A Hybrid Recommender System to Enrollment for Elective Subjects in Engineering Students using Classification Algorithms

Jerson Erick Herrera Rivera[1]
National University of San Agustin
Arequipa, Peru

***Abstract*—One of the main problems that engineering university students face is making the correct decision regarding the lines of elective subjects to enroll based on available information (preferences, syllabus, schedules, subject content, possible academic performance, teacher, curriculum, and others). Under these circumstances, this research work seeks to develop a Hybrid Recommender System. For this, a model based on the Content-based approach of all the subjects that has been studied is developed (using Natural Language Processing and the statistical measures Term Frequency and Inverse Term Frequency), giving it appropriate relevance with the grades that the student has achieved. In addition, a model based on a Collaborative Filtering approach is developed, establishing relationships between different students, identifying similar academic behaviors. Thus, the system will recommend to the student in which lines of elective subjects to enroll to obtain better results in the academic field. The given recommendation will be obtained from machine learning models (XGBoost and k-NN) based on the similarity between the contents of each subject with respect to the line of elective subject and based on the academic relationship between all the students. To achieve the objective, data from engineering students between 2011 and 2016 has been analyzed. The results obtained indicate that the recommendations reach a MAP-k of 82.14% and a precision of 91.83%.**

***Keywords*—Hybrid; recommender system; academic performance; term frequency; inverse term frequency; natural language processing; k-NN; XGBoost; MAP-k**

I. Introduction

Systems Engineering students at National University of San Agustin, follow a curricular mesh of a considerable number of subjects, many of which are requirements of each other, also considering that they have lines of elective subjects from the second semester of the fourth year of university. There are 3 lines of elective subjects:

1) Line A: a) Electronic Business, b) Advanced Topics in Databases, c) Management of Information Systems and Technologies.

2) Line B: a) Introduction to the Development of Entertainment Software, b) Computer Graphics, Computational Vision and Multimedia, c) Development of Software for Games.

3) Line C: a) Introduction to the Development of New Platforms, b) Advanced Development in New Platforms, c) Emerging Platforms.

Of the lines of elective subjects previously described, every student is obliged to follow only two of them. The first subject of each line does not have a requirement to be taken, but the second and third have as a requirement the previous subject of each line of elective subjects.

Students must choose the most convenient lines of elective subjects for them, according to different criteria (interesting subjects according to their preferences, subjects in which their performance is higher, etc.). However, decision-making involves tasks that need time to be analyzed and include activities such as: searching the contents of each subjects of each line of elective subjects, examining carefully the syllabus, requesting access to the curriculum to analyze the content of each subject involved, review the statistics of the subject, or ask for advice from different students who already have the experience of the subject, although the comments may be too subjective depending on the experience.

The decision to choose in which lines of elective subjects to enroll brings with it some restrictions during the university studies. For example, the line chosen must be completed, thus, if the student chooses lines A and B, he must enroll and pass all the subjects on those lines, and otherwise he will not be able to obtain the degree of graduate or bachelor. Another complication is that the student cannot choose the lines again once they have enrolled in one; this means that there is no possibility of changing lines of elective subjects once they have been chosen the first time.

In addition to the previously described restrictions, the disapproval and dropout rate is high compared to other professional schools within the university. Therefore, there is a need for a tool that adequately suggests to the students in which lines of elective subjects they should enroll based on their preferences and performance in all the subjects they have previously studied, and based on the choice of students with academic behavior similar to the student obtaining an objective and exact recommendation; all this making use of the tools and techniques of: 1) a Content-based Recommender System, 2) a Collaborative Filtering Recommender System, and 3) a Hybrid Recommender System, generated from the results of the Content-based Model and the Collaborative Filtering Model.

This research seeks to solve a very common problem among university students by analyzing student performance and analyzing teaching content, tools that are part of Educational Data Mining (EDM) [1] [2] [3] [4] that is focused on the discovery of knowledge that involves education and data mining. EDM can be applied to discover patterns in data sets to automate the decision-making process of teachers, students and educational authorities [5].

The paper has been organized in the following way. Section 2 describes some basic concepts about Recommender Systems. Section 3 gives an overview of works related to Recommender Systems in education, and some using Hybrid Models. Section 4 describes the proposed solution, objectives, architecture, techniques, and methods used in the research. Section 5 details the procedure for developing the Hybrid Recommender System: Content-based model, the Collaborative Filtering model, and the hybridization. Section 6 details the accuracy levels achieved with the Content-based Model, the Collaborative Filtering Model, and the Hybrid Model. Finally, Section 7 describes the conclusions reached in this study, and details some guidelines on future works.

II. Recommender Systems

A. Recommender System

Recommender Systems (RS) are software tools and techniques providing suggestions for items to be of use to a user [6]. In [7], RS are defined as any system that provides individualization of the recommendation results and leads to a process that allows users to build interesting or useful objects in a wide range of possible options in a customized way. RS are specifically targeted at people who lack the professional knowledge or expertise to determine the potentially overwhelming number of alternative products a website has to offer [8].

Clearly, the functionality of RS is similar to the social recommendation and information reduction process, which is useless or uninteresting for the user. The main objective of the RS is to provide support to users in making their decisions (online). In particular, the goal is to provide accessible, high-quality recommendations for a large community of users with common features [9].

The basic RS models work with two types of data [10] [11], which are: 1) the user-item interactions, and appraisals associated with the items provided by the user and other users, such as ratings or buying behavior, and 2) the attribute information and description about the users and items such as text profiles or keywords. Methods which use the former are referred to as methods of Collaborative Filtering, while methods that use the latter are referred to as Content-Based recommender methods. Some RS combine these different aspects to create hybrid systems. Hybrids systems can incorporate the strengths of various types of RS to build approaches than can more robustly perform in a wide variety of settings.

B. Content-based

Content-based Recommender Systems (CBRSs) rely on item and user descriptions (content) to construct item representations and user profiles to suggest items similar to those already liked by a target user in the past. The basic process of producing content-based recommendations is to match the attributes of the target user profile with the attributes of the items in which preferences and interests are stored [6]. The main assumption behind this model is that the behavior of a user remains unchanged over time; hence, the content of past user actions may be used to predict the desired content of future actions [7].

At the most basic level, CBRS relies on two data sources: 1) The first data source is a description of different items in terms of content-centered attributes (for instance, a representation could be the manufacturer's text description of an item), and 2) the second data source is a user profile generated from user feedback about different items [10].

To determine the similarity between items, it is necessary to encode the content of each item, for this the TF-IDF matrix is used. TF (term frequency) describes how often a certain term appears in a document (assuming that important words appear more often). IDF (inverse document frequency) is the measure that is combined with the TF; their goal is to reduce the weight of terms that appear very often in all documents. The idea is that these very frequent terms are not useful to discriminate between documents, so more weight should be given to the words that appear in a few documents. To measure the similarity from the TF-IDF matrix it is necessary to use cosine similarity (1). This metric measures the similarity between two n-dimensional vectors based on the angle between them. The similarity between two items *a* and *b* is formally defined as follows:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a}.\vec{b}}{|\vec{a}|*|\vec{b}|} \quad (1)$$

C. Collaborative Filtering

Collaborative Filtering Recommender System (CFRS) is based on the assumption that similar users prefer similar items or that a user expresses similar preferences for similar items [7]. The basic idea of collaborative filtering methods is that these unspecified ratings can be imputed because the observed ratings are often highly correlated across various users and items. Most of the models for collaborative filtering focus on leveraging either inter-item correlations or inter-user correlations for the prediction process [10].

Euclidean distance (2) is the simplest and most common example of measure used to estimate the distance between two points and identify similar users or items, where n is the number of dimensions (attributes) and x_k and y_k are the kth attributes (components) of data objects x and y, respectively.

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (2)$$

CFRS methods are categorized into two general classes, namely model-based and memory-based [7]. Model-based algorithms use the underlying data to learn a probabilistic model, such as a cluster model or a Bayesian network model; subsequently they make predictions using the model. Memory-based methods store and access raw preference information in computer memory to find similar users or items, and make predictions as required. Based on a set of user ratings about

items, they seek to induce a model for each user based on a collection of user rating about items that would allow the classification of unseen items into two or more classes, each of which corresponds to specific points in the accepted rating scale.

D. Hybrid Methods

Hybrid methods combine two or more recommendation techniques to achieve better performance and to take out drawbacks of each technique separately [7]. It is notable that these different systems use different input types, which have distinct strengths and weaknesses [10]. Some recommender systems, such as content-based systems, are more effective in cold-start settings where there is no significant amount of data. Other recommender systems, such as collaborative methods, are more effective where there is significant amount of data, Usually, CFRS methods are combined with CBRS methods. According to [1], Hybrid RS could be classified into the following categories: 1) combining separate recommenders, 2) a single unifying recommendation model, 3) adding collaborative features to content-based models, and 4) adding content-based features to collaborative models.

According to [9], there are three base hybridization designs: monolithic, parallelized, and pipelined hybrids. The first design incorporates several recommendation techniques in the implementation of a single algorithm. The parallelized design, needs at least two models that produce recommendations independently, later combined with weighted, mixed, and switching strategies. The third design is when the output of one recommender becomes part of the input of the subsequent one.

III. Related Works

In [12], it is identified as a problem for university students to make the right decision regarding their academic itinerary based on available information (subjects, schedules, classrooms and teachers). This work proposes the use of an RS based on data mining techniques to help students in this type of decision. They worked with real data corresponding to seven years of the School of Systems Engineering of the University of Lima. After four tests an accuracy of 77.3% was achieved. They used the Decision Tree technique, which were created from a school database, to generate rules. Finally, the system generates recommendations based on these rules.

In [13] an Intelligent RS framework was designed that can predict the academic performance of the first year of tertiary education students, thus guiding the management of the educational institution in its decision-making on early intervention strategies. They used data obtained from the student archives of Babcock University, Nigeria. From such students, information was taken related to their family, pre-university educational performance and the result of the university entrance exam. For the study they used Decision Trees and Multilayer Perceptron to generate models; reaching an accuracy of 96.78%. Similarly, in [14], students' background information is used to analyze their performance in the first year of study.

Michael O'Mahony and Barry Smith in [15], have developed an Enrollment RS at the University of Dublin, where students learn 12 modules per year, of which 10 are specific to the area of study; and 2 modules are elective from the broader curriculum. Thus, the authors developed the system based on collaborative filtering and content-based methods. The first suggested elective modules based on past choices of students with similar behavior. The second made use of associated text fields detailing the module description and learning outcomes. After this, it is calculated the similarity between modules and determine which ones would be recommended to the student.

In [16], Vialardi et al. propose an enrollment RS based on the student's academic performance record. The system works with two attributes: a) inherent difficulty of a given course and b) measure of a student's competence for a given course based on grades obtained in similar courses. Different data mining methods were evaluated: C4.5, k-NN, Naïve Bayes, Bagging and Boosting, to achieve the best result for this application domain. They concluded that Bagging is the method that guarantees prediction accuracy.

In [17], AACORN is presented, a case-based system that recommends courses to students at DePaul University. Each student's information is organized based on four characteristics: the student's academic program, the curriculum, the student's general grade point average, and the student's history of courses. The system reuses the past experience of students to infer the appropriate courses that a student can enroll in the next study period. Two students in the same program and with similar interests are likely to take the same courses many times. In this way, a student seeking a recommendation can use the experience of students who have completed the program as if it were a template. Each course found in the template that the student has not taken is probably a good course to enroll in.

In [18] and [19], clustering (k-means) and association rules (a priori algorithm) are used to recommend courses to students in e-learning systems. Besides, it is developed an algorithm that combined both. As a result, it is concluded that the combined model generated more and better rules, which allows recommending different combinations of courses to the student, unlike the association rules model that only generated an association rule for the recommendation.

In [20], a Hybrid Recommender System based on machine learning is proposed to recommend Massive Online Open Courses (MOOC's). It makes use of implicit evaluations on the courses, to determine the behavior of each student and generate recommendations for users with similar preferences. The system is trained with a descending gradient. The main drawback found is how computationally expensive it is to make recommendations in real time. To solve this problem, the neighborhood concept is proposed, and with it the use of clustering techniques.

In [21], a hybrid multiple criteria RS applied to the recommendation of university courses is presented (information from the University of Cordova during three years) using CBRS and CFRS methods. The proposed model combines student and course information using configurable weightings to determine the relevance of each criterion. In this way, a genetic algorithm has been implemented in which the relevance of each criterion in the recommendations can be

controlled, as well as obtaining the best configuration of all the parameters used in the RS.

In [22], an element-based and user-based CFRS methods have been combined with a boosted CBRS method. The hybrid model adds average ratings as content based on collaborative filtering in the last step to make the final recommendation list more relevant to the user. The proposed hybrid algorithm was tested on two real-world datasets: 1) MovieLens dataset, and 2) dataset consisting of student scores at a Turkish university. The model was validated with k-fold cross-validation and a survey among students.

IV. The Proposal

The previously analyzed models use student information (educational and personal background, grade history), but do not give too much importance to the characteristics of the subjects themselves. The closest approach is to work with the inherent difficulty of a given subject [16] or to work with courses content [21]. The proposal is based on the use of characteristics of the subjects. It is difficult to describe a subject as quantitative variables. However, using natural language processing and word vectorization, it is possible to represent words or sentences as a vector of real numbers. In this way, the content of each subject can be represented as nominal values and used as input in a prediction system. Additionally, subjects can be objectively compared based on words/phrases that represent them with other subjects and even with student interests [23].

The main objective of this study is to design a Recommendation System architecture that adequately suggests to students in which lines of elective subjects they should enroll based on the student's profile, the subject's profile and the interactions between them, obtaining an objective and exact recommendation. To achieve this goal, it is necessary: 1) collect and structure the data of students, subjects and enrollments, 2) generate a Collaborative Filtering Recommender System, 3) generate a Content-based Recommender System, 4) generate a Hybrid Recommender System based on the results of the previous models, and 5) validate the accuracy of each generated model.

The proposed system has involved the analysis, design, implementation and validation of a Hybrid Recommender System that will allow student to know what are the most convenient lines of elective subjects for them, so they can follow subjects according to their preferences and in which they could perform better academically.

Fig. 1 shows the sequence of activities to be carried out for the development of the proposal and to achieve the stated objectives. The information of subjects was obtained from the contents that are described in detail in the curriculum (curriculum of 2002, 2013 and 2017). The information of students and enrollment is used to generate the academic performance matrices with the grades obtained in enrollments between 2013 and 2016. A CFRS is developed from the identification of the 10 students with academic performance most similar to the student, and analyze which lines of elective subjects are more convenient (with greater weight to the most similar students) and generate an ordered list of recommended subjects. A CBRS is developed based on identifying the 10 subjects in which it showed higher performance contrasting them with a TF-IDF matrix, generating an ordered list of recommended subjects. A hybrid RS is developed from the lists recommended by the 2 previous systems. Each ranked list is trained in different classification algorithms (Decision Trees, Logistic Regression, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naïve Bayes, Support Vector Classification, and XGBoost) to generate two models: 1) a model to predict the first line of elective subjects, and 2) a model to predict the second line of elective subjects; with both models generate a new recommendation. Each RS (CFRS, CBRS, Hybrid) was validated with the metrics: MAP-k (Mean Average Precision at k), precision and recall. MAP-k is a metric used to validate the precision in RS when the recommendation is treated as a ranked list, where it is rewarded for getting many "correct" or relevant recommendations, and it is rewarded for having them at the top of the list (better ranking).

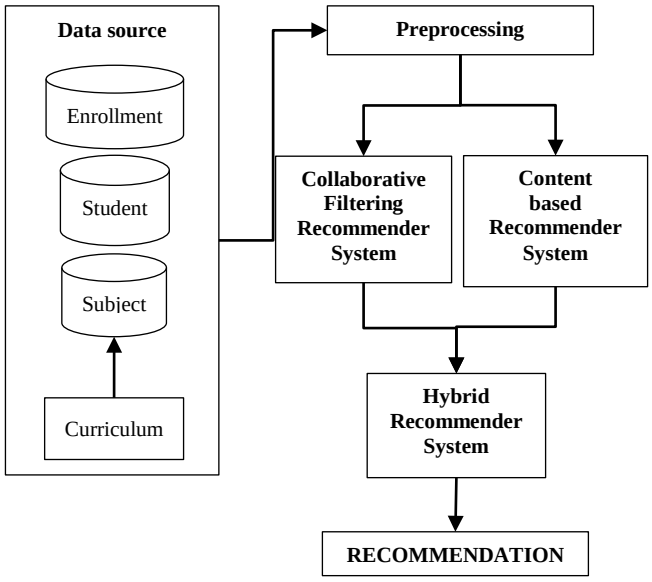

Fig. 1. Activities for the Development of the Proposed Hybrid Recommnder System.

V. Implementation of Proposed Model

For the development of the proposal, the Python programming language is used, together with the Jupyter development environment [24]. Additionally, it was necessary to incorporate the Python libraries: pandas, numpy, tika, pickle, sklearn, scipy, spacy, xgboost.

The enrollment (including grades), subject and student's data was stored in an .mdb file. Likewise, the content of the subjects was obtained from the curriculum, specifically in the section that gathers all the contents.

A. Collaborative Filtering Recommender System

To adequately represent the academic performance achieved by the student, it is necessary to normalize their grade. For this, the min-max normalization (3) is used, within each class (set of students who share the same subject, group,

academic term, and year). Thus, the normalized grade represents adequately the performance within a specific class. For example, if the student has a grade of 16, while the maximum grade for the class is 18 and the minimum grade is 8, then the normalized grade is 0.6.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3)$$

Later, a table is generated from the record of each student's grades. Each column of the new table represents each one of the subjects present in the curriculum of 2002 and 2013 (102 columns), and each row indicates the normalized grade that each student has obtained in each of those subjects (935 rows).

To identify which students have similar academic performance, the Euclidean distance (2) between each row is calculated. For the recommendation, the 10 most similar students to the target student are identified. From this, the elective subjects that these 10 students have chosen are identified, generating a weighted ordered list of lines of elective subjects that are going to be recommended to the target student.

B. Content-based Recommender System

To generate a CBRS it is essential to have coded information on each subject. For this, the contents of each subject stored in the curriculum are used. Analyzing these texts involves lemmatization or stemming processes. Stemming and lemmatization are techniques of Text Normalization, indispensable to process the content of each subject. Stemming is a process used in removing derivational suffixes as well as inflections so that word variants can be conflated into the same roots (the roots do not have to be words of a language). On the other hand, lemmatization uses vocabulary and morphological analysis of word and tries to remove inflectional endings, thereby returning words to their dictionary form. In [25] y [26], they have compared both techniques and agree that in comparison with stemming, lemmatization produced higher precision. Consequently, this paper uses the lemmatization technique.

During the lemmatization process, the spacy library is used to lemmatize the content of each subject. It is also necessary to convert the text to lowercase, and conjunctions, prepositions, punctuation marks, stop words are eliminated (a text normalization technique, which uses vocabulary and morphological analysis of word and tries to remove inflectional endings, thus returning words to their dictionary form). With the lemmatized content of each subject, and using the TfidfVectorizer object [27] from the sklearn library, a vocabulary of features common to all documents is generated (vocabulary with 100 words), and most importantly, a matrix of TF-IDF features. The documents are encoded by TF-IDF matrix as vectors in a Euclidean space, where the dimensions of the space correspond to the features that appear in the vocabulary.

Once the TF-IDF matrix has been created, the next step is to identify the subjects in which the student has obtained the best performance; this is measured with the normalized grade. The 10 best-valued subjects are analyzed from the lemmatize content, and attached to the TF-IDF matrix. The new record is represented as a vector in the TF-IDF matrix and represents the student profile. The cosine similarity (1) is used to identify which elective subjects are the most similar to the student profile. Thus, a weighted ordered list of recommended lines of elective subjects is generated and recommended to the student.

C. Hybrid Recommender System

The previous systems generate a weighted list of lines of elective subjects according to academic behavior similar to other students (CFRS) and academic performance according to personal preferences (CBRS). Both are weighted lists that reflect the relevance of each line, choosing the two lines with the greatest relevance.

The Hybrid RS takes the weights for each line as input to various classification algorithms that will predict which lines of elective subjects to recommend. Two models are created for each algorithm: 1) a model to predict the most relevant line of elective subjects as a first option and 2) a second model to predict the second line of elective subjects as a second option; remembering that students have to choose two lines of elective subjects from the three offered by the university.

The classification algorithms used are as follows: Logistic Regression, Decision Trees, k-Nearest Neighbors (k-NN), Linear Discriminant Analysis, Gaussian Naïve Bayes, Support Vector Classification from Support Vector Machine (SVM), and XGBoost. For modeling, the dataset was divided into training data and testing data (30% for the first model, and 25% for the second model, thus avoiding overfitting). In Table I it can be seen that the first model to predict the first option, achieves better predictions (60% precision in testing data) with the algorithm XGBoost and Logistic Regression. While Table II shows the precision in the second model (to predict the second option) that it achieves better predictions (77% precision in testing data) with the k-NN algorithm. Therefore, the Hybrid RS uses the XGBoost algorithm (better performance and greater adaptability to different datasets compared to Logistic Regression) to recommend the first option of line of elective subjects, while the k-NN algorithm will recommend the second option.

TABLE I. PRECISION IN CLASSIFICATION ALGORITHMS TO OPTION 1

	Algorithm Precision	
Algorithm	***Training***	***Testing***
Logistic Regression	0.82	0.60
Decision Trees	1.00	0.47
k-Nearest Neighbors	0.62	0.40
Linear Discriminant Analysis	0.76	0.47
Gaussian Naive Bayes	0.74	0.40
Support Vector Classification	0.44	0.40
XGBoost	1.00	0.60

TABLE II. PRECISION IN CLASSIFICATION ALGORITHMS TO OPTION 2

	Algorithm Precision	
Algorithm	***Training***	***Testing***
Logistic Regression	0.67	0.69
Decision Trees	1.00	0.54
k-Nearest Neighbors	0.69	0.77
Linear Discriminant Analysis	0.61	0.62
Gaussian Naive Bayes	0.67	0.54
Support Vector Classification	0.61	0.54
XGBoost	1.00	0.62

VI. RESULTS

Throughout the study, three Recommender Systems have been developed: a Collaborative Filtering Recommender System, a Content-based Recommender System, and a Hybrid Recommender System. The results of each of them are summarized in Table III. The precision in CFRS is 54% and in CBRS it is 63%, however, in the Hybrid RS, based on the output of the other two systems, the precision reached 91%, the 68 and 70% improved the precision of the CBRS and CFRS, respectively. Furthermore, using the MAP-k metric that takes into account the ranked position of the recommendation, CFRS reached 35%; CBRS, 55%; and the Hybrid RS, 82%. Again, the Hybrid RS obtained more accurate results (improved by 130% over CFRS, and 47% over CBRS). Finally, the proposed system reaches the following metrics: the precision is equal to 0.91, recall is 0.83, and F1 is 0.87.

TABLE III. RESULTS IN PROPOSED RECOMMENDER SYSTEMS

	Recommender System Precision	
Algorithm	***Precision***	***MAP-k***
Collaborative Filtering Recommender System	0.540	0.352
Content-based Recommender System	0.632	0.556
Hybrid Recommender System	0.918	0.821

VII. CONCLUSIONS AND FUTURE WORKS

The study developed a CFRS (using a student-subject matrix with the grade obtained), a CBRS (using a TF-IDF matrix and generation of a student profile) and a Hybrid RS (using a classification algorithm with the results of CFRS and CBRS as input). The precision level achieved by the first two models was regular, while after hybridization, the results improved considerably. In this way it was proven that hybrid models take the advantages of CRFS and CBRS, and overcome the disadvantages of them by working individually. Thus, the recommendation of lines of elective subjects to choose during enrollment, reflects adequately the relationship between students, subjects, academic performance and student preferences. Therefore, the recommendations generated by the proposal support objectively the students' decision during the enrollment.

Given that the levels of precision reached by the Content-based Recommender System are greater than the Collaborative Filtering Recommender System, it can be suggested that the attitude of the students towards a given course (student preferences) is highly relevant when recommending a line of elective subjects.

Despite the fact that all the developed models do not take the time as a relevant variable in the design or validation, it is important to update the input data of each model, to regenerate the models guaranteeing their validity. It would have positive effects adding behavioral information of the students in the face of the subjects (attendance, partial exams, assignments, etc.) in addition to the final grade in the subject.

REFERENCES

[1] Romero and S. Ventura, "Data mining in education," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 1, pp. 12–27, 2013.

[2] R. Jindal and M. D. Borah, "A Survey on Educational Data Mining and Research Trends," Int. J. Database Manag. Syst., vol. 5, no. 3, pp. 53–73, 2013.

[3] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert Syst. Appl., vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

[4] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," Comput. Educ., vol. 51, no. 1, pp. 368–384, 2008.

[5] A. Sheshasaayee and M. Nazreen Bee, "E-learning: Mode to improve the quality of educational system," Smart Innov. Syst. Technol., vol. 78, pp. 559–566, 2018.

[6] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook, Second Edi., vol. 54. 2015.

[7] A. S. Lampropoulos and G. A. Tsihrintzis, Machine Learning Paradigms Applications in Recommender Systems. 2015.

[8] A. Klašnja-Milićević, M. Ivanović, and A. Nanopoulos, "Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions," Artif. Intell. Rev., vol. 44, no. 4, pp. 571–604, 2015.

[9] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender Systems: An Introduction, vol. 40. Cambridge, 2011.

[10] C. C. Aggarwal, Recommender Systems The TextBook, vol. 40, no. 3. Springer, 2016.

[11] C. Cobos et al., "A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes," Inf. Process. Manag., vol. 49, no. 3, pp. 607–625, 2013.

[12] C. Vialardi, J. Bravo, L. Shafti, and A. Ortigosa, "Recommendation in Higher Education Using Data Mining Techniques," Proc. 2nd Int. Conf. Educ. Data Min., pp. 191–199, 2009.

[13] M. Goga, S. Kuyoro, and N. Goga, "A Recommender for Improving the Student Academic Performance," Procedia - Soc. Behav. Sci., vol. 180, no. November 2014, pp. 1481–1488, 2015.

[14] M. Joshi, P. Bhalchandra, A. Muley, and P. Wasnik, "Analyzing students performance using Academic Analytics," Proc. 2016 Int. Conf. ICT Business, Ind. Gov. ICTBIG 2016, pp. 0–3, 2017.

[15] M. P. O'Mahony and B. Smyth, "A recommender system for on-line course enrolment: an initial study," Proc. 2007 ACM Conf. Recomm. Syst. - RecSys '07, p. 133, 2007.

[16] C. Vialardi et al., "A data mining approach to guide students through the enrollment process based on academic performance," User Model. User-adapt. Interact., vol. 21, no. 1–2, pp. 217–248, 2011.

[17] R. Burke, "Aacorn: A CBR recommender for academic advising," Tech. Rep. TR05-015, 2005.

[18] S. B. Aher and L. Lobo, "Applicability of data mining algorithms for recommendation system in e-learning," Proc. Int. Conf. Adv. Comput. Commun. Informatics - ICACCI '12, p. 1034, 2012.

[19] S. B. Aher and L. M. R. J. Lobo, "Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data," Knowledge-Based Syst., vol. 51, pp. 1–14, 2013.

[20] V. Garg and R. Tiwari, "Hybrid massive open online course (MOOC) recommendation system using machine learning," Int. Conf. Recent Trends Eng. Sci. Technol. - (ICRTEST 2016), pp. 1–5, 2016.

[21] A. Esteban, A. Zafra, and C. Romero, "Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization," Knowledge-Based Syst., vol. 194, p. 105385, 2020.

[22] S. Çapraz and S. Temizer, "A Content Boosted Hybrid Recommender System," ICCIT 2017, 2017.

[23] J. Herrera Rivera, "Content based recommender system for enrollment of elective subjects in engineering careers at the public University of Arequipa," Proc. LACCEI Int. Multi-conference Eng. Educ. Technol., 2019.

[24] M. Ragan-Kelley et al., "The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication.," in AGU Fall Meeting Abstracts, 2014.

[25] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," Int. Conf. Inf. Knowl. Manag. Proc., pp. 625–633, 2004.

[26] V. Balakrishnan and E. Lloyd-yemoh, "Stemming and lemmatization: A comparison of retrieval performances," pp. 174–179, 2014.

[27] D. Cournapeau and M. Brucher, "TfidfVectorizer," 2007. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. [Accessed: 11-Jan-2019].