IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94115

CAREER COUNSELING SYSTEM USING MACHINE LEARNING

SWASTI BHUTANI

[a]Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

Abstract: As more and more opportunities are originating in today's technical world engineering students are getting more options to choose their field of interest from. Hence, it becomes very important for them to be aware of their interests and capabilities at early stages of their career. This will help them start early, and target their efforts in improving their performance. They can assess themselves on all grounds and work on their weaknesses. Even the recruiters assess students in these aspects before recruiting them for a particular job role. This will help both candidates and recruiters to analyze and evaluate the candidate's performance in various areas and suggest the best job profile for him.This paper mainly concentrates on the career area prediction of computer science domain candidates.

KEYWORDS- Career Counseling, SVM, Decision Tree, Feature Selection, XG Boost

INTRODUCTION

In today's extremely ambitious technical world, students need to be ahead of their times to stay competitive. They need to have the forethought to plan their career. Organizing and designing their career path in early stages of their education can help them make targeted efforts towards their goals. Hence it becomes imperative to consistently evaluate their performance, identify their interests and analyze how close they are to their goals. This benefits them in improving themselves, motivating themselves to a better career path if their capabilities are not up to standard to reach their goal and evaluate themselves before going to the career peak point. Even the recruiters while recruiting candidates into their companies evaluate candidates on various parameters and draw a final conclusion to select an employee or not and if selected, finds a best suited role and career area for him. Some of the many types of roles for which recruiters look for candidates are Database administrator, Business Process Analyst, Developer, Testing Manager, Networks Manager, Data scientist and so on. All these roles require some essential knowledge to be placed in them. Recruiters analyze these skills, talents and interests and place the candidate in the right job role suited for them. Various third party performance evaluation portals like Co-Cubes, AMCAT are already using these career recommendation systems too. They only take into account factors like technical abilities and psychometry of students. This career prediction system also considers students' abilities in sports, academics and their hobbies, interests, competitions, skills and knowledge. After evaluating all the factors the total number of parameters that were considered as inputs are 36, and there are 15 job roles. As the input parameters and final classes of output are large in number, advanced machine learning algorithms like SVM, Random Forest decision tree, OneHot encoding, XG boost are used.

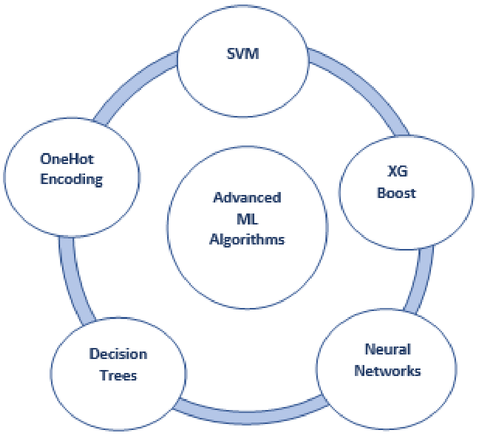

Figure 1: Overview of various Advanced Machine Learning Algorithms

DOI: 10.17148/IARJSET.2022.94115

Machine Learning as the name suggests is about training machines to learn the given inputs and respond to new inputs or scenarios based on their knowledge of previous information automatically without explicitly providing any programme or human intervention. Computers are given the ability to learn and make decisions using statistical techniques. This aims at reducing human involvement in machine dependable problems and scenarios. It solves complex problems with ease and negligible human intervention. NLP, classification, prediction, image recognition, medical diagnosis, algorithm building, self-driving cars are various applications of machine learning. In this project classification and prediction algorithms are used. Majority of 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 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 finally they are made into groups based on these characteristics then it is called unsupervised. Classification algorithms are supervised. Based on the properties of the provided input parameters a predefined class label is assigned to them. There are other alternatives like clustering and regression. After assessing the type of the problem the apt model is chosen. In this project algorithms like SVM, OneHot encoding, Decision tree and XG boost are used. After the data is trained and tested, most accurate results given by the algorithms used After training and testing the data with these we take into consideration the most accurate results given algorithm for our further processing. So, the initial task done is predicting the output using all algorithms proposed above and later analyzing the results and there on continued with the most accurate algorithm. So finally, this paper deals with various advanced machine learning algorithms that involve classification and prediction and are used to improve the accuracy for better prediction, reliability and analyzing these algorithms performance.

2. IMPLEMENTATION

2.1 MINOR PROJECT:

Data collection, processing and encoding was done in the minor project itself.

OneHot encoding was used for data encoding. It is a technique by which categorical 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.

The most relevant features were selected and all the redundant, irrelevant, or noisy features were removed using Chi square method. The formula for calculating chi square value is given as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where χ 2 = Chi-Square value, Oi = Observed frequency, Ei = Expected frequency

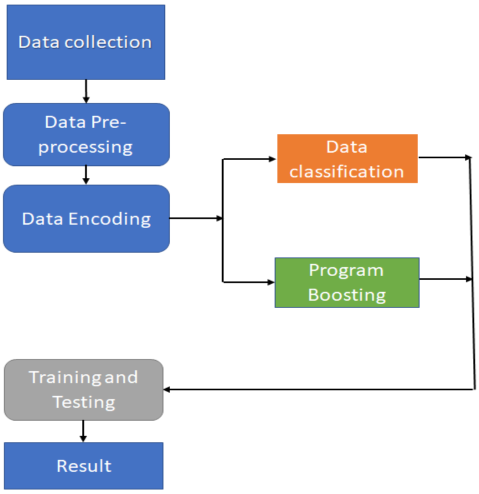

Figure 2: Process Flow Diagram of proposed system

DOI: 10.17148/IARJSET.2022.94115

2.2 SVM- Support Vector Machine :

This algorithm follows a typical procedure to classify the data set provided. Each data item is plotted in an n-dimensional space where n represents the number of features and the value of each feature is represented by a particular coordinate. The classification is done by a hyper-plane which separates these points into two classes with a wide gap in between.

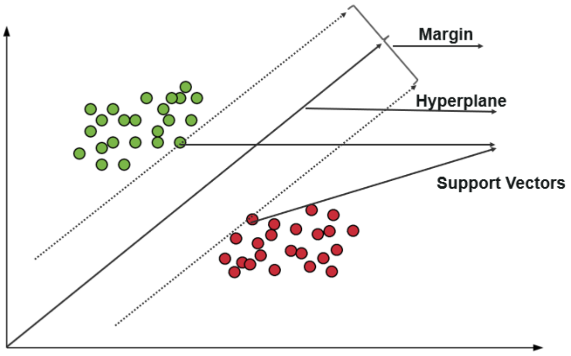

Fig 3. Support Vector Machine

2.3 Decision tree:

Decision trees are the most popular and widely used machine learning algorithm for classification problems. They are the foundation blocks for many other advanced machine learning algorithms including XGBoost, Random Forest and bagging. Infact the XGBoost algorithm that we use further to improve the performance of our program is an advanced version of the decision tree itself. A node here denotes an input variable (x) and a split on that variable assuming the variable is numerical. The leaf which are also called the terminal nodes of the tree possess an output variable (y) which is necessary for prediction. The algorithm follows a fixed procedure which starts by selecting a root node. Entropy of each node is calculated before the split. The node with less entropy is selected and the process of splitting the node is reiterated. Entropy is the measure of uncertainty or randomness of data.

2.4 XGBoost:

XGBoost denotes eXtreme Gradient Boosting. XGBoost is an implementation of gradient boosting algorithms. It is available in many forms like tools, library et cetera. It mainly focuses on model performance and computational time. It greatly reduces the time and greatly lifts the performance of the model. 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 trees and continued training which supports further boosting an already fitted model on the fresh data. Gradient boosting is a technique where new models are made that can predict the errors or remains of previous models and then added together to make the final prediction. They use gradient descent algorithms to reduce loss during adding of new models. They support both classification and regression type of challenges.
In this project as you will see this algorithm boosted the accuracy of the results of the decision tree from 70% to 93.333%.

2.5 Training and Testing:

Finally after processing of data and training the very next task is obviously testing. This is where performance of the algorithm, quality of data, and required output all appear. From the huge data set collected 80 percent of the data is utilized for training and 20 percent of the data is reserved for testing. Training as discussed before is the process of making the machine to learn and giving it the capability to make further predictions based on the training it took.Where as testing means already having a predefined data set with output also previously labeled and the model is tested whether it is working properly or not and is giving the right prediction or not. If the maximum number of predictions are right then the model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model. Also

DOI: 10.17148/IARJSET.2022.94115

further new sets of inputs and the predictions made by the model will be kept on adding to the dataset which makes the dataset more powerful and accurate.

Result:

The data is trained and tested with all three algorithms and out of all XGBoost gave more accuracy with 93.333 percent and then the SVM with feature selection 88.33 percent accuracy. As XGBoost gives the highest accuracy, all further data predictions are chosen to be followed with XGBoost algorithm.

The results obtained with and without feature selection were also compared and a graph was plotted depicting the same.

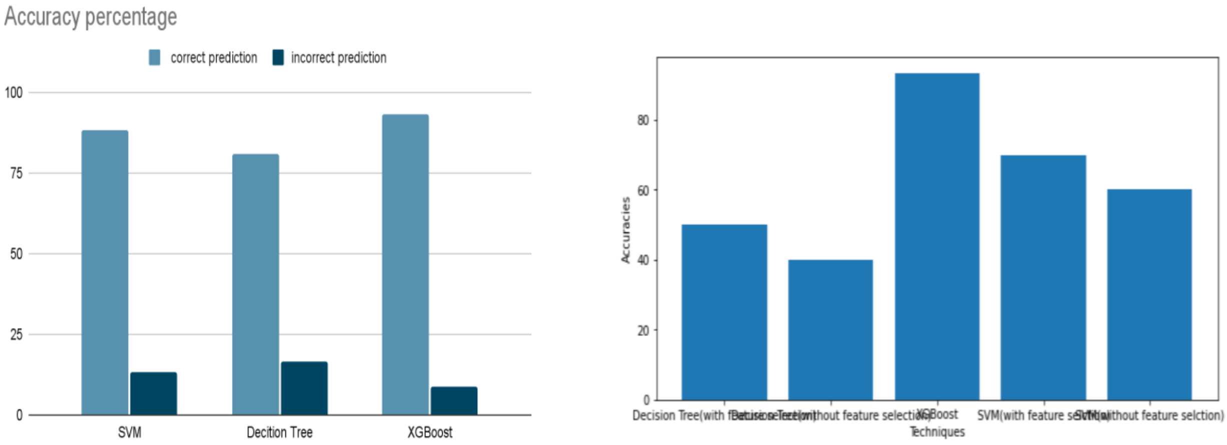

Fig 4. Bar charts of results obtained

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