IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11822

Predicting Academic Performance Based on Social Activities

R. Chandra Kiran[1], Saranya[2]

ME final year, Department of Computer Science and Engineering, CSI College of Engineering,

Ketti, TamilNadu, India[1]

Asst. Professor, Department of Computer Science and Engineering, CSI College of Engineering,

Ketti, TamilNadu, India[2]

Abstract: Predictive modelling is an important part of learning analytics, whose main objective is toestimate student success, in terms of performance, knowledge, score or grade. The data used for thepredictive model can be either state-based data (e.g., demographics, psychological traits, past performance)or event-driven data (i.e., based on student activity). The latter can be derived from students' interactionswith educational systems and resources; learning management systems are a widely analysed data source,while social media-based learning environments are scarcely explored.Data is collected from a Web ApplicationsDesign course, in which students use wiki, blog and microblogging tools, for communication andcollaboration activities in a project-based learning scenario.In addition to the novel settings and performance indicators, an innovative regression algorithm is used for grade prediction. Very good correlation coefficients are obtained and 85% of predictions are within one point of the actual grade, outperforming classic regression algorithms.

Keywords: component, formatting, style, styling, insert (key words)

I. INTRODUCTION

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue. Data mining is also known as data discovery and knowledge discovery. The major steps involved in a data mining process are:

Extract, transform and load data into a data warehouse

Store and managedata in a multidimensional databases

Provide data access to businessanalysts using application software

Present analyzed data in easily understandable forms, such as graphs The first step in data mining isathering relevant data critical for business. Company data is either transactional, non-operational or metadata.

Transactional data deals with day-to-day operations like sales, inventory and cost etc. Non-operational data isnormally forecast, while metadata is concerned with logical database design. Patterns and relationships among data elements render relevant information,which may increase organizational revenue. Organizations with a strongconsumer focus deal with data mining techniques providing clear pictures ofproducts sold, price, competition and customer demographics. For instance,the retail giant Wal-Mart transmits all its relevant information to a datawarehouse with terabytes of data. This data can easily be accessed bysuppliers enabling them to identify customer buying patterns. They cangenerate patterns on shopping habits, most shopped days, and most sought forproducts and other data utilizing data mining techniques. The second step in data mining is selecting a suitable algorithm - a mechanism producing a datamining model. The general working of the algorithm involves identifyingtrends in a set of data and using the output for parameter definition.1.1.1Educational data mining: Educational data mining (EDM) describes aresearch field concerned with the application of data mining, machinelearning and statistics to information generated from educational settings(e.g., universities and intelligent tutoring systems). At a high level, the fieldseeks to develop and improve methods for exploring this data, which oftenhas multiple levels of meaningful hierarchy, in order to discover new insightsabout how people learn in the context of such settings. In doing so, EDM hascontributed to theories of learning investigated by researchers in educational psychology and the learning sciences. The field is closely tied to that oflearning analytics, and the two have been compared and contrasted.Educational data mining refers to techniques, tools, and research designed forautomatically extracting meaning from large repositories of data generated byor related to people's learning activities in educational

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11822

settings. Quite often,this data is extensive, finegrained, and precise. For example, several learningmanagement systems (LMSs) track information such as when each student accessed each learning object, how many times they accessed it, and howmany minutes the learning object was displayed on the user's computerscreen. As another example, intelligent tutoring systems record data everytime a learner submits a solution to a problem; they may collect the time ofthe submission, whether or not the solution matches the expected solution, theamount of time that has passed since the last submission, the order in which solution components were entered into the interface, etc. The precision of thisdata is such that even a fairly short session with a computer-based learningenvironment (e.g., 30 minutes) may produce a large amount of process datafor analysis. In other cases, the data is less fine-grained. For example, astudent's university transcript may contain a temporally ordered list of coursestaken by the student, the grade that the student earned in each course, andwhen the student selected or changed his or her academic major. EDMleverages both types of data to discover meaningful information aboutdifferent types of learners and how they learn, the structure of domainknowledge, and the effect of instructional strategies embedded within variouslearning environments. These analyses provide new information that would be difficult to discern by looking at the raw data. For example, analyzing datafrom an LMS may reveal a relationship between the learning objects that astudent accessed during the course and their final course grade. Similarly,analyzing student transcript data may reveal a relationship between a student'sgrade in a particular course and their decision to change their academic major.

Objectives

Learning analytics (LA) is a growing research area, whichaims at selecting, analysing and reporting student data (intheir interaction with the online learning environment),finding patterns in student behaviour, displaying relevantinformation in suggestive formats; the end goal is theprediction of student performance, the optimization of theeducational platform and the implementation of personalizedinterventions Various educational tasks can be supported by learninganalytics, as identified in analysis and visualization ofdata providing feedback for supporting instructorsproviding recommendations for students; predicting student'sperformance; student modelling; detecting undesirable studentbehaviours; grouping students; social network analysis;developing concept maps; constructing courseware; planningand scheduling.

Existing System

According to the Society of Learning Analytics Research1, LA can be defined as "themeasurement, collection, analysis and reporting of data aboutlearners and their contexts Analysis and visualization of data providing feedback for supporting instructors providing recommendations for studentspredicting student's performance student modeling; detecting undesirable student behaviors; grouping students; social network analysis developing concept maps in

Disadvantages

This is highly interdisciplinary, including machine learning techniques, educational data mining, statistical analysis, social network analysis, natural language processing .

particular the instructor can be advised about students at-risk, who are in need of more assistance

Proposed System

Data is collected from a Web Applications Design course, in which students use wiki, blog and micro blogging tools, for communication and collaboration activities in a project-based learning scenario. From a pedagogical perspective, the results indicate that, as a general rule, a higher engagement with social media tools correlates with a higher final grade. Important part of learning analytics, whose main objective is to estimate student success, in terms of performance, knowledge, score or grade. The data used for the predictive model can be either state-based data

Advantages

It is worth mentioning that the performance of the generalized predictive modelThe differences in instructional conditions and technology use, even in the context of the same discipline, may influence the prediction of academic success; in addition, the individual differences of the students involved in the studies Any attempt at generalizability needs to carefully consider the pedagogical and disciplinary context of the predictive model. Therefore, further studies are needed for comparing courses with different internal and external conditions

II. MODULES AND DESCRIPTION

1. Instructional Scenario
2. Data Collection And Preprocessing
3. Data Analysis
4. Data Prediction

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11822

INSTRUCTIONAL SCENARIO

Our study took place in the context of an undergraduate course for Computer Science students, on Web Applications Design (WAD). The instructional approach was project based learning (PBL) in which the students had to work collaboratively on various complex, challenging and authentic tasks, over extended periods of time; learning was organized around team projects, while the teacher played the role of a facilitator. More specifically, the students collaborated in teams of around 4 peers in order to build a complex web application of their choice (e.g., a virtual bookstore, an online auction website, a professional social network, an online travel agency, etc.).

The project spanned over the whole semester and the evaluation took into account both the final product and the continuous collaborative work. Since PBL has a strong social component, the increasingly popular social media tools appear suitable for communication and collaboration support in PBL framework. Hence, we implemented our PBL scenario with the help of several social media tools (wiki, blog, and microblogging tool) integrated in our social learning environment, called eMUSE. More specifically, a blended learning approach was used consisting of weekly face-to-face meetings between each team and the instructor (for checking the project progress, providing feedback and answering questions), while students had to rely on the social media tools for the rest of the time, as a support for their communication and collaboration activities.

In particular, MediaWiki was used for collaborative writing tasks, for gathering and organizing the team knowledge-base and resources, and for documenting the project. Blogger was used for reporting the progress of each project, similar to a "learning diary" in terms of publishing ideas and resources, as well as for providing feedback and solutions to peer problems.

Each team had its own blog, but inter-team cooperation was encouraged as well. Twitter was meant to foster additional connections between peers and to encourage the posting of short news, announcements, questions, and status updates regarding each project. The eMUSE social learning platform provides an integration point for the social media tools, together with additional support for both students and teachers: basic administrative services, learner tracking and data visualizations, as well as evaluation and grading functionalities. eMUSE also offers data collection mechanisms, as detailed in the next subsection.

Of course, students could choose to use additional communication channels for working on their projects, including face to face meetings, phone calls, chats, email, document sharing or other social media tools. Obviously, these could not be monitored by eMUSE; this means that a part of learner data may not be collected, which is a general limitation of learning analytics approaches based on student activity indicators. In order to mitigate this problem in our PBL scenario, we provided specific instructions to the learners at the beginning of the semester: students were clearly informed that their collaborative learning activity needs to be documented on the social media tools integrated in eMUSE, so that it can be assessed by the instructor. We therefore expect that a large part of the students' communication and collaboration activities indeed took place on the three recommended social media tools.

DATA COLLECTION AND PREPROCESSING

The instructional scenario described above has been applied over 6 consecutive winter semesters (Year 1: 2010/2011 –Year 6: 2015/2016), with 4th year undergraduate students in Computer Science from the University of Craiova, Romania. Small improvements and refinements were made from one year to the consecutive one, based on students' feedback and instructor experience. A total of 343 students, enrolled in the WAD course, participated in this study. All student actions on the three social media tools were monitored and recorded in the eMUSE platform.

The system retrieves learner actions from each of the disparate Web 2.0 tools (by means of open APIs or Atom/RSS feeds) and stores them in a local database, together with a description and an associated timestamp. Thus, a total of almost 19000 social media contributions were recorded: 2609 blog posts and comments, 5470 tweets, 10895 wiki page revisions and file uploads.

Based on these actions, a set of 14 numeric features were computed for each student:

NO_BLOG_POSTS(the number of blog posts)

NO_BLOG_COM(the number of blog comments)

AVG_BLOG_POST_LENGTH (the average length of a blog post)

AVG_BLOG_COM_LENGTH (the average length of a blog comment)

NO_ACTIVE_DAYS_BLOG(the number of days in which a student was active on the blog, i.e., wrote at least a post or a comment)

NO_ACTIVE_DAYS_BLOG_POST (the number of days in which a student wrote at least a post on the blog)

NO_ACTIVE_DAYS_BLOG_COM (the number of days in which a student wrote at least a comment on the blog)

NO_TWEETS(the number of tweets)

NO_ACTIVE_DAYS_TWITTER (the number of days in which a student was active on Twitter, i.e., posted at least a tweet)

NO_WIKI_REV(the number of wiki page revisions)

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11822

NO_WIKI_FILES(the number of files uploaded on the wiki)
NO_ACTIVE_DAYS_WIKI (the number of days in which a student was active on the wiki, i.e., revised at least a page or uploaded at least a file)
NO_ACTIVE_DAYS_WIKI_REV (the number of days in which a student revised at least a wiki page)
NO_ACTIVE_DAYS_WIKI_FILES (the number of days in which a student uploaded at least a file on the wiki)

Students' performance at the end of the semester was assessed on a 1 to 10 scale, with 5 being the minimum passing grade; the evaluation took into consideration both the final project, as well as each student's continuous collaborative work. Our aim is to predict this final grade based on the above set of features, using the LMNNR algorithm, as described in the next section.

DATA ANALYSIS

In this section, we present the results obtained with the LMNNR algorithm, in comparison with those of the algorithms that provided the best results in, namely Random Forestwith 100 trees and k-Nearest Neighbors, with k obtained by cross-validation and inverse-distance weighting of the neighbors. In, an additional setting for kNN implemented in Weka was that mean absolute error was used when doing cross-validation. In this paper, we use mean squared error (MSE) instead, as we observed that it slightly improves the accuracy of the kNN model.
For the analysis of individual years, we chose 3 neighbors and 1 prototype for LMNNR, because it is a simple model that also provides good results. A drawback of LMNNR training is that it sometimes converges into local optima. Several examples of convergence, plotting the value of the obtained objective function F against the corresponding MSE. It can be seen that the lowest value of F does lead to the lowest value of the MSE, but there are also many situations when the obtained MSE is not so good, even for low values of F. That is why we run the algorithm several times and retain the best results. Even if this requires additional training time, we consider that the quality of the obtained results, which are clearly better than those found by the other models, compensate for this inconvenience. In a previous work, an evolutionary algorithm was used for training, but the gradient-based method used here is much faster, although it needs multiple starting points.
One of the main motivations for analyzing learning data is the ability to predict student behavior early in the course, therefore we also assessed how useful the trained model is for predicting student behavior in future years. Thus, we trained the algorithm on data from earlier years and tested its prediction performance on data from later years.

DATA PREDICTION

Academic Performance:
The function of the predictive system is to make predictions of the students' marks at the end of the term, based on the data collected from the interactive web. These data, selected and organized into features, are normalized and given to a machine learning algorithm as input. The predictive system classifies the student expected performance (measured as a mark in percentage) in three possible classes: high performance (expected mark > 80.5%), medium performance (57.5% < expected mark < 80.5%) or low performance (expected mark < 57.5%).
The reason to split the output in three classes is to get an adequate performance out of the classification algorithm, adjusted to the size of the sample (the sample has only 336 students, so three equilibrated classes of 112 individuals are proposed). The balanced distribution of the students in the mark range explains the selected intervals for each class. This is a very effective and efficient Machine Learning algorithm that works very well for general datasets like the one under analysis.
Social activities:
Only a few prior studies have investigated the impact of social media activity on academic performance, despite the growing availability of such data and undisputed presence of these media in our daily lives. The majority of existing studies found a decrease in academic performance with increasing time spent on social media. However, not all studies confirm this result. In some studies, time spent on social media was found to be unrelated to academic performance or even a had positive effect on performance.
There is a growing interest in the relationship between social interactions (especially online social interactions) and academic performance. In the relevant literature there exist two dominant approaches. The first approach focuses on the relation between own performance and that of peers, based on a hypothesis of similarity in peer achievement. The similarity between pairs of individuals connected via social ties are attributed to various aspects: selection into friendships by similarity (i.e., homophily); influence by social peers (also known as peer effect); and correlated shocks (e.g., being exposed to the same teacher). As noted by the issue of separating these effects is inherently difficult.
The second approach emphasizes the positive influence of having a central position in the social network between students. The majority of results in the existing research which measure social networks are, however, based on self-reports and therefore subject to various biases that are in many ways mitigated by using smartphones to measure the social network. However, it should be noted that surveys and observational studies often measure very different aspects of

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

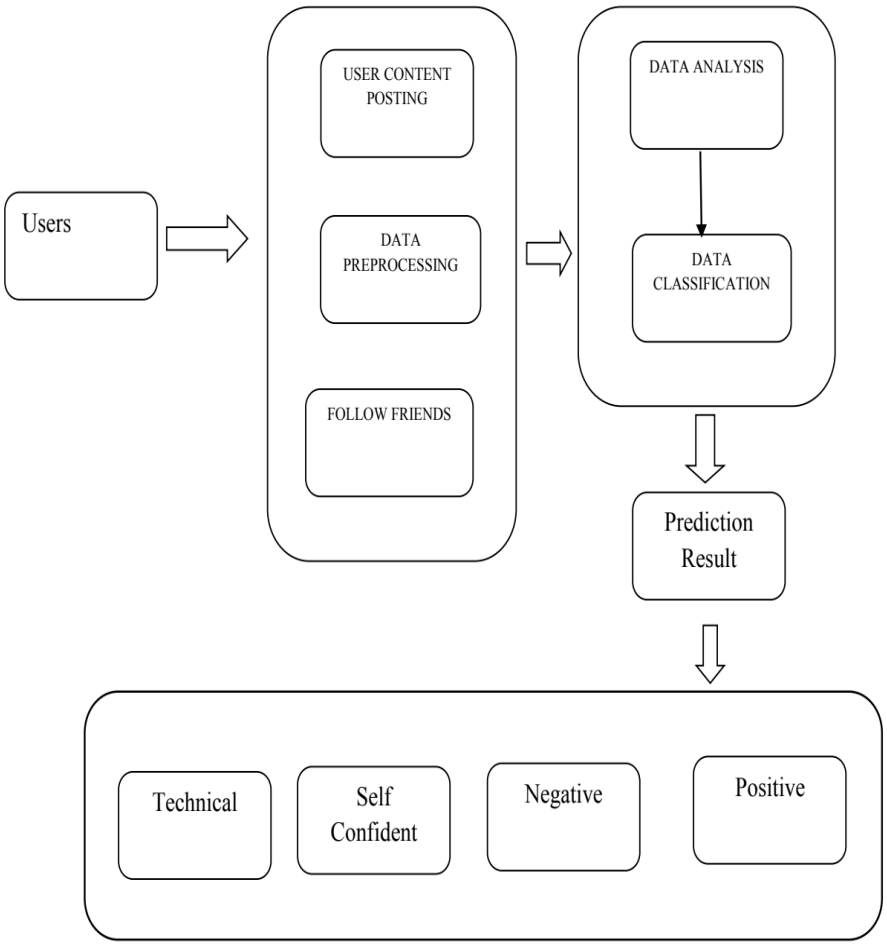
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11822

reality. For instance, in the case of assessing tie strengths, observational studies may be more accurate in quantifying duration and frequency variables of a relationship, while surveys can provide qualitative insights into depth and intimacy.

III. ARCHITECTURAL DIAGRAM



IV. SYSTEM SPECIFICATION

Software Requirements:

Operating System	:	Windows 7 / 8
Language	:	Java, J2EE
Developing Tool	:	Netbeans 7.2.1
Technologies	:	JSP, Servlet
Backend	:	MySql Server

Hardware Requirements:

Processor	:	Dual Core
Ram	:	2GB
Hard Disk	:	160 GB Space

V. INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11822

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
Select methods for presenting information.
Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the Future. Signal important events, opportunities, problems, or warnings.
Trigger an action.
Confirm an action.

VI. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.
Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

VII. CONCLUSTION

The project has shown that students' actions on social media tools are good predictors of academic performance. The innovative algorithm proved verysuitable for our prediction problem, outperforming classic regression algorithms.Very good correlation coefficients were obtained and much more of predictionswere within only 1 point of the actual grade. From a pedagogical perspective, theresults indicate that, as a general rule, a higher engagement with social mediatools correlates with a higher final grade. This is inline with several previousstudies, which found that online participation is a strong indicator of studentperformance and improves learning effectiveness. Nevertheless, there are alsocontradictory studies, which concluded that students learned equally wellregardless of their level of online participation . At the same time, the body ofliterature specifically focused on students' active participation on social media isscarce, hence the novelty and added value of our study. are also contradictorystudies, which concluded that students learned equally well regardless of theirlevel of online participation. At the same time, the body of literature specificallyfocused on students' active participation on social media is scarce, hence thenovelty and added value of our study. It is worth mentioning that the performance of the generalized predictive model is slightly lower than theperformance of each individual year model. This is in line with the findings in ,which addresses the issue of aggregating trace data from different courses forcreating one generalized model for academic success prediction. The differencesin instructional conditions and technology use, even in the context of the samediscipline, may influence the prediction of academic success; in addition, theindividual differences of the students involved in the studies (e.g., meta cognitiveand motivational factors) may have an impact on the learning analytics results.Any attempt at generalizability needs to carefully consider the pedagogical anddisciplinary context of the predictive model. Therefore, further studies are neededfor comparing courses with different internal and external conditions. Hence,investigating the LMNNR algorithm performance on student data collected fromdifferent courses and instructional scenarios is an interesting research direction.Furthermore, combining the predictive analytics approach proposed here withour previous work on social network analytics and discourse analytics could leadto a more comprehensive perspective on the social learning process andenvironment.

REFERENCES

[1] M. Abdous, W. He, and C.-J. Yen, ''Using data mining for predicting
relationships between online question theme and final grade,'' Edu. Technol.Soc., vol. 15, no. 3, pp. 77–88, 2012.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11822

[2] J.W.AlsteteandN.J.Beutell,''Performanceindicatorsinonlinedistance learningcourses: a study of management education,'' Quality Assurance Educ., vol. 12,no. 1, pp. 6–14, 2004.

[3] O. Ardaiz-Villanueva, X. Nicuesa-Chacón, O. Brene-Artazcoz, M. L. S. deAcedo Lizarraga, and M. T. S. de Acedo Baquedano, ''Evaluation of computertools for idea generation and team formation in project-based learning,'' Comput.Educ., vol. 56, no. 3, pp. 700–711, 2011.

[4] K. C. Assi, H. Labelle, and F. Cheriet, ''Modified large margin nearestneighbor metric learning for regression,'' IEEE Signal Process. Lett., vol. 21, no.3, pp. 292–296, Mar. 2014.

[5] R. S. Baker and P. S. Inventado, ''Educational data mining and learning analytics,'' in Learning Analytics, J. A. Larusson and B. White, Eds. New York,NY, USA: Springer-Verlag, 2014, pp. 61–75.

[6] R. Barber and M. Sharkey, ''Course correction: Using analytics to predictcoursesuccess,''inProc.2ndInt.Conf.Learn.Anal.Knowl.(LAK),2012, pp.259–262.

[7] A. Becheru and E. Popescu, ''Using social network analysis to investigate students' collaboration patterns in eMUSE platform,'' in Proc. ICSTCC, Oct.2017, pp. 266–271.