropouts [18]. This phenomenon has become a big problem for many countries and their educational systems. Various kinds of researches and analyses have been conducted to detect the main reasons of dropping out. The research results have confirmed

[1]* Corresponding author: Bruno Trstenjak, B .J. Jelacica 22a, 40000 Cakovec, Croatia
Tel.: +385 40 396 990 Fax: +385 40 396 980 E-mail address: btrstenjak@mev.hr

that the student motivation is not the only reason that affects the dropout rate [18]. It was found out that various socio-demographic attributes have an impact on the success of the study, influence the level of overcoming a study program [26] . [22] in their research analysed the importance and impact of various features in the success of the study.

Data mining in higher education in the research field is called Education Data Mining (EDM). The importance of education for the society has encouraged EDM to become a new growing research field. [20] conducted a survey on EDM between 1995 and 2005. The conducted research tackled the application of data mining in traditional educational systems, in particular web-based courses, well-known learning content management systems and adaptive and intelligent web-based educational systems. Educational data mining uses many machine learning techniques to discover data association, to perform classification and clustering over data. [24] applied the research using various machine learning techniques to improve graduate students' performance and overcome the problem of low grades of graduate students. Conducted evaluations and obtained results showed how useful data mining can be used in higher education, particularly to improve the performance of graduate students. [23] presented the research in which the classification is used as data mining technique to predict the numbers of enrolled students. They presented a new approach so they can manage and prepare necessary resources for the newly enrolled students.

This paper proposed a new web prediction framework for college selection. In the paper we describe the structure of the proposed framework and functions of its major components. The framework provides future students with a possibility to select the study program at the college where they want to study. The basic conceptual idea in this research was to create a framework that unifies the information of a large number of colleges and the information about alumni and their achievements during their study into a single entity. The framework performs the prediction based on student graduation rates and socio-demographic data. Each college has a full autonomy to create its own prediction model, the possibility to independently determine student characteristics and dataset features. The result of the prediction is a list of colleges and their study programs, whereas student will achieve the best academic success. A web framework is based on the hybrid CBR classification model. The hybrid model merges three machine learning techniques: Information Gain (IG), k-means and Case Based Reasoning (CBR). The framework, except hybrid CBR models, uses the knowledge of experts from the higher education field.

The rest of the paper is organized as follows; some related work to concept of different models used to predict a student's academic success in Section 2. Section 3 presents the structure of framework and the methods used. Section 4 presents the achieved results of the case study and discussion. Finally, the last Section are the conclusions and recommendations for future research.

2. Literature Review

The literature review in this section presents machine learning frameworks used for prediction of a student's academic success.

Nowadays, when the information in high education is increasing, data mining (DM) is a suitable technique to discover new information and knowledge about students. DM provides various methods for data analysis and conducting the prediction process. The ability to predict student performance is very important in educational environments. One of the major challenges in HEI is to improve students' academic performance. With the development of machine learning techniques and hybrid models, different frameworks aiming to support the education management and better decision making, and as well as to optimize business processes were simultaneously developed.

[13] presented system for Higher Educational Data Mining System (HEDMS) is concerned with the developing methods that discover useful knowledge from data that extracted from educational system. The objective of HEDMS is to build a DM system for academic instructor for education system. HEDMS consists of several components; data gathering, preparing data to discover knowledge, data pre-processing, applying data mining techniques and visualization results.

[7] presented an Naive Bayes algorithm (NB) approach to predict graduating cumulative Grade Point Average based on applicant data collected from the surveys conducted during the summer semester at the University of Tuzla. The Naive Bayes algorithm is used to discover the most suited way to predict student's success. The achieved results indicate that the Naïve Bayes classifier outperforms in prediction decision tree, indicated that a good classifier model has to be both accurate and comprehensible for professors.

[9] propose the framework which uses a hybrid model of neural network and decision tree classifier. The framework predicts the likelihood of the university a student may enter by analysing his academic merits, background and the

university admission criteria from historical records. The system was tested with the data from Macau secondary school students, and the suitable universities that match the students' profiles were predicted. [25] presented a framework designed to predict students’ performance in an online course. The framework is used to identify at an early stage of the course those students who have a high risk of failing. The concept of framework is based on the k-Nearest Neighbours machine learning technique. The research was focused on the prediction of students’ performance in a touch typing course.

[2] proposed a framework to predict the performance of the first year bachelor students in a Computer Science course. The prediction is based on socio-demographic data of students. Naïve Bayes classifier was used.

[8] presented a framework for predicting students’ academic performance of the first year bachelor students in a Computer Science course. This study conducts a comparative analysis of three classification techniques: Decision Tree, Naïve Bayes, and Rule Based applied in the prediction model. The model allows the lecturers to take early actions to help and assist the poor and average category students to improve their results.

[21] propose a simple framework called Faculty Support System that enables faculty to analyse their student performance in a course. The concept of the framework is based on association rule mining to identify the factors influencing the result of students. The framework uses C4.5 decision tree algorithm in the process of prediction. The authors conclude that the proposed system enables easy implementation and is suitable for faculties who do not have any knowledge of data mining technique.

[14] proposes a conceptual framework to support Student Relationship Management (SRM) for Thai universities. The framework has implemented 3 internal processes: data preparation, data analysis and predictive modelling. The main used techniques in the framework are Association Rule and Fuzzy Logic and Rough Set. The proposed framework is designed to work in an online environment.

[6] propose an intelligent framework for monitoring individual or group performance during activity and problem based learning tasks. The framework is used to evaluate teaching approaches and methodologies, identify weaknesses and provide more personalized feedback on the learner’s progress. The framework uses a fuzzy LS technique based on extracting fuzzy weighted If-Then rules from student learning activity and performance data. The rules can be used

to highlight associations between activity / engagement characteristics and achieved performance on tasks. Case study demonstrates how the fuzzy LS approach can be used to effectively monitor students' progress.

3. The proposed framework and research methodology

The objective of our study is to develop web prediction framework based on hybrid CBR model for college selection. In this section, we will first explain the work principle of our proposed framework. We will present the structure of the proposed framework and the major components of the framework structure. The second part of the section describes the structure of the hybrid CBR model and used machine learning techniques.

3.1. Overview of the work principle

The framework is designed for two types of users, colleges and future students. The framework enables colleges to create their own prediction model. Students have a possibility to perform the prediction process, to select colleges and study programs which are most suitable to them based on their characteristics. The framework is designed as a web environment that is accessed via the Internet. Fig. 1 illustrates the framework components and their corresponding function in detail. The structure of the frameworks is composed of two main components: (i) Component for creating the hybrid CBR model and (ii) Web application component. The first component enables creating a hybrid prediction model and the other is used in the process of prediction. The framework was developed in an object-oriented environment. Java programming language was used, which enables online dynamic creation of the objects necessary for the functionality of the framework and hybrid CBR model. The framework core is based on our own developed Java libraries.

3.1.1. The component for creating the hybrid CBR model

The component for creating hybrid CBR model is the most complex part of the framework. It is intended for colleges, allowing them to form their predictive models. Each college can autonomously create a prediction model and independently define the features that describe their students.

Creating a predictive model begins with uploading a dataset in the framework. The dataset contains information about students, their socio-demographic data, ratings, achieved ECTS (European Credit Transfer and Accumulation System) credits the student achieved at the end of the academic year. Each piece of data about a student is a feature that is used in the prediction. The framework in the process of creating a predictive model uses four internal modules. All modules in Fig. 1 are marked with Roman numbers:

I. Web wizard,

II. Data pre-processing,

III. Data clustering and ranking features,

IV. CBR classification generator.

After receiving the input file, the framework starts the wizard in the form of a website that guides the user in the application process. The wizard, after checking the syntactic correctness of the data, requests from the user additional explanations related to the properties of the features. To ensure forming a high quality predictive model, it is necessary to define certain parameters that characterize students. During the application process two properties should be determined for each feature. The first property is the feature data type (numeric, categorical, ordinal). The second property is the feature's importance, its impact on prediction results. This step is very important for the correct creation of a hybrid classification model. In this step the college can change the weight value of each feature. From these weight values, the framework will create a domain driven model (DDM) for each study. DDM is a knowledge base of experts adapted to the structure of framework [3], their opinions about the importance and influence of a certain feature which describes the students characteristic on their performance in the study. In the last step, the wizard requests from the user to specify label description for each feature. The label descriptions framework is used during the creation of web forms, which are used by the main web application interface for online prediction.

Data pre-processing is a process of transforming the raw data into a suitable format ready to be used by a data mining process. Pre-processing includes various activities such as: data cleaning, normalization, transformation, feature extraction and selection, etc. [11]. The module adjusts the source data from the dataset to the syntax supported by the framework. This is an extremely important internal process that prevents incorrect generation of internal files, required for the formation of a hybrid model.

The third and the fourth module belong to the process of forming a hybrid CBR classifier. In our hybrid model three machine learning techniques were implemented: (i) technique for features ranking and determining a weight value to all features, (ii) clustering technique and (iii) classification technique. The third module performed ranking features and clustering dataset instances. For ranking of attributes the framework uses Information Gain (IG) technique [1]. With features ranking, each feature is assigned a weight value. The framework uses weight values in the process of classification. Parallel to the features ranking in the framework, clustering of instances is being carried out. Data clustering is performed with the k-means technique [10]. With clustering, instances are grouped according to their properties and thereby additionally enrich the original data set with information.

The last module is the CBR classification generator. The task of the generator is to generate various files required for dynamic creation of the hybrid CBR model. The hybrid CBR model can dynamically change its properties based on the selected college. Case Based Reasoning (CBR) classification technique is used in the framework. In the CBR classification generator four internal processes are carried out. The first internal process is responsible for forming the case instances and their recording in the case database. Case instances formed in such a way use only the CBR classification model which belongs exclusively to the college which started the process of forming a prediction model. In this way, the framework unites different prediction models and gives autonomy to each college. After formation of the case database, the framework begins forming a DDM file with information about the weight features values. One DDM file belongs to each college and its study program. The DDM file contains the knowledge defined by the college expert, the knowledge that indicates the expert opinion about the level of influence the features have on prediction results. In other words, which features (characteristics of students) and their values are important to successfully complete the selected study program. DDM has a role of a correction factor in the process of CBR classification. In our study, the values were determined by surveying experts from various fields of economy who work in our institution. The values are in the range from 0-1. Value 0 determines a feature which is not important for the process of prediction and is not used. That attribute is ignored in the prediction process. Value 1 determines a feature which has a maximum impact on the result of prediction. The feature must come into the process of prediction. Value between 0 and 1 represents the corrective factor in the process of determining a feature weight value.

The framework is designed for prediction in the web environment. To quality assurance of web application that uses this framework, it is necessary to ensure synchronization between the prediction model and public web forms used by future students. The hybrid CBR model is then synchronized with the selected web form. For the purposes of such synchronization, the generator creates a web form file which is adapted to the features properties. A web form has a mechanism for controlling online input data. The last internal process is responsible for creating various internal files required to create the CBR classifier. The files contained information about all the properties that define one CBR classifier. Based on the properties the classifier selects case instances. With this process ends the formation of a hybrid CBR model adjusted to the college dataset characteristics.

3.1.2. Web application component

The second component is intended for future students. The component allows users college selection and performs the prediction process. Fig. 2 illustrates the prediction process. The prediction process begins with selection of the college and study program by a future student. Based on selection, the framework loads a web form which belongs to the college. A web form has been created in the process of creating a hybrid classification model. After data entering, the web form sends a prediction request to the framework and its hybrid classification model. On the basis of the prediction request, the framework performed selection of a primary classifier, the classifier created by the college. After selecting a primary classifier, the framework selects a potential acceptable hybrid CBR models from other colleges. The framework selects hybrid models whose features at least 80% match the features of the primary classifier. In the case when the selected model has a larger number of features, the framework in the prediction process takes as many features as a primary classifier. The framework uses the formed list of colleges to dynamically load the model properties into the CBR classifier. For each selected college the framework carried out prediction, obtained results are sent to the application web form. The results of prediction are values expressed in percentage. The values indicate to what extent the college and its study program are suitable for the student. A higher value indicates a greater likelihood for the future student to successfully complete the study.

3.2. Hybrid CBR model representation

Data mining (DM) classification is one of the methods that can be used in the prediction process. In data classification, the prediction process can use different machine learning techniques. According to classification approach, techniques can be grouped into two groups: single and hybrid classification techniques. The first group includes a standard machine learning techniques such as: Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine, etc. The second approach used in this study performs the classification using a concept of hybrid classifiers. The concept of hybrid classifiers is based on combining two or more machine learning techniques. [15] applied a survey on hybrid classification models. In the survey, a tree way to build hybrid classifiers is emphasized. The first one involves cascading different classifiers which can be organized into several levels. [4] presented a new architecture of cascaded classifiers to handle multi-class tasks. In the hybrid classifier the cascades are broken into some sub-stages; each contains a number of classifiers. The second approach is based on a combination of clustering technique and classification technique. [27] presented such a concept of a hybrid model for prediction of the web service quality. The last type of the hybrid classifier is the result of integrating several machine learning techniques into a single classification model. [28] applied the hybrid model to predict disease, diagnosis diabetes Type-2. A hybrid model named Modified Glowworm Swarm Optimization-Improved Relevance Vector Machine (MGSO-IRVM) is based on Association Rule Mining (ARM). Hybrid MGSO-IRVM classifier is used to generate the association rules based on the frequent item sets, which avoids rule redundancy and conflicts during the rule mining process.

3.2.1. Case Based Reasoning (CBR)

The principle of CBR method is based on solving new problems by observing the similarity with the previously solved problems. CBR method uses a problem-solving approach analogous to the way of problem solving by man when he draws on his experiences. Each CBR system contains an embedded library of the cases that were resolved in the past. This is something like collecting life experiences in the domain of the problem. Each case represents a description of the problem with its associated solution. CBR method with a built-in function of similarities tries to find the most similar case from the library. The retrieved cases from the library are used to suggest a solution. If the proposed solution is not satisfactory, the method tries to revise selected cases and find a new solution. The method adds a new revised case to the cases library and thereby expands the knowledge base. The whole execution cycle of the algorithm can be divided into four main steps: Retrieve, Reuse, Revise, Retain [19].

CBR performs measurement of similarity on the local and global level. Local similarity refers to the measurement of the similarity between pairs of features. Global similarity refers to the comparison of the similarity between all the features that make up the object. Measuring similarity can be shown by the following Eq. (1):

$$Similarity(T,S)=\sum_{i=1}^{n} f\left(T_j,S_i\right)\times w_i, \tag{1}$$

Where

T= target case

S= source case

n= number of features in each case

I= individual feature from 1 to n

f= similarity function for features I in cases T and S

w= importance weighting of feature I

3.2.2. *Information Gain (IG)*

Entropy is a measure of disorderliness of the system. IG method calculates the value of the features information. The value is defined as the amount of information, provided by the feature items for the class. IG uses the following expression for the calculation:

IG (Class, Feature) = H (Class) – H (Class |Feature)

where H is entropy, which is defined by using the following Eq. (2)

$$Entropy(S)=-\sum_{j=1}^{m} p_j \log_2 p_j \tag{2}$$

where p is probability, for which a particular value occurs in the sample space S.

Entropy value ranges from 0 to 1. Value 0 means that all variable instances have the same value, value 1 equals the number of instances of each value. Entropy shows how the attribute values are distributed and indicates the "purity" of features. High entropy indicates the uniform distribution of attributes. Opposite to that, low entropy indicates the distribution which is concentrated around the point [1].

3.2.3. k-means

K-means is one of the simplest unsupervised learning algorithms used for solving clustering problems. Let X={x_i, i=1,..,n} be a set of n dimensional objects, which should be classified into k clusters, C={c_j, j=1,..,k}. The algorithm determines the quality of the clustering calculating square error between the mean of the cluster and the points in the cluster. The goal of the algorithm is to minimize the sum of the squared error over all K clusters. The quality is determined by following the error function, as shown in Eq. (3) [12]:

$$E = \sum_{j=1}^{k} \sum_{x \in C_j} \left| x_i - \mu_j \right|^2 \tag{3}$$

where E is a sum of the squared error of all objects, μ_j indicates the average of cluster C_j. $|x_i-\mu_j|^2$ is a chosen distance measure between data point xi and the centroids value. The algorithm can use different methods to calculate the distance (Euclidean, Manhattan, Minkowski, etc.) [17].

4. Experimental results and discussion

In order to assess quality framework evaluation, several tests were conducted. In order to determine the quality of framework evaluation of prediction accuracy, measuring the impact of irrelevant parameters and evaluation a speed performance of framework were conducted. Fig. 3 shows the framework interface, web form for ~~a~~ student data entering and displaying the obtained results.

4.1. Dataset

In this study the datasets with five different study programs have been used. The study programs belong to the field of social and technical sciences, as shown in Table 1. All datasets are part of the national database of ISVU system. The database contains a large number of different data about students organised into several groups: personal demographic information, demographic information about the parents and the type of residence, information on passed exams, data on high school achievement, information about the current student status, student's social status, student food grants, etc. The system uses more than 40 features to describe one student. Using datasets from the same system provide the ability to compare the features properties which constitute data instances. For training and testing, the data of the students enrolled in the program in the last two academic years 2013- 2014 have been used. As the target data, ECTS credits which the student achieved at the end of the academic year have been used. The colleges were classified in two classes, "suitable" and "unsuitable".

4.2. Evaluation of prediction accuracy

The first part of the case study was focused on measuring the prediction accuracy of the framework. Table 2 shows the achieved framework prediction results, where CBR is prediction accuracy, PR is classification precision and SE is sensitivity achieved with the framework. During the evaluation, the framework achieved a high degree of prediction accuracy and a high degree of sensitivity. The results show equalized values of prediction accuracy in all data sets. These results indicate a good concept of the hybrid CBR model. The framework demonstrates high prediction quality regardless the dataset size and the number of features.

In order to obtain complete details about the characteristics of the hybrid CBR model, additional analysis of framework performances using the ROC curve was performed in the study. ROC (Receiver Operating Characteristic) analysis is usually associated with the classifier and is used as an alternative metric to compare the performance of the classifier [5]. Figure 4 shows the ROC curve of framework prediction.

The next phase of evaluation has offered an answer to the question of how much the framework is sensitive to dataset size and to irrelevant features. In the framework development strategy it was decided for colleges to enable autonomy in defining the features that describe students. In the selection process, the college can make a mistake and put irrelevant features on the list of the selected features. Various studies have shown that irrelevant features may have a negative impact on the prediction accuracy [16]. For this reason, DDM, which can make corrections in the list of the selected features that participate in the prediction, is embedded in the framework. In order to determine the impact of irrelevant features on the framework, four features were added in the original dataset. The evaluation was carried out on the same ISVU dataset. In the dataset, the following features were added: E-mail, name, surname, street name of residence. The features are taken because it is obvious that they cannot have any influence on the result of prediction. For example, the student name does not guarantee his performance during the study. Table 3 shows the evaluation results.

The measured results confirm that irrelevant features have a significant impact on the framework and the accuracy of prediction. Figure 4 shows the ROC curve of prediction. The difference in the measured accuracy of the prediction is 10%. Such large difference should not be too easily accepted. The results indicate the importance of the process of creating a dataset.

4.3. Evaluation framework prediction performance

The objective of the framework performance evaluation should provide answers to which the framework components that consume the most CPU resources, determine the level of consumption of memory in a situation of high levels of the load. Performance measurement was conducted on a virtual server installed on a standard computer (CPU i3 1.9 GHz, memory 4GB). Fig. 5 shows the time spent in the process of prediction. In Fig. 6 two curves are shown; the first one represents CPU times spent during the creating a hybrid CBR model and loading a case database into internal memory. The curve presents relation between prediction speed and database size (or dataset). The second curve represents the time spent for college's prediction. Archived results show that the creation time is in the range of 300-600 ms when the framework uses a database with 1700 case instances. The framework conducted high speed prediction, average time was approximately 400 ms. Observing the prediction in the framework level, recorded times should be aggregated. The aggregation result represents duration of the entire prediction process.

Fig. 7 shows load of the framework internal memory during simulation on the virtual server. During the framework testing, the framework was loaded with 200 virtual users tasks at the same time. In this way, we tried to simulate the real situation in which the framework can be used. The measured results show no overload of memory and critical situations regardless the frequent creation of hybrid CBR models.

5. Conclusions

This paper has introduced a new approach for prediction framework. The framework allows future students, based on college selection, to predict their performance during the future study. The system allows colleges the autonomy in forming their own model of prediction. The concept of frameworks offers colleges the autonomy in forming their own hybrid model. A detailed evaluation and performance tests confirm that framework concept is well-designed and as such achieves good results.

Results from our experiments suggest that the proposed framework based on hybrid CBR model approach possesses good properties from the standpoint of quality. This property gives the framework a note of universality. Our test concept is certainly a good starting point for further development. This research will be followed by additional testing of the framework in online environment, involving more colleges from various educational fields. Future testing of different types of datasets will indicate the elements that will need modification, with the main purpose to achieve better results.

References

[1] B. Azhagusundari and A. S., Thanamani, "Feature selection based on Information Gain," *International Journal of Innovative Technology and Exploring Engineering*, vol. 2, no. 2, pp. 18-21, 2013.

[2] A.A. Aziz, N. H. Ismail, F. Ahmad, and H. Hassan, "A Framework for students' academic performance analisis using Naive Bayes classifier," *Jurnal Teknologi (Sciences & Engineering)*, vol. 75, no. 3, pp. 13-19, 2014.

[3] G. Azkune, A. Almeida, D. López-de-Ipiña, and L. Chen, "Extending knowledge-driven activity models through data-driven learning techniques," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3115–3128, 2015.

[4] M. Behroozi and R. Boostani, "Presenting a new cascade structure for multiclass problems," in *Electronics, Computer and Computation (ICECCO)*, Ankara, 2013, pp. 192-195.

[5] S. Bernard, C. Chatelain, S. Adama, and R. Sabourin, "The multiclass ROC Front method for cost-sensitive classification," *Pattern Recognition*, vol. 52, pp. 46-60, 2015.

[6] F. Doctor and R. Iqbal, "An intelligent framework for monitoring student performance using Fuzzy Rule-Based linguistic summarisation," in *FUZZ-IEEE 2012, IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, 2012, pp. 1-8.

[7] L. Dole and J. Rajurkar, "A Decision support system for predicting student performance," *International Journal of Innovative Research in Computer and Communication*, vol. 2, no. 12, pp. 7232-7237, 2014.

[8] A. Fadhilah, I. H. Nur, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415-6426, 2015.

[9] S. Fong, S. Yain-Whar, and R. P. Biuk-Aghai, "Applying a hybrid model of neural network and decision tree classifier for predicting university admission," in *The Seventh International Conference on Information, Communications & Signal Processing*, Macau, 2009, pp. 1-5.

[10] S. Ghosh and S. K. Dubey, "Comparative analysis of K-means and Fuzzy C-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, pp. 35-39, 2013.

[11] J. Huang, Y.F. Lib, and M. Xie, "An empirical analysis of data preprocessing for machine learning-base software cost estimation," *Information and Software Technology*, vol. 67, pp. 108-127, 2015.

[12] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[13] A. E. Khedr and A. I. El Seddawy, "A proposed data mining framework for higher education system," *International Journal of Computer Applications*, vol. 113, no. 7, pp. 24-31, 2015.

[14] K. Kongsakun, C. C. Fung, and W. Philuek, "An intelligent recommendation system framework for student relationship management," in *8th International Conference on e-Business*, Bangkok, Thailand, 2009, pp. 87-91.

[15] W. Y. Lin, Y. H. Hu, and C. F. Tsai, "Machine learning in financial crisis prediction: A survey," *Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 421-436, 2012.

[16] A. Meshkin and H. Ghafuri, "Prediction of relative solvent accesibility by Support Vector Regression and Best-First method," *Experimental and Clinical Sciences*, vol. 9, pp. 29-38, 2010.

[17] V. R. Patel and R. G. Mehta, "Data clustering: Integrating different distance measures with modified k-means algorithm," *Advances in Intelligent and Soft Computing*, vol. 131, pp. 691-700, 2012.

[18] L. Paura and I. Arhipova, "Cause analysis of students' dropout rate in higher education study program," *Procedia - Social and Behavioral Sciences*, vol. 109, pp. 1282–1286, 2014.

[19] M. M. Richter and R. Weber, "Case-Based Reasoning: A Textbook," in *Basic CBR Elements*.: Springer Science & Business Media, 2013, pp. 17-34.

[20] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[21] J. Shana and T. Venkatachalam, "A Framework for dynamic faculty support system to analyze student course data," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 7, pp. 478-48, 2012.

[22] J. Shana and T. Venkatachalam, "Identifying key performance indicators and predicting the result from student data," *International Journal of Computer Applications*, vol. 25, no. 9, pp. 45-48, 2011.

[23] B. Shannaq, Y. Rafael, and V. Alexandro, "Student relationship in higher education using data mining techniques," *Global Journal of Computer Science and Technology*, vol. 10, no. 11, pp. 54-59, 2010.

[24] M. M. A. Tair and A. M. El-Halees, "Mining educational data to improve students' performance: A Case study," *International Journal of Information and Communication Technology Research*, vol. 2, no. 2, pp. 140-146, 2012.

[25] T. Tanner and H. Toivonen, "Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment," *International Journal of Learning Technology*, vol. 5, no. 4, pp. 356-377, 2010.

[26] T. Thiele, A. Singleton, D. Pope, and D. Stanistreet, "Predicting students' academic performance based on school and socio-demographic characteristics," *Studies in Higher Education*, vol. 41, no. 8, pp. 1424–1446, 2016.

[27] B. Trstenjak and D. Donko, "Predicting quality of web service using IKS Hybrid model," in *Telecommunications (BIHTEL)*, Sarajevo, Bosna i Hercegovina, 2014, pp. 1-6.

[28] K. Vembandasamy and T. Karthikeyan, "An improved FP-Growth with Hybrid MGSO-IRVM classifier approach used for Type-2 Diabetes mellitus diagnosis," *International Journal of Engineering and Technology*, vol. 7, no. 6, pp. 2293-2303, 2016.

Table 1. Overview of datasets and their properties.

Label	**College**	**Description**	**Feature No.**	**Size**	**Class No.**
D1	Computer engineering	technical science	15	280	2
D2	Management of tourism	social science	18	350	2
D3	Sustainable development	technical science	15	180	2
D4	Preschool Education Studies	social science	30	787	2
D5	Teacher Education Studies	social science	18	1250	2

Table 2. Overview archived prediction accuracy results.

Label	**CBR**	**PR**	**SE**
D1	95.86	0.90	0.91
D2	96.67	0.93	0.93
D3	96.20	0.90	0.91
D4	97.70	0.96	0.98
D5	92.86	0.82	0.80

Table 3. The achieved prediction results with and without irrelevant features.

Dataset	**CBR**	**PR**	**SE**
Dataset with irrelevant features	82.20	0.83	0.83
Dataset without irrelevant features	**92.86**	0.82	0.80

Figure captions:

1. Framework structure.
2. A diagram of prediction process in the web application.
3. The framework interface.
4. Framework ROC characteristics.
5. ROC curves of prediction with and without irrelevant features.
6. Analysis of the consumption time in the prediction process.
7. Load of the framework internal memory.

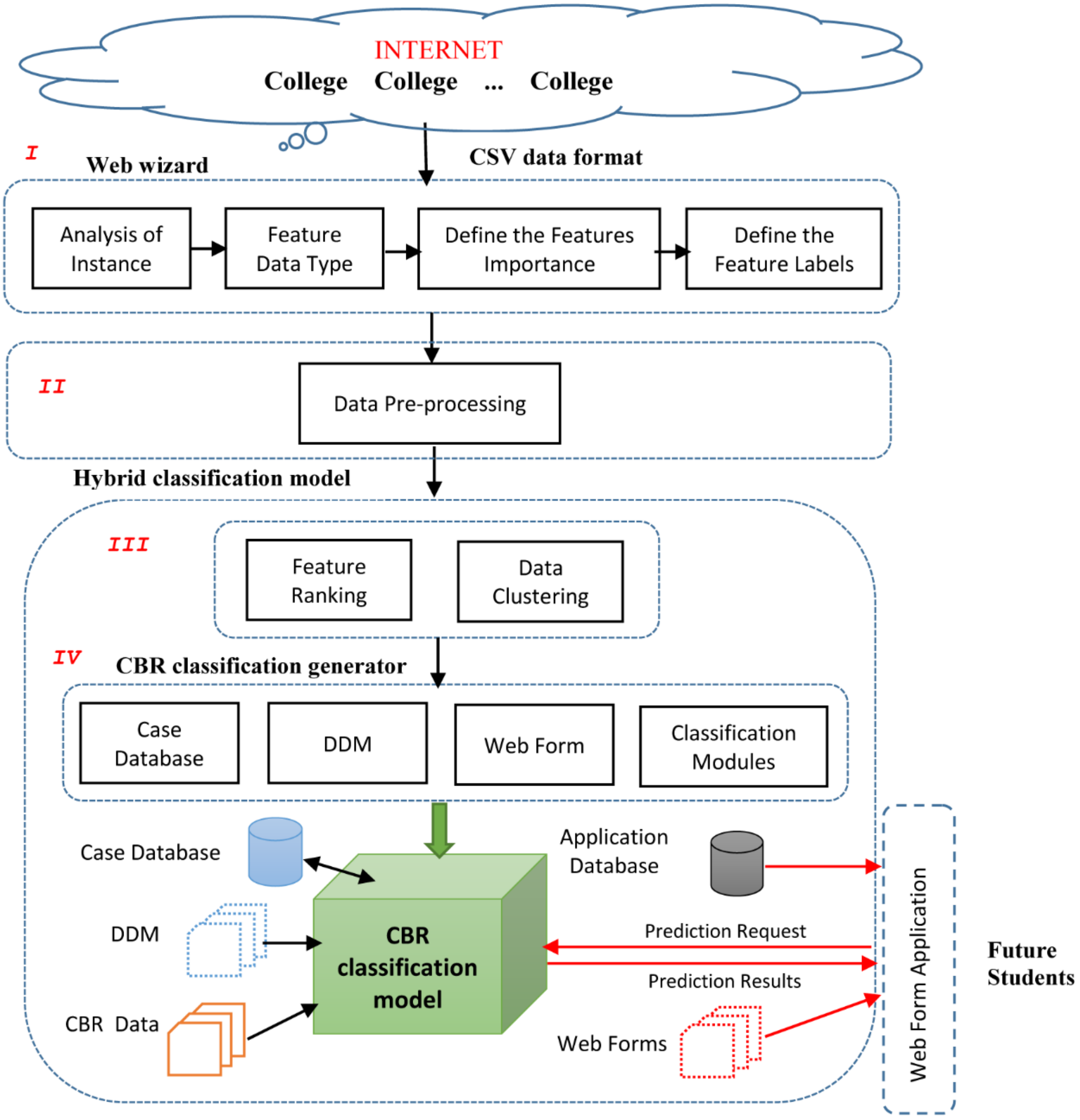


Fig. 1 Framework structure.

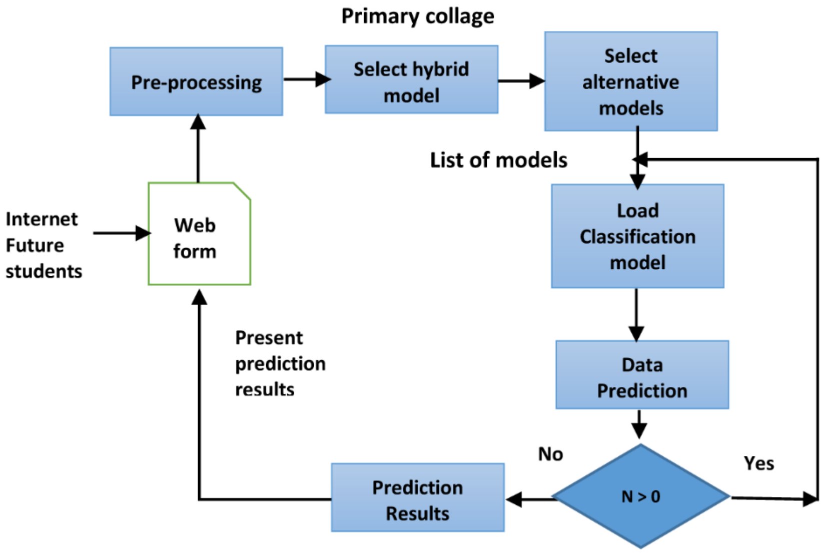
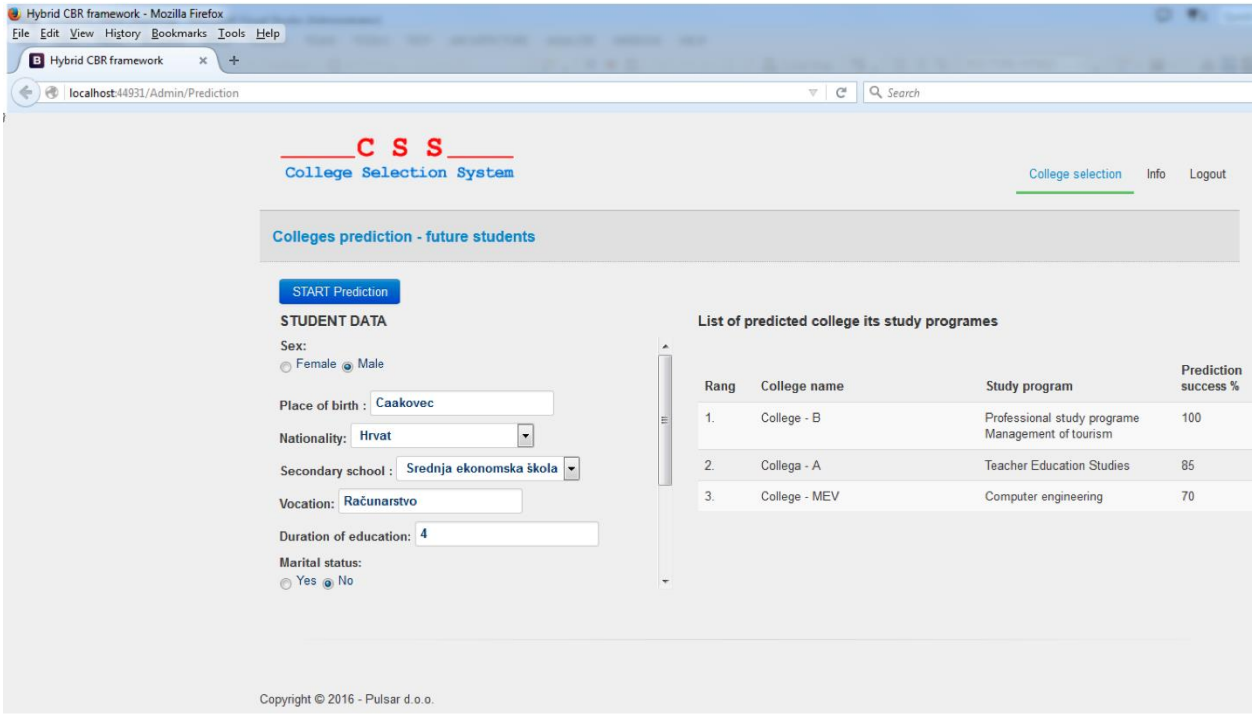


Fig. 2 Diagram of prediction process in the web application.


Hybrid CBR framework - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Hybrid CBR framework
localhost:44931/Admin/Prediction
Search
C S S
College Selection System
College selection
Info
Logout
START Prediction

Colleges prediction - future students

STUDENT DATA

Sex:
○ Female ● Male

Place of birth : Caakovec

Nationality: Hrvat

Secondary school : Srednja ekonomska škola

Vocation: Računarstvo

Duration of education: 4

Marital status:
○ Yes ● No

List of predicted college its study programes

Rang	College name	Study program	Prediction success %
1.	College - B	Professional study programe Management of tourism	100
2.	Collega - A	Teacher Education Studies	85
3.	College - MEV	Computer engineering	70

Copyright © 2016 - Pulsar d.o.o.

Fig. 3 The framework interface.

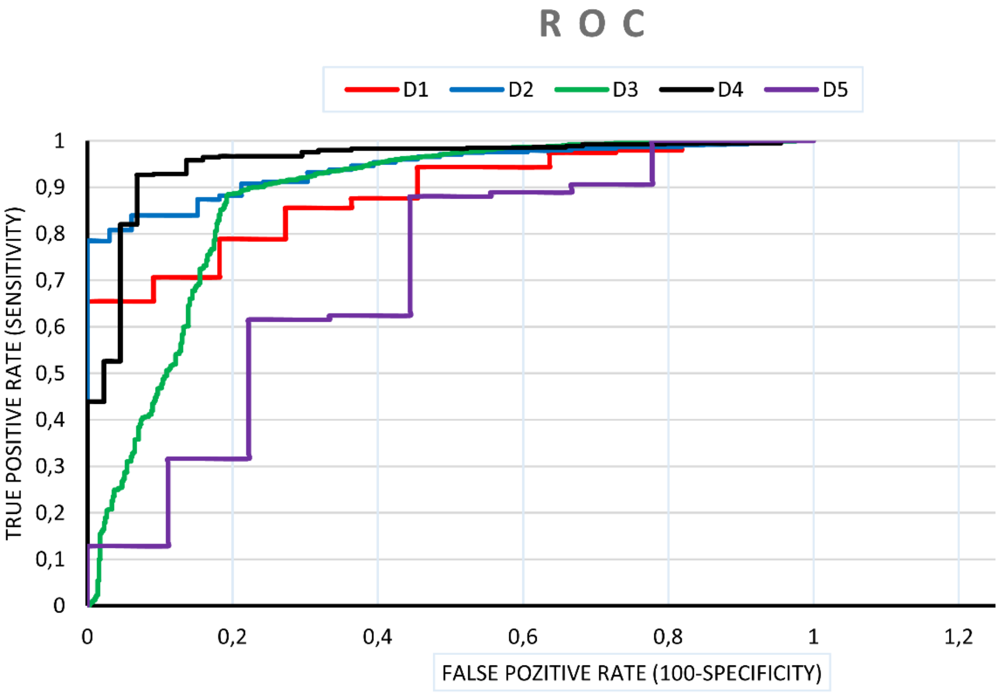


Fig. 4 Framework ROC characteristics.

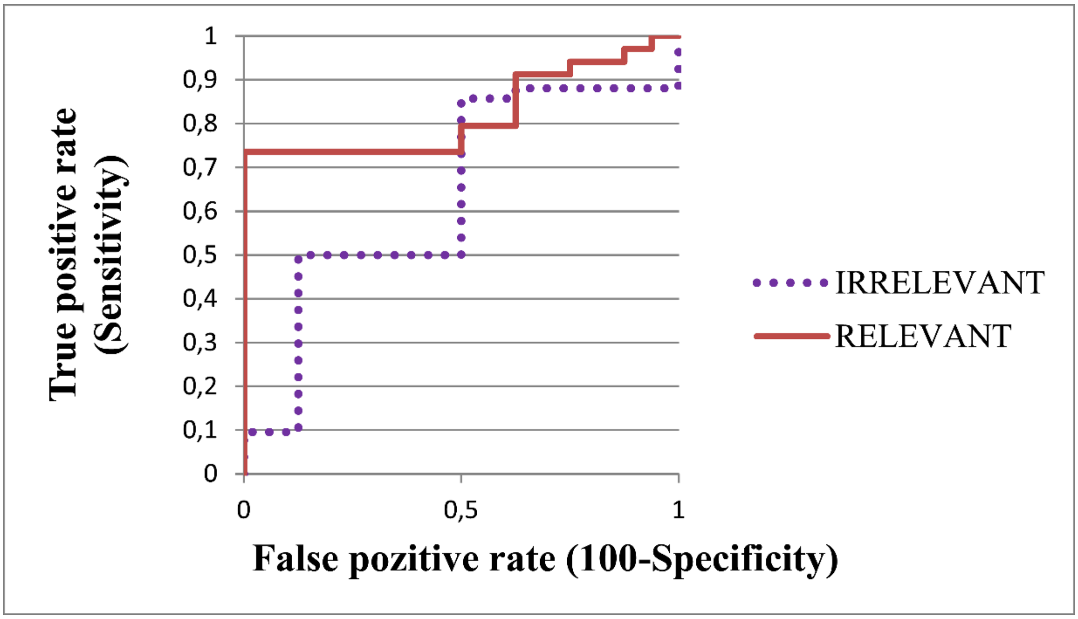


Fig. 5 ROC curves of prediction with and without irrelevant features.

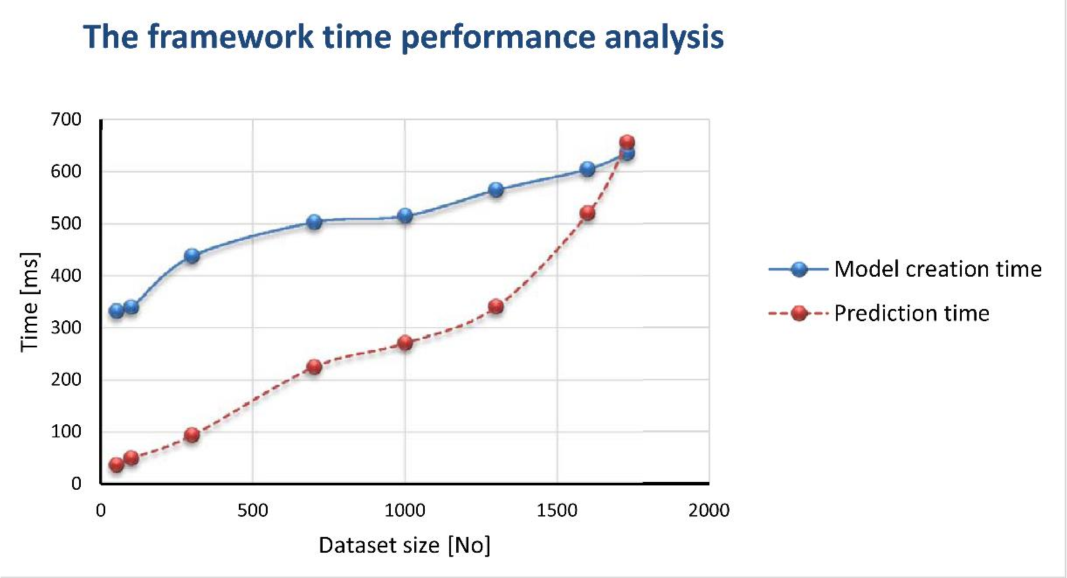


Fig. 6 Analysis of the consumption time in the prediction process.

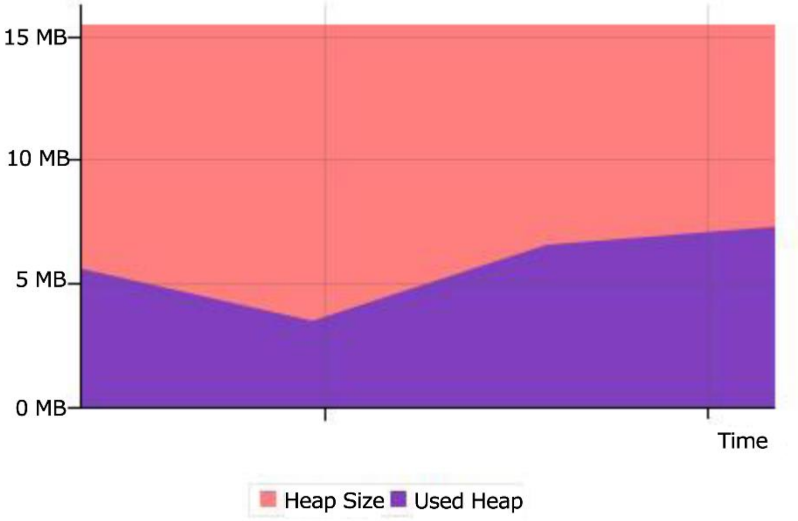


Fig. 7 Load of the framework internal memory