STUDENT CAREER PREDICTION SYSTEMUSING MACHINE LEARNING

Mr.Gsai Krishna[1], K.ajay[2], raghavendrasrinivas[3], prem kiran[4],vinayreddy CH[5].

Associate Professor[1], Students[2,3,4,5].

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING[1,2,3,4,5.]

MALLA REDDY INSTITUTE OF TECHNOLOGY AND SCIENCE, HYDERABAD, INDIA.

Email ID:gasikrishna5555@gmail.com, ajaykyadaveni2001@gmail.com[2] , remellaraghavendra6@gmail.com[3],premkiran3107@gmail.com[4],

vinayreddyvinay4518@gmail.com[5].

1.Abstract:

There are many good schools and colleges in India. But most of the students are dropping their education Becauseof various reasons. There are a lot of reasons, some of the students have some financial problem with their family, some ofthe students don't have interest towards their next level of education, and some students think about gender and some aboutrural areas don't have good schools and educators. In today's world choosing an appropriate career path is one of the mostimportant decisions but with the increase in the number of career options and opportunities, it makes this decision evenmore difficult for the students. Different people suggest different career paths to the student but at last the student has toselect their career themselves. According to the survey conducted by the Council of Scientific and Industrial Research's(CSIR), about 40% of students are confused about their career selections. This may lead the students to wrong careerselection and then working in an area which was not meant for them, this leads their career in wrong path, and this maynot be good for their future career. Therefore, it is very important to take the right decision regarding your future career at the right time. So, this proposed method deals with 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 ismade up of the concepts of Mathematics and Science. This paper deals with the student's career prediction by usingvarious 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 students'careers waspredicted by questions and answers. But this method takes a lot of time. Now, computing technologies play important role in various fields. Machine learning is one of the newest computing techniques. In this digital world machine learning is used in various fields and industries such as image processing, classification, clinical analysis, regression andmore and more. It has the capability of developing and studying automation without being explicit. Machine learning is of threetypes i.e supervised machine learning, unsupervised machine learning and reinforcement machine learning algorithms. In simplewords Machine learning is the science of learning and behaving like humans. It is very important to analyze the ability of thestudents and they should be directed in the right pathway. This research work includes the concepts of machine learning which areapplied to detect the next level of education of the students. This prediction is very important for almost all types of educationalinstitutions, recruiters and so on. Based on the result of this prediction accuracy, the educational institutions find the people withlow 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 theyare interested in going to the job or they interested to do their higher studies. This research work mainly focuses on the career predictionof 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 results. These classifiers are implemented with the help of pythonprogramming language, because most of the real time problems are easily implemented by using this programming language. Nextsection deals with the views and approaches that are used by various authors in career prediction research domain.

2. Literature Review:

A recent fact provides information using student's data based on their behavioral aspects to forecast the career path. Min Nie etal. 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 behaviouraldata which are collected from 4000 students [1]. Mining student's educational data is also one of the important tasks in the educationfield. In the beginning days data mining methods were used in the education field by using a smaller number of arguments, because lowrecord 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 usingvarious arguments and the students are classified as low, high and medium type. To execute this process the authors K. B. Eashwaret al., combined SVM and K-means methods. A SVM concept is used for classification purposes and K-mean technique is mainlyused for clustering the student's data [2]. Predicting the student performance level is one of the important tasks in education domain.

Data mining concepts are used to predict the student's performance by using various parameters. Ankita Kadambande et al., uses

semantic rules and SVM concepts for the predictions. Semantic rules are used to improve the quality of educational content andconvey educational action to every student. Here the authors help the students by providing better suggestions and recommendationsfor improving student's performance level in forthcoming exams. This system will provide help to low level and high-level studentsand also to increase the student's interest in their education. The main purpose of this research work is to increase the quality oflearning measures and support the students by forecasting their academic level which will help the students to a great extent[3].

3. Previous Research:

In the past two decades, we have seen a large number of high-quality works using students' academic performance and learningbehavioural data to predict outcome variables, such as standard test score, dropout from school, college enrollment, and majorsubject choices. For example, Feng, Heffernan, and Koedinger (2009) investigated how students' interaction data extracted fromthe Assistment 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 theBayesian Information Criterion (BIC) to compare linear regression models with different groups of predictors. They presented thatstudents' 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 descriptive variables in college enrollment research, such as family background, career aspiration, andassessment scores, San Pedro, Baker, Bowers, and Heffernan (2013) studied how student online learning behaviours observed inmiddle 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 predict STEMand 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 correspondencebetween student affect and behavioural engagement and scores on a high-stakes math exam. Using eight machine learning models,they created a set of affect and engagement behaviour detectors to estimate the probability that a student is in a state of boredom,engaged concentration, confusion, and so on. They further created a model to predict students' math exam scores and showed thatthe constructed detectors helped the model achieve high prediction accuracies.

Knowles (2015) described how to create a state-wide dropout early warning system that can accurately predict the probability ofgraduation for high school students in the State of Wisconsin. The paper systematically demonstrated the workflow of the wholesystem, from data cleansing to model training and searching. To balance the trade-off between the correct classification of dropoutsand false alarms, the receiver-operating characteristics (ROC) metric is used to identify the best models from a large collection ofaspirants, from linear logistic regression models to complex nonlinear models, such as support vector machines. This work was alsoimplemented 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 riskof 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 largest roles inprediction, which provided information to educators that can be used in individualized involvements.

4. Problem Identification:

Selecting an appropriate career is one of the most important decisions and with the increase in the number of career paths andopportunities, has made this decision quite difficult for them. According to the survey conducted by the Council of Scientific andIndustrial Research's (CSIR), about 40% of students are confused about their career options. This may result in the selection of the wrongcareer and then working in a field which was not meant for them, thus reducing the productivity of human resourcess. Therefore, itis quite important to take the right decision regarding the career at an appropriate age to prevent the consequences that result due towrong career selection. This system is a web application that will help students studying in high schools to select a course for theircareer. The system will suggest the student, a career option based on their personality, skills, interest and their capacity to take up
the course.

5. Methodology:

Machine Learning is a technique in which the machines are trained in such a way that it achieves the ability to respond to specificinput or scenario based on the past knowledge it has learnt. Simply, it gives computers the ability to learn by using statisticaltechniques. With the help of Machine learning the computers gains ability to act inspite of being explicitly programmed. This aimsat reducing the human involvement in machine dependable problems and scenarios. This helps in solving very difficult tasksand problems very easily and without involving much human effort. Various applications of machine learning include NLP,classification, image recognition, prediction, medical diagnosis, algorithm building, self-driving cars and much more. This paperinvolves classification and prediction. Let us see the details of classification and prediction. Most of the problems in machinelearning can be resolved using supervised and unsupervised learning. If the final class labels are previously known and all the otherdata items are to be assigned with one of the available class labels, then it is called supervised. And if the final output classes andsets are not known and it is done by identifying the similarity between data point and their characteristics and then they are madeinto groups based on these characteristics then it is called unsupervised. Classification falls under supervision. On the basis of the
andregression as well. Based on the type of problem the appropriate model is chosen.

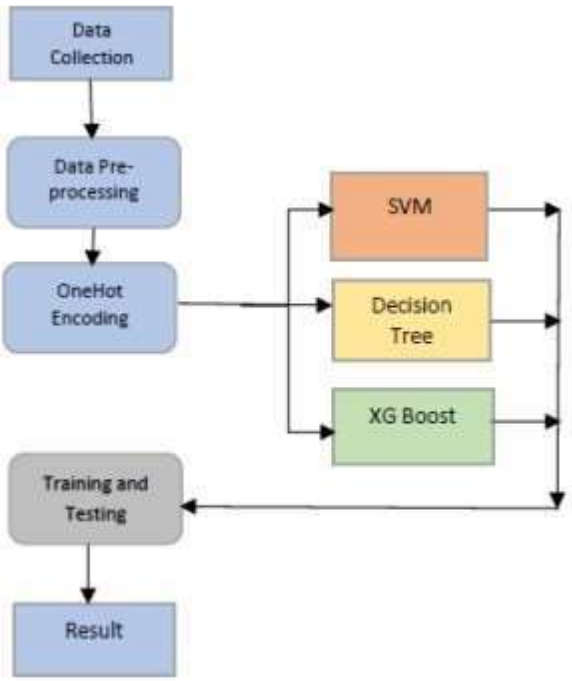

Fig.1: Process Flow Diagram of Proposed System.

Here algorithms like SVM, OneHot encoding, Decision tree and XG boost are used. After training and testing the data with thesewe consider the most accurate results given by the algorithm for our further processing. So, the initial task done is predicting the outputusing all algorithms proposed above and later analyzing the results and there after we continue with the most accurate algorithm.So, this paper deals with various advanced machine learning algorithms that involve classification and prediction and are used toimprove the accuracy for better prediction, reliability and analyzing these algorithms performance.

6. Implementation:

A. Data Collection:

Data Collection is one of the most important tasks for any project of machine learning. Because the input we feed to the machinelearning algorithms is data. So, the algorithms efficiency and accuracy depend on the correctness and quality of data collected.Accurate the data accurate will be the output. For student career prediction many parameters like students' academic scores in variousspecializations, subjects, programming and analytical capabilities, memory, personal details like relationship, interests,competitions, hackathons, sports, workshops, certifications, books interested and many more are required. All these factors playan important role in deciding a student's progress towards a career area and all these are taken into consideration. Data is collected inmany ways. Some data is collected from employees working in different organizations, some amount of data is randomly generatedand other from college alumni database. Totally nearly 15 thousand records with 35 columns of data are 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 inan unorganized format and there may be a lot of null values, invalid data values and unwanted data. Cleaning of all these improperdata and replacing them with appropriate or approximate data and removing null and missing data and replacing them with somefixed alternate values are the basic steps in pre-processing of data. Even data collected may completely contain garbage values. Itmay not be in the exact format or way which is meant to be. All such data must be verified and replaced with alternate values to makedata meaningful and useful for further processing. Data must be kept in a organized format.

C. OneHot Encoding:

OneHot Encoding is a technique by which categorial values present in the collected data are converted into numerical or otherordinal format so that they can be provided to machine learning algorithms and get better results of prediction. OneHotencodingsimply transforms categorial values into a form that best fits as input to be feeded to various machine learning algorithms. Thisalgorithm works good with almost all machine learning algorithms. Few algorithms like random forest handle categorical valuesproperly. In such cases this encoding is not required. The process of OneHot encoding seems to be difficult but most modern daymachine learning algorithms take care of that. The process is easily explained here: For example, in a set of data if there are valueslike yes and no, integer encoder assigns values like 1 and 0 to them. This process can be followed as long as we continue the fixedvalues as 1 for yes and 0 for no. As long as we allocate or assign these fixed numbers to these particular labels, this is called integer encoding. But here consistency is very important because if we invert the 12 | P a g e encoding later, we should get back thelabels 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 thenbe represented with vector [1,1] where the zeroth index is given the value 1. Similarly, label 'no' encoded as '0' will be representedlike [0,0] which represents the first index is represented with value 0. For example, [pillow, rat, fight, rat] becomes [0,1,2,1]. Thisis imparting an ordinal property to the variable, i.e. pillow < ret < fight. As this is ordinally characteristic and is usually not required,OneHot encoding is required for correct representation of distinct elements of a variable. It makes representation of categorialVariables 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 regressionand classification type of problems. The main applications of SVM can be found in various classification problems. The typicalProcedure of the algorithm is first each data item is to be plotted in a n-dimensional space, where n is the number of features andthe value of each feature being the value of that particular coordinate. Next step is to classify the hyper-plane that separates the twoclasses very finely.

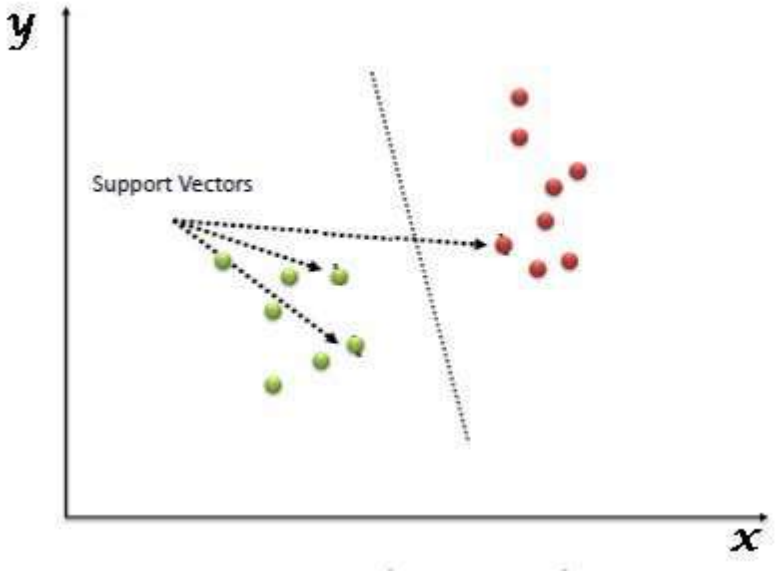

Fig. 2: Support Vector Machine

SVM algorithms are practically implemented using kernels. There are three types of SVM's. In linear SVM hyperplane is calculatedor found by transforming the problem using linear algebra. The concept is that SVM can be rephrased by using the inner productof two observations. The sum of the multiplication of each pair of inputs is called the inner product of two vectors. The equation fordot 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 kernelcan 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 isK(x,xi) = exp(-gamma * sum((x – xi^2)).

2. XG Boost:

XGBoost denotes eXtreme Gradient Boosting. XGBoost is an implementation of gradient boosting algorithms. It is available in manyforms like tools, library etcetera. It specifically focuses on model performance and computational time. It reduces the time and liftsthe performance of the model greatly. Its implementation has the features of scikit-learn and R implementations and also have newly added features like regularization. Regularized gradient boosting means gradient boosting with both L1 and L2 typeregularizations.
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 trainingwhich supports further boost an already fitted model on the fresh data. Gradient boosting is a technique where new models arecreated that can predict the errors or remains of previous models and then added together to make the final prediction.
They usegradient descendent algorithms to reduce loss during addition of new models.
They support both classification and regression typeof problems. In the training part generally, an objective function is defined. For example,

$$obj=i=1\ \Sigma l(yi,yi^{(t)})+ i=1\ \Sigma\ \Omega(fi)$$

3. Decision Tree:

Decision Tree is an extremely popular and one of the simple and easy techniques to implement machine learning classificationproblems. Decision trees are the 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 treesare 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, possesses an output variable (y) which is vital for prediction. The typical scenariothat a decision tree follows is first selecting 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. Then split the node and reiterate the process. Theprocess 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.

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

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

7. Result:

The data was trained and tested with all three algorithms and out of all SVM gave more accuracy with 90.3 percent and then the
XG Boost with 88.33 percent accuracy. As SVM gave the maximum accuracy, for all further data predictions SVM will be followed.
So, finally, we concluded that SVM provides better result. So, this algorithm can be used in the background and the new

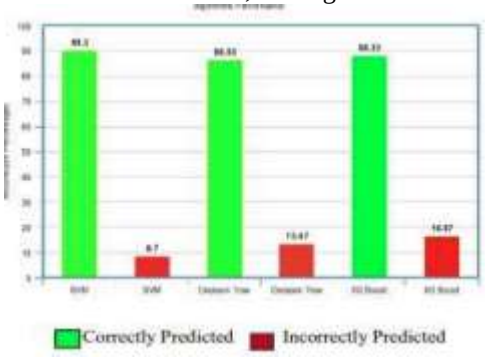

Fig. 3: Final Output Graph

8. Future Scope: 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 processes. 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 Selection", 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 Selection", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.

[4]. 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.

[5]. 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.

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

[7]. Bo Guo, Rui Zhang, "Predicting Students Performance in Educational Data Mining", 2015 International Symposium on Educational Technology.

[8]. 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.