Machine Learning for Predictions in Academics

Shashi Sharma, Sunil Kumar Pandey, Kumkum Garg

Abstract— ***In recent years, a lot of data has been generated about students, which can be utilized for deciding the career path of the student. This paper discusses some of the machine learning techniques which can be used to predict the performance of a student and help to decide his/her career path. Some of the key Machine Learning (ML) algorithms applied in our research work are Linear Regression, Logistics Regression, Support Vector machine, Naïve Bayes Classifier and K- means Clustering. The aim of this paper is to predict the student career path using Machine Learning algorithms. We compare the efficiencies of different ML classification algorithms on a real dataset obtained from University students.***

Keywords— Machine Learning, Student Performance

I. INTRODUCTION

In recent years, application of the Internet has increased tremendously, resulting in the availability of huge amounts of data. This data come primarily from excessive use of social media, but also from IoT systems and other day-to-day digital transactions. Most of this data is useless and carries no useful information. To generate information from big data and analyse it for a particular purpose, Machine Learning techniques are used.

Machine Learning (ML) is a subfield of Artificial Intelligence. It is the study from examples and experience, without explicitly programming computers. Machine learning focuses on the development of computer programs that can access data and use it for learning for deriving relevant functions [1].

Arthur Samuel defined Machine Learning (ML) as follows: “Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.’’ More recently, in 1997, Tom Mitchell of Carnegie Mellon University, went further into the engineering concept and said that “A computer program is said to learn from experience E with respect to a task T and some performance measure P, if its performance on T, as measured by P, improves with experience E” [1].

ML is categorized as either supervised or unsupervised, just as humans learn through supervised or unsupervised ways. A child learns many things in life through his own experience, without being told, or a research scholar

studies and discovers or ‘learns’ new concepts. The supervised learning of human beings comes through their elders like parents and friends, or through teachers.

Classification algorithms are useful for predicting outputs that are discrete. In other words, they are useful when the answer to a question falls under a predictable set of probable outcomes. Here the system is required to answer with yes-or-no prediction; for example, ‘Is this cancer?’

ML uses data to the generic algorithm rather than writing program and develops logic based on the data given. The goal of ML is not to make perfect guesses, because it deals with domains where there is no such thing. Its goal is to make the guesses good enough to be useful. Sine ML works on the principles of statistics, the training samples given to the machine have to be sufficient in number and random.

In Rajasthan, every year 15-18 Lakh students complete their 12th standard from RBSE and CBSE Boards. Data about these students are used to counsel them and prepare them for better career opportunities, depending on their capability. Growing volumes and varieties of data have made it difficult to analyse manually. Machine Learning techniques help us to reduce this human effort.

II. LITERATURE REVIEW

In this section, we review some papers in related areas.

Vaidu et al implemented ML techniques to predict employability skills, based on students’ performance. They implemented K- Nearest Neighbors (KNN) and Naïve Bayes to classify students into several groups. Both the algorithms are used to predict of the employability of students. The results obtained for KNN is 95.33% and for the Naïve Bayes, it is 67.67% accuracy. [2]

Iqbal et al discuss different ML techniques to predict grades in different courses. They use matrix Factorization, Regression and Classification models like collective filtering and Restricted Boltzmann Machine (RBM) techniques to analyse data collected from Information Technology University (ITU), Pakistan. They evaluated the performance of the students who got admission in the Bachelor’s Program, using ML techniques. The RBM technique was found to be the best among different ML techniques. [3]

Byung-Hak et al have proposed a Deep Leaning based algorithm GritNet for forecasting the future performance of students. GritNet gives better results as compared to standard Logistic Regression according to this paper. It takes student data from Udacity Nanodegree Programs. [4] Jie et al have proposed a ML approach for predicting the student performance in degree programs.

Manuscript published on January 30, 2020.
* Correspondence Author

Shashi Sharma*, School of Computing Skills, Bhartiya Skill Development University, Jaipur, India. pathak.shashi26@gmail.com

Sunil Kumar Pandey, School of Computing Skills Bhartiya Skill Development University, Jaipur, India. skphind@gmail.com

Prof.(Dr.) Kumkum Garg, Dean Academics Bhartiya Skill Development University, Jaipur, India kumkum.garg@ruj-bsdu.in

This investigation is based on past and present performance. The proposed system uses a bi-layered structure comprising of multiple base predictor and a data driven approach, based on latent factor models for constructing an efficient base predict.

This paper shows that the proposed method achieves better performance as compared to benchmark approaches. [5]

Pojon Murat examines ML algorithms to predict student performance. He uses three different techniques, viz., Linear Regression, Decision Trees and Naïve Bayes Classification on two different datasets: row version and feature engineered version. The best technique is Naïve Bayes for the first dataset with 98% accuracy and Decision Trees for second dataset with 78% accuracy. [6]

Singh et al discuss ML techniques to predict the subject wise academic performance of engineering students. They predict subject scores in ongoing courses by analysing subjects based on the previous semester. For this purpose, they use two classification techniques, Naïve Bayesian and C4.5 Decision Tree classifier. The result shows that C4.5 Decision Tree has higher accuracy than Naïve Bayes. [7]

BendengnuKsung et al propose a Deep Neural Network model for prediction of student performance and class category. The paper compares the accuracy of Deep Neural network (DNN) with existing different Machine Learning techniques like Decision Tree (J48), Naïve Bayes and ANN. This model achieves accuracy of 84.3% and is better than other Machine Learning Algorithms. [8]

Pushpa S et al predicted final results whether the student will pass or fail on the basis of students' performance in the previous semester using Machine Learning Algorithms. This paper used four algorithms: Support Vector Machine, Naïve Bayes, Random Forest and Gradient Boosting. The accuracy of Random Forest 89.06% was greater than other algorithms. [9]

Gerritsen L. et al predicted student performance from Learning Management System data in the context of educational data mining using Neural Networks. The dataset used for this paper was a Moodle log file containing 4601 students' information. This paper compared the performance of Neural Networks against six other classifiers. These algorithms were Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector machine and Logistics regression. This paper showed that the accuracy of the Neural network was better than six other classifiers. [10]

Martín S. et al analyse the performance of four Machine Learning algorithms with different perspective, to predict dropout in university students. Four algorithms, Random forest, Neural Networks, Support Vector Machines and Logistics Regression are used. The dataset uses those students who enrolled in a degree program at the Instituto Tecnológico de Costa Rica (ITCR) between the years 2011 and 2016. The Random Forest algorithm with two variables is the best for predicting dropouts. [11]

III. MACHINE LEARNING TECHNIQUES

Five machine learning techniques have been used in this paper. These are as follows:

Logistic Regression: It is a classification and not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on a known set of independent variables. In other words, it predicts the possibility of occurrence of an event by fitting data to a logit function. Therefore, it is also known as **logit regression**. It predicts the probability and its output values lie between 0 and 1.

Decision Tree: Decision tree is a supervised learning algorithm which is mostly used for classification problems. Decision tree is used for both categorical and continuous dependent variables. In this algorithm, we divide the population into two or more similar sets. A decision tree is a graphical representation that makes use of branching methodology to represent all possible outcomes of a decision based on certain conditions.

Support Vector Machine: It classifies the data into different classes by finding a line (hyperplane) which divides the training data set into classes. This algorithm plots each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Support vectors are data points that are closer to the hyperplane. Using these support vectors, we maximize the margin of the classifier.

KNN: K Nearest Neighbour is used for both classification and regression problems. It is more commonly used in classification problems in. In KNN algorithm, we store all available cases and categorizes new cases by a majority vote of its k neighbors. The case is given to the class is most common amongst its K nearest neighbors. It measured by a distance function.

Naïve Bayes: Naive Bayes method is a classification technique based on Bayes' theorem. It is an assumption of independence between predictors. In simple term, a Naive Bayes classifier predicts that the presence of a particular feature in a class is unrelated to the presence of any other feature. [12]

IV. DATA COLLECTION AND PREPROCESSING

The original data has been collected from Bhartiya Skill Development University for the students enrolled for the year 2018-19 and 2019-20. It is a dataset of approximately 300 records. The original data includes the students' personal, Acedmic and Financial details. It has 21 labels.

Table I. Student Dataset

Attributes	Data Type	Description
Age	Numeric	Student Age
Gender	Nominal	Student Gender
Fedu	Nominal	Father Education
Foccu	Nominal	Father Occupation
Medu	Nominal	Mother Education
Moccu	Nominal	Mother Occupation
Fincome	Numeric	Family Income
NOS	Numeric	No. of Siblings
Address	Nominal	Belongs to
MoE	Nominal	Medium of Education

10th	Numeric	10th Marks
12th	Numeric	12th Marks
StIntforJob	Nominal	Student Interest for job
Health	Nominal	Health

ST	Numeric	Study Time in week
STS	Numeric	Study time on Sunday
IntInflence	Nominal	Internet Influence while choosing career
AreaofInterest	Nominal	Area of student
Why this course	Nominal	Why choose this course
Branch	Nominal	Course

Data Preprocessing is a technique that is used to convert the raw data into a clean and standard dataset. Datasets are required to be clean in order for the algorithm to give accurate results. We have applied the following procedure on raw data for preprocessing.

- Conversion of nominal values into numeric values.
- Irrelevant attributes like student name, DOB, address are not helpful in prediction, hence is removed from the training data set.
- Redundant attributes mostly give the identical information, hence is also eliminated.

V. METHODOLOGY

In this paper, we compared the efficiency of different Machine Learning classification algorithms on a particular dataset. The prediction is made on the basis of the physical parameters of the student. In validation, we check whether the predictions are accurate or not. We can plot the results and compare them with the actual values, i.e., calculate the distance between the predictions and actual values. Lesser this distance, more accurate will be the predictions.

Python is a popular platform used for scientific research and development of production systems. It is a language with number of modules, packages and libraries. Python and its libraries like NumPy, SciPy, Scikit-Learn, Matplotlib are used in data science and data analysis. They are also broadly used for creating scalable machine learning algorithms. Python implements machine learning techniques such as Classification, Regression, Recommendation, and Clustering [13].

After data preprocessing, a dataset is created by splitting it into two parts: 80% of which is used to train the models and 20% is held back as a test dataset. Both training dataset and test dataset are One-hot encoded, i.e., nominal variables are converted into numerical form to be provided to different Machine Learning algorithms for effective prediction. K is Chosen for K-fold cross validation to estimate the accuracy of different models.

V. EXPERIMENTAL RESULTS

The experiment was run on a system with Microsoft Windows 10 operating system with the configuration of 64GB RAM and 4 Intel cores. Tool python3 was used to run different Machine Learning algorithms. matplotlib library was used to visualize the inner working of the model. We used the accuracy measure for evaluating the quality of the classifier. The purpose of accuracy is to achieve a higher value. We used five different machine learning algorithm, Decision Tree, Naive Bayes, Logistics Regression, K-nearest neighbor and Support Vector Machine on the same dataset. These classification algorithms were run in python with a cross validation of 10 folds. The final classification accuracy is considered and compared with each other.

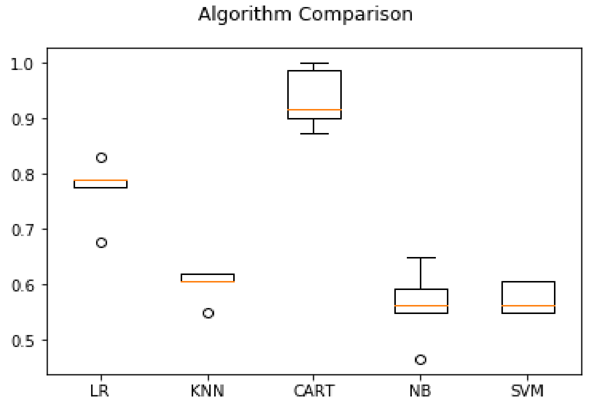

Figure 1. Algorithm Comparison

A whisker plot is used to show the accuracy scores of each cross validation of 10 folds for each Machine Learning algorithm, as given in Fig. 1.

Table II. Accuracy Comparison

Classifier	Accuracy
Logistics Regression (LR)	0.771831
K-Nearest Neighbor (KNN)	0.600000
Decision Tree (CART)	0.935211
Naïve Bayes (NB)	0.563380
Support Vector Machine (SVM)	0.574648

It is seen that the Decision Tree is provides the highest accuracy.

VI. CONCLUSION

Machine Learning algorithms have beeb used in this paper for predicting the career path of students. We compared the performance of five different Machine Learning techniques in academics' context. These algorithms are Logistics Regression, Naïve Bayes, K-Nearest Neighbor, Decision Tree, Support Vector machine. It showed that the accuracy of the Decision Tree is better than other classifiers in our domain.

In ML, there is no specific model or algorithm which can give 100% result for particular dataset. We need to understand the data before we apply any algorithm and build our model.

REFERENCES

1. Mccrea, N., "An Introduction to Machine Learning Theory and Its Applications: A Visual Tutorial with Examples", https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer

2. Vaidu, G., and Sornalakshmi, K., "Applying Machine Learning Algorithms for student employability prediction using R," International *Journal of Pharmaceutical Sciences Review and Research*, pp. 38-41, 05, March 2017. [Online]. Available: http://globalresearchonline.net/journalcontents/v43-1/11.pdf
3. Iqbal, Z, Qadir, J., and Kamiran, F., "Machine Learning based student grade prediction: A case study," 17 Aug 2017. [Online]. Available: https://arxiv.org/pdf/1708.08744.pdf
4. Kim. B, Vizitei, E., and Ganapathi, V., "GritNet: Student performance prediction with Deep learning," 19 Apr 2018. [Online]. Available: https://arxiv.org/abs/1804.07405
5. Xu, J., Horoon, K., and Schaar, V., "A Machine Learning Approach for Tracking and predicting student performance in degree program," *IEEE Journal of Selected Topics in Signal Processing*, Vol 11, pp. 742-753, Aug. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7894238/
6. Pojon Murat, "Machine Learning to predict student performance," 2017. https://tampub.uta.fi/bitstream/handle/10024/101646/GRADU-1498472565.pdf
7. Singh, M., and Singh, J., "Machine Learning Techniques for prediction of subject scores: A comparative study", International *Journal of Computer Science and Network*, Vol 2, issue 4, pp. 77-80, August 2013. [Online]. Available: https://pdfs.semanticscholar.org/2368/3634d0999020d6a90bf79fa605ceebe90891.pdf
8. BendengnuKsung, and Prabhu, P., "Students performance prediction using Deep Neural Network," *International Journal of Applied Engineering Research*, Vol 13, Number 2, pp. 1171-1176, 2018. [Online]. Available: https://www.ripublication.com/ijaer18/ijaerv13n2_46.pdf
9. Pushpa, S., Manjunath, T., Mrunal, T., Singh, A., and Suhas, C., "Class result prediction using machine learning," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Bangalore, 2017, pp. 1208-1212.
10. Gerritsen L. and Conijn R., "Predicting student performance with Neural Networks," dissertation, Dept. Humanities, Tilburg University, The Netherlands, May 2017.
11. Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M., "Perspectives to Predict Dropout in University Students with Machine Learning," *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, San Carlos, 2018, pp. 1-6.
12. https://www.dezyre.com/article/top-10-machine-learning-algorithms/202
13. Python Introduction, https://www.w3schools.com/python/python_intro.asp

AUTHORS PROFILE

Shashi Sharma, She has more than 9 years of experience in academics. She has 12 publications in her national and international journals. She guided M.Tech dissertations. Her teaching interest include Machine Learning, Statistics with R, programming in C, C++, Design and analysis of algorithm etc. She has done her M.Tech in computer Science with hons. degree.

Sunil Kumar Pandey, He has more than 16 years of experience including Academics and Industry. He is a Certified professional of ERP/SAP, Six Sigma, SPSS, Minitab, Big Data & Cloud Computing (IBM), Moodle LMS, Digital Marketing and SEO. He has authored books on Oracle, Data Structures and IT Trends. He has over 50 research papers and articles publications to his credit. He is member of International Societies-NSBE-Alexandria, USA, IEEE Computer Society-USA, Indo- Nepal Intellectual Society-Nepal, IAENG-Hongkong, IFETS-USA. Apart from it, he is the technical author & reviewer of the Asia's #1 IT Magazine –Developer IQ since 2005.

Prof.(Dr.) Kumkum Garg, She is a gold-medalist alumnus of IIT Roorkee and has a doctorate from Imperial College London, UK. She has been a faculty member at IIT Roorkee for 34 years. She served as Head of its Information Superhighway Centre, from 2005 to 2006. Dr. Garg had also served as the Director, Manipal Institute of Technology, Manipal University in Karnatakabetween 2010-12 and founding Dean, Faculty of Engineering and Pro Vice Chancellor at Manipal University Jaipur, from 2012 to 2017. Currently, she holds the position of Dean, Faculty of Informatics and Automation at Bhartiya Skill Development University, Jaipur.

Dr. Garg is a Senior Member of IEEE, Fellow of Institution of Engineers (I) and Life Member of various Professional Societies, including ISTE, SMATAC and ISCEE. She has over 47 years of experience of teaching and research and has successfully undertaken a number of Sponsored Research and Consultancy projects from CISCO USA, AICTE, DRDO, MIT GoI and Govt. of Uttarakhand. She has been on the panel of experts for Computer Science and Engineering/IT education and state-wide area networking for various Government and private organizations.

She has served as Guest Editor, Hindawi publications USA, in 2015 and as UGC expert nominee on UGC SAP committee (DRSII) for Dept. of Computer Science, Punjabi University, Patiala. Currently, she is a Member of the Editorial Board of the Journal of Institution of Engineers (India), Series B.