Carrier Finder using Machine Learning

1* Shivam Jaiswal

Computer Science And Technology

Shri Ram Institute Of Technology Jabalpur(M.P.), India

Shivamsrit980@gmail.com

2* RamJiDas Gautam

Computer Science And Technology

Shri Ram Institute Of Technology Jabalpur(M.P.), India

ramjidasgautamm@gmail.com

3* Rohit Tiwari

Computer Science And Technology

Shri Ram Institute Of Technology Jabalpur(M.P), India

rt084705@gmail.com.

Prof Akriti Pathak

Computer Science and Engineering

Shri Ram Institute of Technology Jabalpur

Prof Rajendra Arakh

Computer Science and Engineering

Shri Ram Institute of Technology Jabalpur

Abstract

To determine the career field that best suits their interests and abilities, students should evaluate their strengths and identify their interests while pursuing their academic courses. This will help them perform better and foster their interests to enable them to pursue their desired careers. Moreover, career recommender systems assist recruiters in selecting candidates based on their performance and other evaluations. They thoroughly evaluate candidates in all relevant areas to determine which job role suits them best. This paper primarily focuses on predicting computer career areas.

Keywords—*Student Career Prediction, Decision Tree, Machine Learning, SVM, One Hot Encoder, XG Boost*

Introduction

In today's world, competition is fiercely increasing daily. In the technological world of today, in particular, it is excessively heavy. In order to compete and accomplish the objective, students must be planned and coordinated from the very beginning of their schooling.

Therefore, it is critical to continuously assess their performance, determine their areas of interest, gauge their progress toward their objectives, and determine whether they are headed in the correct direction. This aids in their self-improvement, self-motivation to pursue a better career path in the event that their abilities fall short of their objectives, and self-evaluation prior to reaching the career peak point.

Additionally, recruiters assess candidates based on a variety of factors before deciding whether or not to hire them. If hired, they then locate the ideal position and field of expertise for the chosen candidate. Database administrators, business process analysts, developers, testing managers, networks managers, data scientists, and other positions are among the many different kinds of jobs. To be placed in any of these roles, there are prerequisites

that must be met. Thus, after considering the candidate's abilities, interests, and talents, recruiters match them with the ideal position. Because recommendations are made based on inputs as they are provided, these prediction systems make hiring very simple.

Various third-party performance evaluation portals, such as Co-Cubes and AMCAT, already use these kinds of prediction and recommendation systems for jobs and careers. They simply take into account the students' psychometry and technical prowess. These portals evaluate students technically and recommend jobs for both students and businesses based on how well they perform. However, in this case, a number of variables, such as students' athletic, intellectual, and Additionally taken into account are interests, skills, knowledge, competitions, and hobbies. Thirty-six parameters in total were considered as inputs after accounting for all the factors. Additionally, there will only be 15 job roles in total. Normal algorithms and typical programming cannot provide the best possible output classification and prediction due to the large number of final output classes and input parameters. Thus, sophisticated machine learning algorithms are employed, including SVM, Random Forest decision tree, OneHot encoding, and XG boost.

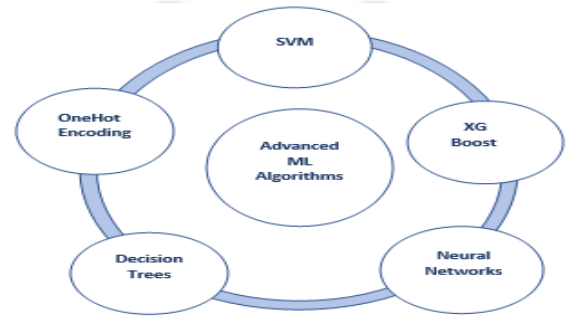


Figure 1: Overview of various Advanced Machine Learning Algorithms.

Literature review

Senior high school, also known as senior secondary school, is a higher education level under the K–12 Curriculum Program for students between the ages of 15 and 18, according to DepEd Memorandum No. 76's 2016 Senior High School Operations Manual. It usually occurs immediately following junior high school graduation and occurs prior to going to college or getting a job. Pupils focus on their academic or professional paths in order to prepare for college or the workforce. The aim of this educational phase is for the student to acquire more specialized knowledge and abilities. Accountancy, Business Management (ABM), Humanities and Social Sciences (HUMSS), Science, Technology, Engineering, and Mathematics (STEM), General Academic Strand (GAS), and other academic tracks that align with college and workforce demands requirements may include career pathways such as the Technical-Vocational-Livelihood (TVL) strand, which offers specialization in Agri-Fishery Arts, Information and Communications Technology (ICT), Industrial Arts, and Home Economics, as well as senior high school in some educational systems. These classes are designed to provide a strong foundation for students in order to get them ready to take on specific professional pathways or enroll in additional college courses [8].

A. The home economics strand focuses on occupations like housekeeping, fashion design, hospitality, cooking, serving, baking, crafts, massage therapy, wellness, travel services, theme parks, and ecotourism-related attractions [8]. It also concentrates on livelihood businesses like home management and fashion design.

PROPOSED METHODOLOGY

B. Research Design:- **In order to investigate the relationship between two or more quantitative variables from the same group of participants, this study uses a quantitative research method called descriptive correlation, which involves creating a questionnaire. By using surveys to gather information and interact with participants, the researchers are able to focus on similar score patterns rather than average differences.**

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.

- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter", not "webers/m2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm3", not "cc". (*bullet list*)

C. Respondents of the Study

The total number of respondents was Two Hundred Nineteen (219) out of 363 total population of the HE Grade 12 students of Muntinlupa National High School - Senior High School.

D. Sampling Technique

To choose participants from a particular population, this study used systematic sampling techniques with an odd and even scheme serving as the sample interval.

E. ***Research Instrument :-***For the purpose of gathering data for a study, the researchers created a survey questionnaire technique. The questionnaire examined the age, gender, and socioeconomic status of students while also analyzing their professional aspirations and abilities. Additionally, it evaluated the students' abilities and career choices for home economics coursework. The questionnaire was created following examining research findings and professional guidance

IMPLEMENTATION

F. Data Collection:

Collection of data is one of the major and most important tasks of any machine learning projects. Because the input we feed to the algorithms is data. So, the algorithms efficiency and accuracy depends upon the correctness and quality of data collected. So as the data same will be the output. For student career prediction many parameters are required like student academic scores in various subjects, specializations, programming and analytical capabilities, memory, personal details like relationship, interests, sports, competitions, hackathons, workshops, certifications, books interested and many more. As all these factors play vital role in deciding student's progress towards a career area, 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 collected through Linked In a pi , some amount of data is randomly generated and other from college alumni database. Totally nearly 20 thousand records with 36 columns of data is collected. The template is designed for, but not limited to, six authors.

G. Data Pre-processing:

Collecting the data is one task and making that data useful is another vital task. Data collected from various means will be in an unorganized format and there may be lot of null values, invalid data values and unwanted data. Cleaning all these 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 preprocessing of data. Even data collected may contain completely garbage values. It may not be in exact format or way that is meant to be. All such cases must be verified and replaced with alternate values to make data meaning meaningful and useful for further processing. Data must be kept in a organized format..

C. OneHot Encoding:

a) OneHot Encoding is a technique by which categorial values present in the data collected are converted into numerical or other ordinal format so that they can be provided to machine learning algorithms and get better results of prediction. Simply OneHot encoding transforms categorial values into a form that best fits as input to feed to various machine learning algorithms. This algorithm works fine with almost all machine learning algorithms. Few algorithms like random forest handle categorical values very well. In such cases OneHot encoding is not required.

Process of One Hot encoding may seem difficult but most modern-day machine learning algorithms take care of that. The process is easily explained here: For example, in a data if there are values like yes and no., integer encoder assigns values to them like 1 and 0. This process can be followed as long as we continue the fixed values for yes as 1 and no as 0. 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 encoding later, we should get back the labels correctly from those integer values especially in the case of prediction. Next step is creating 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 'yes' label encoded as 1 will then be represented with vector [1,1] where the zeroth index is given the value 1. Similarly, 'no' label encoded as '0' will be represented like [0,0] which represents the first index is represented with value 0

MACHINE LEARNING ALGORITHMS

1. **SVM:**

For Support Vector Machine, use SVM. It is a supervised machine learning algorithm that is typically applied to problems involving both regression and classification. The primary uses for this are in different classification issues. The algorithm's usual process starts with plotting each data item in an n-dimensional space, where n is the number of features and the value of each feature is the coordinate. Getting the hyper-plane that precisely divides the two classes is the next step in the classification process.

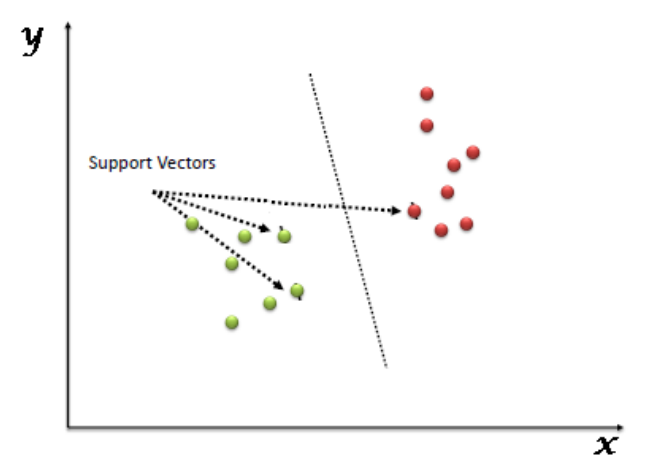


Figure 3: Support Vector Machines Example

the inner product of two vectors is the result of multiplying each pair of inputs by their sum. f(x) = B0 + sum(ai * (x,xi)) is the equation for the dot product of an input (xi) and a support vector (xi).
the inner product of two vectors is the result of multiplying each pair of inputs by their sum. f(x) = B0 + sum(ai * (x,xi)) is the equation for the dot product of an input (xi) and a support vector (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*:***

A stand for eXtreme Gradient Boosting is XGBoost. The application of gradient boosting algorithms is called XGBoost. It can be found in a variety of formats, including libraries and tools. It is primarily concerned with computational time and model performance. Both the time and the model's performance are significantly improved. Its implementation includes recently added features like regularization in addition to the features of R and Sci-kit-learn implementations. Gradient boosting with both L1 and L2-type regularizations is referred to as regularized gradient boosting. The primary best feature that the algorithm's implementation offers is the: Missing value handling automatically with sparsely aware implementation, and it offers block structures to encourage simultaneous tree building and training, thereby supporting the enhancement of an already-fitted model on the new data. Using a technique called gradient boosting, new models are created that can forecast the mistakes or

remnants of earlier models, which are subsequently combined to produce the ultimate forecast. They lessen loss when adding new models by using gradient descendent algorithms. They assist with challenges of both the regression and classification types. Typically, an objective function is defined in the training section. Establish an objective function and work toward its optimization.

obj=i=1 ∑l(yi,yi^(t))+ i=1 ∑ Ω(fi)

3. **Decision Tree**:

One of the most popular and straightforward machine learning classification problems is the use of decision trees. Numerous sophisticated algorithms, including bagging, gradient boosting, and random forests, owe their basic roots to decision trees. The more sophisticated form of the XG Boost algorithm that was previously this broad decision-tree. The decision trees CART, C4.5, C5, and ID3 are frequently utilized. Assuming the variable is numerical, a node represents a split on an input variable (X). An output variable (y) is present in the leaf, also referred to as the terminal nodes of the tree, and is essential for prediction.

First, choosing a root node is the usual process that a decision tree goes through. Prior to the split, determine the entropy or information gain for every node. Choose the node with less entropy or greater information gain. Divide the node even more and repeat the procedure. Until there is no longer any chance of splitting or the entropy is at its lowest, the process is repeated. The metric used to measure data randomness or uncertainty is called entropy. The metric used to quantify the amount of entropy reduction before and after split is called information gain.

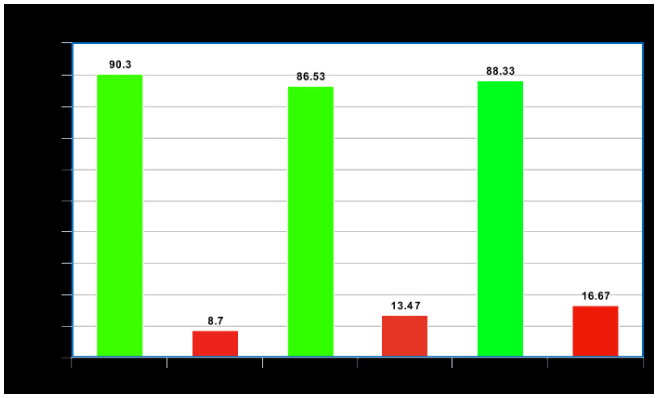
$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

TRAINING AND TESTING:

Testing is, of course, the very next task after data processing and training. This is where the algorithm's performance, the data's quality, and the necessary output all come into play. From the enormous amount of data gathered Twenty percent of the data is set aside for testing, and the remaining eighty percent is used for training. As was previously mentioned, training is the process of teaching a machine to learn and equipping it with the knowledge to make more predictions in light of the training it has received. On the other hand, testing entails using a predefined data set with output that has already been labelled to determine whether the model is functioning correctly and producing the correct prediction. If the greatest number of forecasts come true, then the model will be dependable and have a good accuracy percentage; if not, it would be best to modify the model. Additionally, new inputs and model predictions will continue to be added, enhancing the dataset's strength and precision.

RESULT:

After training and testing the data using all three algorithms, SVM yielded the highest accuracy (90.3%), followed by XG Boost (88.33%). Since SVM produced the best accuracy, SVM will be used for all subsequent data predictions. And now at last, a web application is designed to provide the student's input parameters, after which the final forecast is produced and shown. SVM is the background algorithm in use, and new predictions are continuously added to the dataset to increase accuracy.



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.

[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] 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

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

[11] 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

[12] 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.

[13] 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.

[14] 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

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

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