JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Review on Student Career Prediction Using Machine Learning

Umesh Kumar Sah[1], Mrs. Awantika Singh[2]

[1]M.Tech Scholar, [2]Assistant Professor

umeshsah95564@gmail.com , awanisingh90@gmail.com

Computer Science & Engineering Department, Kalinga University, Raipur (C.G.), India

Abstract

There are number of good schools and colleges in India. But most of the students are dropping their education because of various reasons. There are many reasons, some of the students have some financial problem with their family, some of the students don't have interest towards their next level of education, some think about the gender and some rural areas don't have good schools and educators. So this proposed method deals whether the students will be going to the next level of higher education or not. This can be decided with the concepts of machine learning which is the subset of artificial intelligence. Machine learning is made up with the concepts of Mathematics and Science. This paper deals with the student's career prediction by using various machine learning algorithms like Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Adaboost. Machine learning algorithms are implemented by using Python programming language.

Keywords: Machine Learning Classifier, Support Vector Machine, Adaboost, Random Forest, Decision Tree.

1. Introduction :

Traditionally student's career can be predicted by using questionnaire. But this method takes lot of time. Now, computing technologies play important role in various fields. Machine learning is one of newest computing technique. In this digital world Machine learning is used in various fields and industries such as image processing, classification, clinical analysis, regression and more and more. It has the capability of developing and studying automation without being explicitly. Machine learning is of three types i.e. supervised machine learning, unsupervised machine learning and reinforcement machine learning algorithms. In simple words Machine learning is the science of learning and behaving like humans. It is very important to analyze the ability of the students and they should be directed in the right path way. In this research work the concepts of machine learning are applied to detect the next level of education of the students. This prediction is important for all type of educational institutions, recruiters and so on. Based up on the result of this prediction accuracy, the educational institutions find the people with low performance and provide the proper training to them to improve their performance. Job providing companies also spend lot of amount for selecting a qualified candidate. The output of the prediction model is also used to find the status of the students, if they are interested to go to the job or they are interested to do their higher studies. This research work mainly focuses on the career prediction of undergraduate level students. Machine learning algorithms such as SVM, DT, RF and Adaboost classifiers are used to construct the model. Among the above classifiers RF produces better result. These classifiers are implemented with the help of python programming language, because most of the real time problems are easily implemented by this programming language. Next section deals with the views and approaches that are used by various authors in career prediction research domain.

2. Literature Review :

A recent fact provides the information using student's data based on their behavioural aspects to forecast the career path. Min Nie et al., proposed a novel model known as ACCBOX (Approach Cluster Centers Based On XGBOOST) to forecast student's career. The final result clearly states that the current method is better than other methods of prediction. This model uses 13 behavioural data which are collected from 4000 students [1]. Mining student's educational data is also one of the important tasks in education field. In the beginning days data mining methods were used in education field by using less number of arguments, because low record maintenance in concern institutions. Recently the large volume of data can be stored on the basis of student. In India 0.3 % people only move forward from their PG level to research level. This prediction task evaluates performance of the students by using various arguments and the students are classified as low, high and medium type. To execute this process the authors K. B. Eashwar et al., combined SVM and K-means methods. A SVM concept is used for classification purpose and K-mean technique is mainly used for clustering the student's data [2].

3. Previous Research :

In the past two decades, we have seen a large number of high-quality works using students' academic performance and learning behavioural data to predict outcome variables, such as standardized test score, dropout from school, college enrollment, and major choice. For example, Feng, Heffernan, and Koedinger (2009) investigated how students' interaction data extracted from the Assessment platform can be used to reliably evaluate students' math 19 Journal of Educational Data Mining, Volume 12, No 2, 2020 proficiency. They were especially interested in building features related to student help seeking behaviours and used the Bayesian Information Criterion (BIC) to compare linear regression models with different groups of predictors. They showed that students' end-of-year exam scores can be better predicted by leveraging the interaction data that reflect assistance requirement, effort, and attendance.

Instead of using traditional explanatory variables in college enrollment research, such as family background, career aspiration, and assessment scores, San Pedro, Baker, Bowers, and Heffernan (2013) studied how student online learning behaviours observed in middle school related to their college choice. They built a logistic regression model using automatically generated affect and engagement features to achieve decent accuracy at predicting college attendance. Their study was further extended to predicting STEM and Non-STEM college major enrollment by San Pedro, Ocumpaugh, Baker, and Heffernan (2014).

Pardos, Baker, San Pedro, Gowda, and Gowda (2014) also studied the Assistments system, but they focused on the correspondence between student affect and behavioural engagement and scores on a high-stakes math exam. They constructed a set of affect and engagement behaviour detectors using eight machine learning models to estimate the probability that a student is in a state of boredom, engaged concentration, confusion, and so on. Further they built a model to predict students' math exam scores and showed that the constructed detectors helped the model achieve high prediction accuracy.

Knowles (2015) described how to create a state wide dropout early warning system that can accurately predict the likelihood of graduation for high school students in the State of Wisconsin. The paper properly demonstrated the workflow of the whole system, from data cleansing to model training and searching. To balance the tradeoff between the correct classification of dropouts and false alarms, the receiver-operating characteristics (ROC) metric is used to identify the best models from a large collection of candidates, from linear logistic regression models to complex nonlinear models, such as support vector machines. This work was also implemented in the open source R package, EWStools (Knowles, 2014).

Baker, Berning, Gowda, Zhang, and Hawn (2019) presented a case study on automatically identifying students that have a high risk of dropping out of high school, using data on students' discipline, attendance, course-taking, and grades. The logistic regression model used in the study helped the authors not only select students at risk, but also found which factors played the greatest roles in prediction, which provided information to educators that can be used in individualized interventions.

4. Methodology :

Machine Learning is a technique in which the machines are trained in such a way that it gains the ability to respond to a particular input or scenario based on the past inputs it has learnt. Simply it the gives computers the ability to learn by using statistical techniques. With the help of Machine learning the computers gains ability to act without explicitly being programmed. This aims at reducing the human involvement in the machine dependable problems and scenarios. This helps in resolving very complex tasks and problems very easily and without involving much human labour. Various applications of machine learning include NLP, classification, prediction, image recognition, medical diagnosis, algorithm building, self-driving cars and much more. In this paper classification and prediction are being done. Let us see the details of classification and prediction. Most of the problems in machine learning can be solved using supervised and unsupervised learning. If the final class labels are previously known and all the other data items are to be assigned with one of the available class labels, then it is called as supervised. And if the final output classes and sets are not known and it is done by identifying the similarity between data point and their characteristics and then they are made into groups based on these characteristics then it is called as unsupervised. Classification falls under supervised. On the basis of the properties of the given Input parameters a predefined class label is assigned. There are other alternatives like clustering and regression as well. Based on the type of problem the appropriate model is chosen.

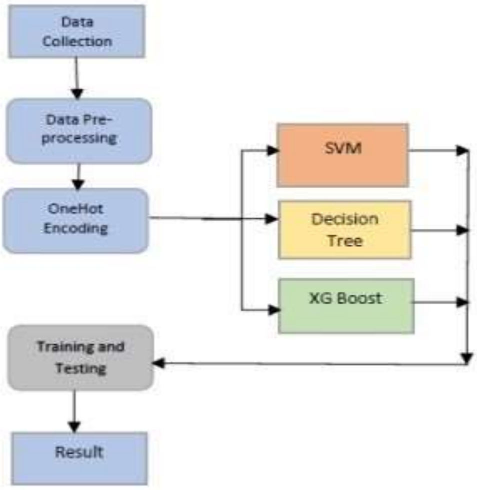

Fig.1 : Process Flow Diagram of Proposed System.

5. Implementation :

A. Data Collection:

Data Collection is one of the most important tasks for any machine learning projects. Because the input we feed to the machine learning algorithms is data. So, the algorithms efficiency and accuracy depends on the correctness and quality of data collected. So as proper the data accurate will be the output. For student career prediction many parameters are required like students academic scores in various specializations, subjects, programming and analytical capabilities, memory, personal details like relationship, interests, competitions, hackathons, sports, workshops, certifications, books interested and many more. As all these factors play vital role in deciding student's progress towards a career area and all these are taken into consideration. Data is collected in many ways. Some data is collected from employees working in different organizations, some amount of data is randomly generated and other from college alumni database. Totally nearly 15 thousand records with 35 columns of data is collected.

B. Data Pre-processing :

Collecting the data is one task and making that data useful is another important task. Data collected from various sources will be in an unorganized format and there may be a lot of null values, invalid data values and unwanted data. Cleaning of all these improper data and replacing them with appropriate or approximate data and removing null and missing data and replacing them with some fixed alternate values are the basic steps in pre-processing of data. Even data collected may completely contain garbage values. It may not be in exact format or way which is meant to be. All such datas must be verified and replaced with alternate values to make data meaningful and useful for further processing. Datas must be kept in a organized format.

C. OneHot Encoding :

OneHot Encoding is a technique by which categorical values present in the collected data are converted into numerical or other ordinal format so that they can be provided to machine learning algorithms and get better results of prediction. OneHot encoding simply transforms categorical values into a form that best fits as input to be feeded to various machine learning algorithms. This algorithm works good with almost all machine learning algorithms. Few algorithms like random forest handle categorical values properly. In such cases this encoding is not required. Process of OneHot encoding seems to be difficult but most modern day machine learning algorithms take care of that. The process is easily explained here: For example, in a set of data if there are values like yes and no, integer encoder assigns values like 1 and 0 to them. This process can be followed as long as we continue the fixed values as 1 for yes and 0 for no. As long as we assign or allocate these fixed numbers to these particular labels, this is called as integer encoding. But here consistency is very important because if we invert the 12 | Page encoding later, we should get back the labels correctly from those integer values especially in the case of prediction. Next step is to create a vector for each integer value. Let us suppose this vector is binary and has a length of 2 for the two possible integer values. The label 'yes' encoded as 1 will then be represented with vector [1,1] where the zeroth index is given the value 1. Similarly label 'no' encoded as '0' will be represented like [0,0] which represents the first index is represented with value 0. For example, [pillow, rat, fight, rat] becomes [0,1,2,1]. This is imparting an ordinal property to the variable , i.e. pillow < ret < fight. As this is ordinally characteristic and is usually not required and so OneHot encoding is required for correct representation of distinct elements of a variable. It makes representation of categorical variables to be more expressive and meaningful.

D. Machine Learning Algorithms :-

1. SVM :

SVM denotes Support Vector Machine. It is a supervised machine learning algorithm which is generally used for both regression and classification type of problems. The main applications of SVM can be found in various classification problems. The typical procedure of the algorithm is first each data item is to be plotted in a n-dimensional space, where n is the number of features and the value of each feature being the value of that particular coordinate. Next step is to classify the hyper-plane that separates the two classes very finely.

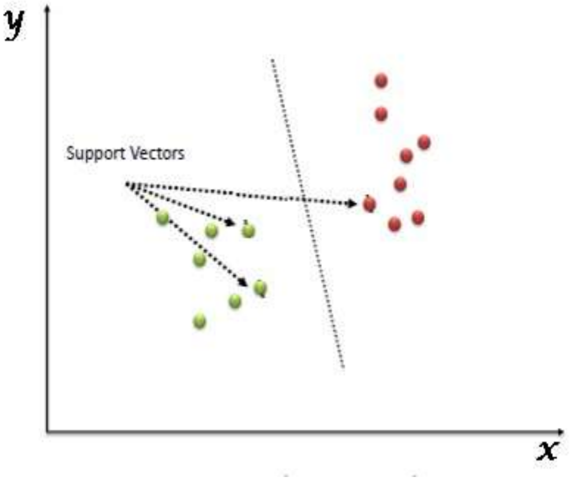

Fig. 2 : Support Vector Machine

SVM algorithms are practically implemented using kernels. There are three types of SVM's. In linear SVM hyperplane is calculated or found by transforming the problem using linear algebra. The concept is that SVM can be rephrased by using the inner product of two observations. The sum of the multiplication of each pair of inputs is called inner product of two vectors. The equation for dot product of a input xi and support vector xi is: f(x) = B0 + sum(ai * (x,xi)). Instead of using the dot-product, a polynomial kernel can be used, for example: K(x,xi) = 1 + sum(x * xi)^d And not only that a more complex radio kernel is also there. The general equation is:

K(x,xi) = exp(-gamma * sum((x – xi^2)).

2. XG Boost :

XGBoost denotes eXtreme Gradient Boosting. XGBoost is implementation of gradient boosting algorithms. It is available in many forms like tool, library etcetera. It specifically focuses on model performance and computational time. It reduces the time and lifts the performance of the model greatly. It's implementation has the features of scikit-learn and R implementations and also have a newly added features like regularization.

Regularized gradient boosting means gradient boosting with both L1 and L2 type regularizations. The main best features that the implementation of the algorithm provides are: Automatic handling of missing values with sparse aware implementation, and it provides block structure to promote parallel construction of tree and continued training which supports further boost an already fitted model on the fresh data. Gradient boosting is a technique where new models are created that can predict the errors or remains of previous models and then added together to make the final prediction. They use gradient descendent algorithms to reduce loss during addition of new models. They support both classification and regression type of problems. In the training part generally an objective function is defined. For example,
obj=i=1 Σl(yi,yi^(t))+ i=1 Σ Ω(fi)

3. Decision Tree :

Decision Tree is an extremely popular and one of the simple and easy technique to implement machine learning classification problems. Decision trees are basic foundation for many advanced algorithms like bagging, gradient boosting and random forest. The XG Boost algorithm mentioned above is the advanced version of this general decision tree. The commonly used decision trees are CART,C4.5,C5 and ID3.A node denotes a input variable (X) and a split on that variable, assuming the variable is numerical. The leaf, also called the terminal nodes of the tree possess an output variable (y) which is vital for prediction. The typical scenario that a decision tree follows is first select a root node. Then calculate information gain or entropy for each of the nodes before the split. Then select the node that has more information gain or less entropy. After then split the node and reiterate the process. The process is iterated until and unless there is no possibility to split or the entropy is minimum. Entropy is the measure of uncertainty or randomness of data. Information gain is the measure of how much entropy is reduced before to after split.

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Conclusion :

A more powerful web application or mobile application can be developed where inputs will not be given directly, instead student parameters will be taken by evaluating students through various evaluations and examining process. Technical, logical, analytical, psychometry, memory based, and general awareness, interests and skill based tests can be designed and parameters can be collected through them so that results certainly will be more accurate and the system will be more reliable to use.

References :

[1] P.KaviPriya, "A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering.

[2] Ali Daud, Naif Radi Aljohani, "Predicting Student Performance using Advanced Learning Analytics", 2017 International World Wide Web Conference Committee (IW3C2).

[3] Marium-E-Jannat, SaymaSultana, Munira Akther, "A Probabilistic Machine Learning Approach for Eligible Candidate Se-lection", International Journal of Computer Applications (0975 – 8887) Volume 144 – No.10, June 2016.

[4] Sudheep Elayidom, Dr.Sumam Mary Idikkula, "Applying Data mining using Statistical Techniques for Career Selec-tion", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.

[5] Dr.Mahendra Tiwari, Manmohan Mishra, "Accuracy Estimation of Classification Algorithms with DEMP Model", International Journal of

Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013.

[6] Ms.Roshani Ade, Dr. P. R. Deshmukh, “An incremental ensemble of classifiers as a technique for prediction of student’s career choice”, 2014 First International Conference on Networks & Soft Computing.

[7] Nikita Gorad, Ishani Zalte, “Career Counselling Using Data Mining”, International Journal of Innovative Research in Computer and Communication Engineering.

[8] Bo Guo, Rui Zhang, “Predicting Students Performance in Educational Data Mining”,2015 International Symposium on Educational Technology
[9] Ali Daud , Naif Radi Aljohani , “Predicting Student Performance using Advanced Learning Analytics” .

[10] Rutvija Pandya Jayati Pandya, “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning“, International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015.