**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Career Guidance System

***[1]Ashutosh Kavishwar, [2]Ankit Gurjar, [3]Ankit Nagar, [4]Aman Uniyara, [5]Pir Mohammad***

[1,2,3,4]Student, Acropolis Institute of Technology and Research

[5]Professor, Acropolis Institute of Technology and Research

**Abstract: -**

As students are going through their academics and pursuing their interested courses, it is very important for them to assess their capabilities and identify their interests so that they will get to know in which career area their interests and capabilities are going to put them in. This will help them in improving their performance and motivating their interests so that they will be directed towards their targeted career and get settled in that. Also recruiters while recruiting the candidates after assessing them in all different aspects, these kind of career recommender systems help them in deciding in which job role the candidate should be kept in based on his/her performance and other evaluations. This paper mainly concentrates on the career area prediction of computer science domain candidates.

***Key-Words:*** *Students, Career, Recommender system, Prediction*

## I. Introduction

Competition in today's society is heavily multiplying day by day. Especially it is too heavy in present day's technical world. So as to compete and reach the goal students need to be planned and organized from initial stages of their education. So it is very important to constantly evaluate their performance, identify their interests and evaluate how close they are to their goal and assess whether they are in the right path that direct towards their targeted. This helps them in improving themselves, motivating themselves to a better career path if their capabilities are not up to the mark to reach their goal and pre evaluate themselves before going to the career peek point.

Not only that recruiters while recruiting people into their companies evaluate candidates on different 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. There are many types of roles like Database administrator, Business Process Analyst, Developer, Testing Manager, Networks Manager, Data scientist and so on. All these roles require some pre-requisite knowledge in them to be placed in them. So, recruiters analyze these skills, talents and interests and place the candidate in the right job role suited for them. These kind of prediction sys-tems make their recruitment tasks very easy because as the inputs are given, recommendation is done based on inputs. Already these type of various career recommendation systems and job role recommendation, prediction systems are being used in various third party performance evaluation portals like Co-Cubes, AMCAT. They only take factors like technical abilities and psychometric of students into consideration. These portals asses the students technically and suggest the students and companies job roles suited on their performance. But here various factors including abilities of students in sports, academics and their hobbies, interests, competitions, skills and knowledge are also taken into consideration. Considering all the factors the total number of parameters that were taken into consideration as inputs are 36. And the final job roles are fixed to 15 in number. As the input parameters and final classes of output are large in number typical programming and normal algorithms cannot give the best possible output classification and prediction. So 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

## 2. Problem Formulation

While graduation signifies the end of college many students can be left wondering, 'what's next?' Many students face a difficult time choosing a career path in college, especially in fields like engineering where there are vast domains to choose from. Thus it is very important for students to assess their capabilities and identify their interests while pursuing their studies so that they are clear about their dreams and the career path leading towards it.

Thus, we have proposed an idea of an ML-based "Career Guidance System" which helps the students (esp. In their pre-final and final year) to decide the job role the candidate should undertake based on his/her performance and other evaluations.

Our proposed idea of the Career recommendation system takes into consideration the students' abilities in academics, technicality, hobbies, interests, psychometric, skills, and knowledge. Taking these inputs machine learning algorithms will be applied and suitable job roles will be suggested.

ML Algorithms used:

● Support Vector Machine (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 this can be found in various classification problems.

● XG Boost

XGBoost denotes eXtreme Gradient Boosting. XGBoost is implementation of gradient boosting algorithms. It is available in many forms like tool, 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. Its implementation has the features of scikit- learn and R implementations and also have a newly added features like regularization.

● Decision Tree

Decision Trees are extremely popular and one of the simple and easy to implement machine learning classification problems. Decision trees laid basic foundation for many advanced algorithms like bagging, gradient boosting and random forest. The XG Boost algorithm discussed above is the advanced version of this general decision tree. The commonly used decision trees are CART, C4.5,C5 and ID3.

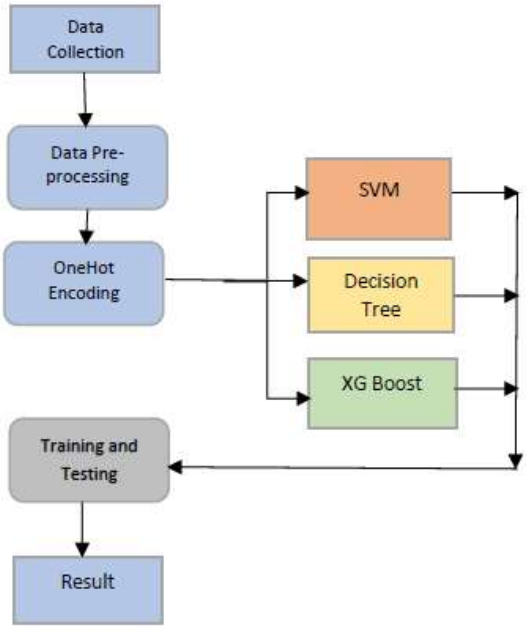



**Figure 2**: Process Flow Diagram

## III. Literature Review

There are various websites and web apps over the internet which helps students to know their suitable career path. But most of those systems only used personality traits as the only factor to predict the career, which might result in an inconsistent answer. Similarly, there are few sites that suggest career based on only the interests of the students. The systems did not use the capacity of the students to know whether they would be able to survive in that field or not. The paper by [1] Beth Dietz-Uhler & Janet E. Hurn suggest the importance of learning analytics in predicting and improving the student's performance which enlightens the importance of student's interest, ability, strengths etc. in their performance. According to the paper by [2] Lokesh

Katore, Bhakti Ratnaparkhi, Jayant Umale, the career prediction accuracy was determined using 12 attributes of students and different classifiers with c4.5 having the highest accuracy of 86%. [3] Another paper by Roshani Ade, P.R.Deshmukh suggested incremental ensemble of classifiers in which the hypothesis from number of classifiers were experimented and by using 'Majority voting rule', the final results was determined. The proposed ensemble algorithm gave an accuracy of 90.8% [4].The paper by Mustafa Agaoglu suggested the importance of different attributes in evaluating the performance of faculty. It also showed the comparison of different classifiers proved that the most accurate classifier was c5.0 which has the maximum attribute usage compares to other classifiers like CART, ANN-Q2H, SVM etc. Also, the suggestions provided by the system are very much generalized and not specific to a university or country/state. The suggestion for course is also generalized. For example, the results of few systems were a group of courses like data analyst, accountant, law etc. Thus, if a student gets such a recommendation then he/she might again get confused as the above specified course belongs to different streams.

## IV. Methodology

***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 col-lected. So as the data same will be the output. For student ca-reer prediction many parameters are required like students 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 stu-dent's progress towards a career area, all these are taken in-to consideration. Data is collected in many ways. Some data is collected from employees working in different organizations, some amount of data is collected through LinkedIn api, 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.

***Data Pre-processing:***

Collecting the data is one task and making that data useful is an-other vital task. Data collected from various means will be in an unorganized format and there may be lot of null values, in-valid 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 pre processing of data. Even data collected may contain completely garbage val-ues. 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.

***OneHot Encoding:***

OneHot Encoding is a technique by which categorial val-ues 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. Simp-ly OneHot encoding transforms categorial values into a form that best fits as input to feed to various machine learning algo-rithms. 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 OneHot 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 espe-cially in the case of prediction. Next step is creating a vector for each integer value.

***Machine Learning Algorithms***

***SVM:***

SVM denotes Support Vector Machine. It is a supervised machine learning algorithm which is generally used for both regres-sion and classification type of problems. The main applica-tions of this can be found in various classification problems. The typical procedure of the algorithm is first each data item is 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 by getting the hyper-plane that separates the two classes very finely.

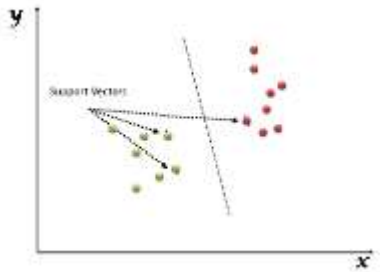


**Figure 3**: Support Vector Machines Example

SVM algorithms practically are implemented using kernels. There are three types of SVM's and in linear SVM hyperplane is calculated or found by transforming the problem using linear alge-bra. The insight is that SVM can be rephrased by using the inner product of two observations. The sum of the multipli-cation 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:

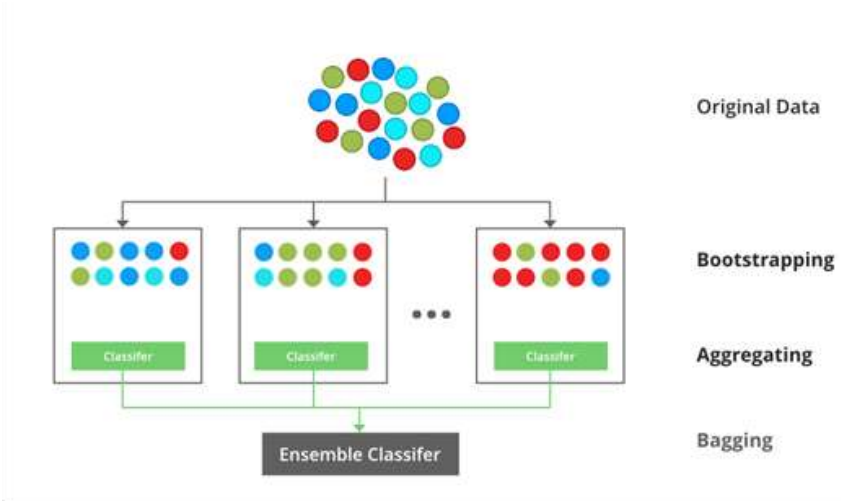
K(x,xi) = exp(-gamma * sum((x – xi^2))

***XG Boost:***

XGBoost denotes eXtreme Gradient Boosting. XGBoost is implementation of gradient boosting algorithms. It is availa-ble in many forms like tool, library et cetera. It mainly focus-es on model performance and computational time. It greatly reduces the time and greatly lifts the performance of the model. 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 gra-dient boosting with both L1 and L2 type regulariza-tions. The main best features that the implementation of the algo-rithm provides are: Automatic handling of missing values with sparse aware implementation, and it provides block struc-ture to promote parallel construction of tree and continued train-ing which supports further boost 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 descendent algorithms to reduce loss during adding of new models. They support both classification and regression type of challenges. In the training part generally an objec-tive function is defined.

Define an objective function and try to optimize it

obj=i=1 Σl(yi,yi^(t))+ i=1 Σ Ω(fi)

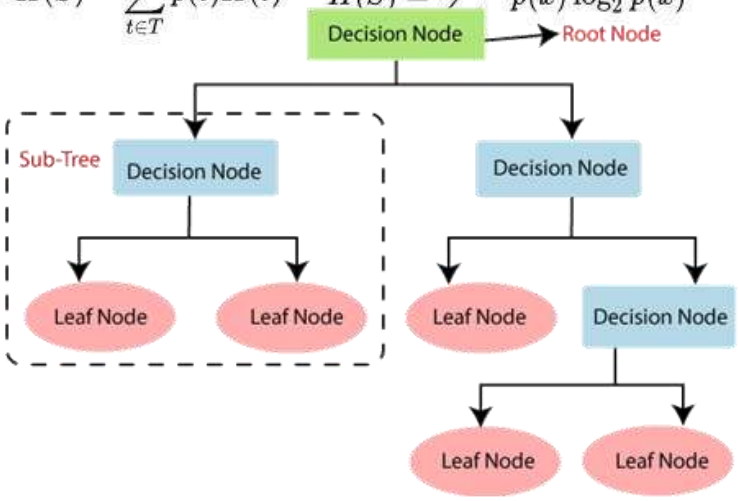



***Decision Tree:***

Decision Trees are extremely popular and one of the simple and easy to implement machine learning classification problems. De-cision trees laid basic foundation for many advanced algo-rithms like bagging, gradient boosting and random forest. The XG Boost algorithm discussed 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 varia-ble (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 vital for prediction.

The typical scenario that a decision tree follows is first selecting a root node. Calculate information gain or entropy for each of the nodes before the split. Select the node that has more infor-mation gain or less entropy. Further split the node and reiterate the process. The process is iterated until there is no possibility to split or the entropy is minimum. Entropy is the metric to meas-ure uncertainty or randomness of data. Information gain is the metric that measures how much entropy is reduced before to after split.

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t) \qquad H(S) = \sum -p(x)\log_2 p(x)$$




***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 appears out. From the huge data set collected 80 percent of the data is utilized for train-ing 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 labelled and the model is tested whether it is working properly or not and is giving the right prediction or not. If maximum number of predictions are right then model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model. Also further new set of inputs and the predictions made by the model will be keep on adding to the dataset which makes dataset more powerful and accurate.

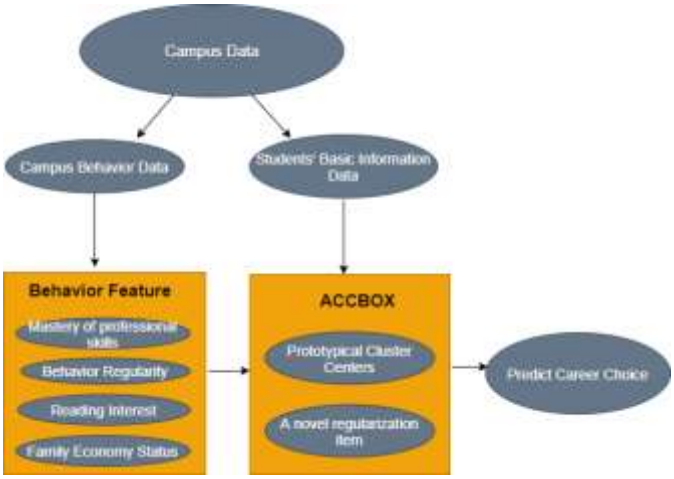
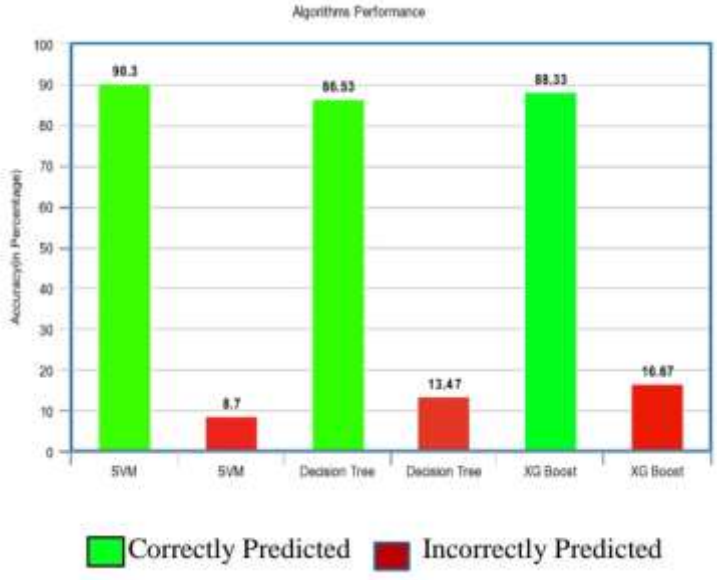



**Figure 1.** Framework of the proposed Model

## V. Result Discussions

The data is 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 highest accuracy, all further data predictions are chosen to be followed with SVM. So, finally a web application is made to give the input parameters of the student and the final prediction is generated and displayed. The background algorithm being used is SVM and the new prediction are keep on adding to the dataset for further more accuracy.




## VI. Conclusion

So, finally a web application is made to give the input parameters of the student and the final prediction is generated and displayed. The background algorithm being used is SVM and the new prediction are keep on adding to the dataset for further more accuracy. A more powerful web application can be developed where inputs are not given directly instead student parameters are taken by evaluating students through various evaluations and examining. Technical, analytical, logical, memory based, psychometric and general awareness, interests and skill based tests can be designed and parameters are collected through them so that results will be certainly accurate and the system will be more reliable to use.

***Acknowledgment***

There are number of people without whom this projects work would not have been feasible. Their high academic standards and personal integrity provided me with continuous guidance and support.

We owe a debt of sincere gratitude, deep sense of reverence and respect to our guide and mentor **Prof. Pir Mohammad,** Professor, AITR, Indore for his motivation, sagacious guidance, constant encouragement, vigilant supervision and valuable critical appreciation throughout this project work, which helped us to successfully complete the project on time.

We express profound gratitude and heartfelt thanks to **Prof. Pir Mohammad**, Professor, AITR Indore for his support, suggestion, and inspiration for carrying out this project. I am very much thankful to other faculty and staff members of IT Dept, AITR Indore for providing me all support, help and advice during the project. We would be failing in our duty if do not acknowledge the support and guidance received from **Dr S C Sharma**, Director, AITR, Indore whenever needed. We take opportunity to convey my regards to the management of Acropolis Institute, Indore for extending academic and administrative support and providing me all necessary facilities for project to achieve our objectives.

References

**[1]** P.KaviPriya**, "**A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques", International Jour-nal 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", Inter-national 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 Jour-nal 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 Compu-ting

**[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 Educa-tional 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", Inter-national Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015.

**[11]** Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest Shiju Sathyadevan and Remya R. Nair

**[12]** Yu Lou, Ran Ren, "A Machine Learning Approach for Future Ca-reer Planning"

**[13]** Gareth James ,Daniela Witten ,Trevor Hastie, "An Introduction to Statistical Learning with Applications in R"

**[14]** Anuj Karpatne, Gowtham Atluri, "Theory- Guided Data Science: A New Paradigm for Scientific Discovery from Data", IEEE trans-actions on knowledge and data engineering, vol.29, no. 10, october 2017.