Analyzing Personality Traits and External Factors for Stem Education Awareness using Machine Learning

Sang C. Suh[1]
Department of Computer Science
Texas A&M University-Commerce
Commerce, TX–USA

Anusha Upadhyaya B.N[2]
Department of Computer Science
Texas A&M University-Commerce
Commerce, TX-USA

Ashwin Nadig N.V[3]
Department of Computer Science
Texas A&M University-Commerce
Commerce, TX-USA

***Abstract*—The purpose of the paper is to present the personality traits and the factors that influence a student to pursue STEM education using machine learning techniques. STEM courses have high regard because they play a vital role in global technology, inventions and the economy. Educational Data Mining helps us to identify patterns and relationships in a large educational database. On the other hand, Machine Learning facilitates decision making process by enabling learning from the dataset. A survey comprising of an extensive variety of questions regarding STEM education was conducted and the opinions of students from various backgrounds and disciplines were collected. A dataset was generated based on the responses from students. Machine Learning algorithms (one class-SVM and KNN) applied on this dataset emphasizes variety of courses offered, research-oriented learning, problem-solving approach, a good career with high paying job are some of the factors which may influence a student to choose STEM course.**

Keywords*—Educational Data Mining (EDM); Science Technology Engineering Management (STEM); Machine Learning (ML); K-Nearest Neighbor (KNN); One class–Support Vector Machine (one class - SVM)

I. Introduction

A. Significance of STEM Education

Education is now one of the fundamentals of living. It is no more just a tool to spread knowledge. Every day in this modern world, a new scientific invention or technology is being introduced. Science, Technology, Engineering and Mathematics have integrated into all aspects of life and it has been more accessible than ever before. Students of all areas of study are directly or indirectly interacting with these aspects, no matter what is the field of study there is some technology to store it, share it and enhance its possibility in various dimensions.

As this era is turning out to be an era of automation and machines are becoming more capable of exhibiting their intelligence and skill, current and next generation of students have to compete with machines also. So the focus of the research is on the attitude, knowledge and abilities of the students regarding STEM discipline [1]. As an individual chooses his/her path for a career based on the interests which evolved from his/her childhood. So it is essential to analyze the distinct quality of each person who is in multiple year of study in college and his/her perspective towards STEM education.

The teaching Science, Technology, Engineering and Mathematics also emphasize integrating other areas of study to be a part of them. A person who teaches these subjects should be able to identify research and explain the significance and uniqueness of this integrity among his/her students. Integration of the skills acquired from STEM courses will play a vital role in understanding and re-structuring of various aspects of life [2]. Thus, a student should be able to understand the local politics and know what is happening around the globe and ready to articulate his/her views on it. Inquiry-based learning model can help students to become scientists in their way to explore new information. STEM courses are potent tools for a student to contrive solutions. It also helps in understanding, being creative and innovative in approach to any given challenge. Furthermore, STEM helps to acquire the necessary skills [1].

B. Machine Learning in Educational Research

The influence of Machine Learning and Artificial Intelligence is already noticeable on the global economy. Thus, it is driving much attention from analysts. ML deals with several algorithms to enhance their performances. It tackles multiple aspects of regression or classification issues in research related to data analytics [2]. *Educational Data Mining* (EDM) deals with extracting necessary information from massive data sets, which are related to students. EDM involves various algorithms like Decision trees, K-Nearest Neighbor, Neural Networks, Support Vector Machines and other Machine Learning algorithms. Analysis, prediction and data-driven problem-solving techniques have shown effective results in forecasting consumer behaviors, fraud detection, intrusion detection and various assessments. The education system also can be one of the significant areas where one can implement data-driven techniques, where it will help to analyze various patterns associated with students [3]. At the same time, statistical methods make it difficult to identify and comprehend.

In this modern world that is driven by data, there is an exponential growth of textual chronicles in the field of education. Every device connected to the internet is an ocean of knowledge. As the online tutoring and skill sharing are becoming more popular each day, this process is producing a massive amount of data daily. Classification of this enormous amount of data is the biggest challenge in data analytics and there have been introduced a variety of methods to handle data volumes. One such method of classification is K-Nearest

Neighbor algorithm which is a supervised learning method [10]. On the other hand, there is an unsupervised learning method called one-class SVM, where a model is trained on the data with only one class. It gleans the features of typical cases and from these features it can predict the cases which are not normal. The main aim of these classifications on educational data is to help the authorities and teachers to improve the performance of a student and assist them in exploring learning and career options with early predictions which are based on previous data.

In this paper, Machine Learning algorithms, one-class SVM and KNN are applied on the data collected from a survey conducted on the students of Texas A&M University-Commerce, where the students from various disciplines gave their responses. The survey consisted of questions regarding their knowledge and opinion about STEM education. Thus, one can able to predict the factors that may influence a student to take up a course associated with STEM or not.

In Section II, we have presented various works which are related to Educational data mining and Machine Learning. Section III describes Machine Learning techniques. Section IV has experimental analysis. Section V has the conclusion and followed by the future work.

II. Related Work

Educational data mining is the new popular topic among the researchers who are conducting researches on higher education in various institutions. Data mining in the education sector helps us to understand multiple hidden traits among the students, and it helps to analyze the factors influencing a student to take up a particular course and dedicate all their time in learning it. STEM courses are one of the major streams in education, where the STEM courses will help a student to have a successful career. So the awareness regarding this plays a vital role in a students' life.

F. Sciarrone [2] conducted Educational data mining (EDM) with Machine Learning techniques to extract the data. Moreover, he described various models of Learning Analytics with EDM.

Thakar, Mehta, and Manisha [3] have described the techniques used in data mining of Educational data and have given a brief description of the work that has been conducted regarding Educational Data Mining.

Bhardwaj and Pal [4] described the techniques to analyze student performance using decision trees, based on the data extracted from student assessments and their final exam results.

Romero and Ventura [5] surveyed the studies carried out on EDM, which used the computational approaches to analyze the educational data to study educational questions, also it elaborately analyzes the various educational environments and data produced by it.

Meyrick [6] compared traditional college methods and STEM-based program and gave an insight into the positive aspects which influence students to attend STEM Courses.

Popenici and Kerr [7] explored the emergence of artificial intelligence techniques in learning and teaching. It explores the connection between education and emerging technologies by the teaching methods of institutions and how students evolve.

B. Yildirim [8] gave an overview of research studies that are conducted on STEM education and the attributes of students associated with it. Several studies are considered for meta-synthesis method. The results of the study emphasize that the students of STEM discipline are more creative and interested in solving problems.

Amra and Maghari [11] gave a student performance prediction model by applying KNN and Naïve Bayesian on an educational dataset of secondary schools, extracted from the ministry of Gaza strip in 2015, where they compared two of the techniques.

Imdad et al. [12] proposed a method for student result classification based on two traditional algorithms KNN and artificial neural network using the data from the Pakistan education board.

Manevitz and Yousef [13] gave a detailed description of different version of SVMs for one class classification with the context information retrieval.

Kruengkrai and Jaruskulchai [14] presented a comprehensive approach of relevant sentence extraction using only positive samples for training. Here they have applied methodology of support vector machines for one class classification. The fundamental goal of one class SVM is to modify data into feature space compatible with Kernel and then discrete them from original with the highest margin.

Ciolacu et al. [16] presented their case study with the primary goal of predicting the final score of students before attending the exams. Authors proposed an early recognition system with actual data extracted from an embedded learning curriculum with the personalized test for each before the start of the semester.

To analyze the personality traits of the students regarding the knowledge about STEM education, a survey was conducted at the Texas A&M University-Commerce, where students were given a set of statements and questionnaire to respond. The poll was designed to explore the knowledge of a student regarding STEM and to know according to them what the positive and negative aspects of STEM education are. Moreover, the survey helped in understanding the perspective of a student regarding STEM education along with various external factors such as the latest technological trends, world economy, etc.

III. Machine Learning Techniques

Machine Learning techniques are categorized into two types, supervised learning and unsupervised learning. In supervised learning, the outcome can be predicted using the previous input and output data, whereas in unsupervised learning, the algorithm recognizes the hidden patterns or the internal structures of the input data. In our research one

algorithm in supervised learning (KNN) and one algorithm in unsupervised learning (One-class SVM) were considered.

A. One-Class Support Vector Machine (One-Class SVM)

Schölkopf et al. [9] presented a support vector method for novelty detection. The data points are separated from the origin and distance is maximized from this hyperplane to the origin. Results were shown as in binary function where regions are captured in the input space with the probability density if data is present. So the function yields +1 for the small region and -1 for the rest of it.

$$\min_{w,\xi_i,\rho} \frac{1}{2} \|w\|^2 + \frac{1}{\upsilon n}\sum_{i=1}^{n} \xi_i - \rho \quad (1)$$

Subject to:

$$\left(w \cdot \phi(x_i)\right) \geq \rho - \xi_i \quad \text{for all } i=1..., n \; \xi \geq 0 \quad \text{for all } i=1., n$$

In equation (1), parameter ν(*nu*) characterizes the solution. It decides the upper bound on the fraction of outliners and the number of training examples is the lower bound used as support vectors.

By a kernel function for dot product calculations, the final decision function is given in equation (2):

$$f(x) = \operatorname{sgn}\left(\left(w \cdot \phi(x_i)\right) - \rho\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i K(x, x_i) - \rho\right) \quad (2)$$

B. K-Nearest Neighbor (KNN) Algorithm

K- Nearest Neighbor (KNN) is one of the supervised machine learning techniques. It is a non-parametric lazy learning algorithm. The main aim of the KNN algorithm is to utilize a database wherein which the data points are divided into various classes to predict the classification of a new sample point.

One of the challenging aspects regarding the KNN algorithm is to finding proper value of k [10], For instance, if k equal to 1, then it is defined as the nearest neighbor algorithm. It is simple and straight forward to implement, where it requires only two parameters, stores all the given cases and classifies new cases depending on the similarity measure. In this research, Minkowski distance metric is given by equation (3).

$$\left(\sum_{i=1}^{k} \left(\left|x_i - y_i\right|\right)^q\right)^{\frac{1}{q}} \quad (3)$$

IV. Experimental Analysis

As described in the previous section, the survey responses are given to identify various perspectives of students regarding STEM. The survey was successful in identifying some of the key factors which influenced them to pursue STEM. Those key factors include the attitude towards STEM, big five personality traits of an individual [15] and the external factors like career opportunities and professional growth by studying STEM.

The responses given by the students were recorded and used to create a dataset. Further, applying the Machine Learning techniques on the dataset, one will be able to predict the factors influencing a student to take up STEM or not.

To apply the two machine learning techniques on this dataset, the following parameters were considered. For one-class SVM, MinMaxScaler is used for feature scaling and radial basis function kernel (RBF) and the ν(*nu*) parameter value is 0.025. Parameters used in KNN algorithm is, StandardScaler is used for feature scaling, Minkowski distant metric for measuring distance and the number of neighbors (k) is 1. Results of the experiment are given in Table I.

TABLE I. Accuracy of Classification

Algorithm	Precision	Recall	Accuracy	F1 Score
One-Class SVM	98.54%	98.5%	98.5%	98.49%
KNN	87.34%	87.25%	87.25%	87.08%

V. Conclusion

The paper has examined the accuracy of features, which are considered to be the influencing factors on a student to choose STEM courses. The analysis of the data using machine learning highlighted the influence of certain personality traits on students to opt STEM courses, for instance, a student who is confident in science is likely to take up math for study. Likewise, the person who is interested in a particular area of studies such as robotics, engineering, psychology, statistics and the person who has ability and passion on technical problem solving is more likely to choose STEM rather than any other course. There are other external factors such as the variety of courses offered, research-oriented learning, extensive area of study and most importantly the opportunities to explore and a good career with high paying jobs have elevated impact on choosing STEM courses for their education.

VI. Future Work

The survey is limited to a particular university, and it reflected the attitude of a certain set of students. In the future, this process can be conducted on a large scale to determine factors influencing current and future students with the involvement of students from various state and educational institutions. This will help to understand the educational and career perspectives of students on national and international levels.

References

[1] Q. C. Pham, R. Madhavan, R. Chatila, L. Righetti, and W. Smart, "The Impact of Robotics and Automation on Working Conditions and Employment [Ethical, Legal, and Societal Issues]," IEEE Robotics & Automation Magazine, vol. 25, no. 2, pp. 126–128, June 2018.

[2] F. Sciarrone, "Machine Learning and Learning Analytics: Integrating Data with Learning," in 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET), April 2018.

[3] P. Thakar, A. Mehta, and Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," International Journal of Computer Applications, vol. 110, no. No. 15, January 2015.

[4] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2011.

[5] Cobal Romero and S. an Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions On Systems Man And Cybernetics, vol. 40, no. 6, pp. 601–618, NOV 2010.

[6] K. M. Meyrick, "How STEM Education Improves Student Learning," Meridian K-12 School Computer Technologies Journal, vol. 14, no. 1, 2011.

[7] S. A. D. Popenici and S. Kerr, "Exploring the impact of artificial intelligence on teaching and learning in higher education," in Research and Practice in Technology Enhanced Learning, 2017.

[8] B. YILDIRIM, "An Analyses and Meta-Synthesis of Research on STEM Education." Journal of Education and Practice, vol. Vol.7, no. No.34, 2016.

[9] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylo, and J. C. Platt, "Support Vector Method for Novelty Detection," in Advances in Neural Information Processing Systems '12.

[10] Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," in 2017 8th International Conference on Information Technology (ICIT). IEEE May 2017.

[11] I. A. A. Amra, A. Y. A. Maghari "Students performance prediction using KNN and Naïve Bayesian" ICIT 2017 — 8th Int. Conf. Inf. Technol. Proc., pp. 909-913 2017.

[12] Imdad, Ulfat, et al. "Classification of students results using KNN and ANN." 2017 13th International Conference on Emerging Technologies (ICET). IEEE, 2017.

[13] Manevitz, Larry M., and Malik Yousef. "One-class SVMs for document classification." Journal of Machine Learning research 2. Dec (2001): 139-154.

[14] Kruengkrai, Canasai, and Chuleerat Jaruskulchai. "Using one-class SVMs for relevant sentence extraction." International Symposium on Communications and Information Technologies. 2003.

[15] Ciolacu, Monica, et al. "Education 4.0-Artificial Intelligence Assisted Higher Education: Early recognition System with Machine Learning to support Students' Success." 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging(SIITME). IEEE, 2018.

[16] L. M. P. Zillig, S. H. Hemenover, and R. A. Dienstbier, "What Do We Assess When We Assess a Big 5Trait?: A Content Analysis of the Affective, Behavioral, and Cognitive Processes Represented in Big 5 Personality Inventories," Personality and Social Psychology Bulletin, vol. 28, no. 6, pp. 847–858, June 2002.