



AN INTELLIGENT CAREER GUIDENCE SYSTEM

¹ Potteti sumalatha, ² Manogna Mandaloji, ³ Manisha Amuda

¹ Associate Professor, Department of Computer Science and Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India.

¹ sumalatha.po@gmail.com

^{2,3} Students, Department of Computer Science and Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India.

² mandaloji.manogna@gmail.com , ³ amudamanisha153@gmail.com

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.

1. 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 directs 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 them 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 prerequisite 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 systems 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 psychometry of students into consideration. These portals assess 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.

2. LITERATURE SURVEY

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 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 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%. 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%.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

3. PROBLEM STATEMENT

In the current scenario, the career prediction system work for getting job

recommendations. Several platforms such as AMCAT, CoCubes, etc. they provide job recommendations. There is no such system exist which takes input and gives recommendations of the suitable job profile.

Many machine learning techniques, such as decision trees, artificial neural networks, matrix factorization, collaborative filters and probabilistic graphical models, have been applied to develop prediction algorithms. Most of this work ignores the continuous effect that students enhance their knowledge over time and follow the prediction as a one-time task.

4. PROPOSED SYSTEM

In this paper classification and prediction are being done. Let us see what is classification and prediction. Majority of problems in machine learning can be solved using supervised and unsupervised learning. If the final class labels are previously know and all the other data items are to be assigned with one of the available class labels, then it is call 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 falls under supervised. Input



parameters are given and based on their properties a predefined class label is assigned. There are other alternatives like clustering and regression. Based on the type of problem the apt model is chosen. However here algorithms like SVM, One Hot encoding, Decision tree and XG boost are used. After training and testing the data with these we take into consideration the most accurate results given algorithm for our further processing. So, 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 involves classification and prediction and are used to improve the accuracy for better prediction, reliability and analyzing these algorithms performance.

5. IMPLEMENTATION

5.1 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 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 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 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 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 pre processing



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 meaningful and useful for further processing. Data must be kept in a organized format.

5.2 OneHot Encoding

OneHot Encoding 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. Simply OneHot encoding transforms categorical 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 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 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. For example [pillow, rat, fight, rat] becomes [0,1,2,1]. This is here imparting an ordinal property to the variable, i.e. pillow < rat < fight. As this is ordinal characteristic and is usually not required and desired 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.

5.3 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. 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 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 the training part generally an objective function is defined

5.4 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 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 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.

6. ARCHITECTURAL DESIGN

Django is based on MVT (Model-View-Template) architecture. MVT is a software design pattern for developing a web application.



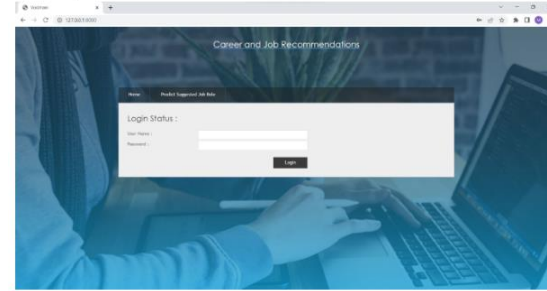
MVT Structure has the following three parts :

Model: Model is going to act as the interface of your data. It is responsible for maintaining data. It is the logical data structure behind the entire application and is represented by a database (generally relational databases such as MySQL, Postgres). To check more, visit – Django Models

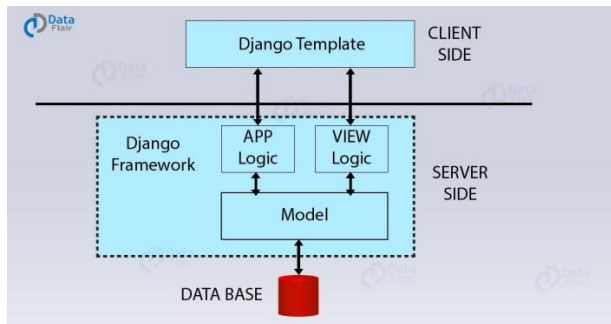
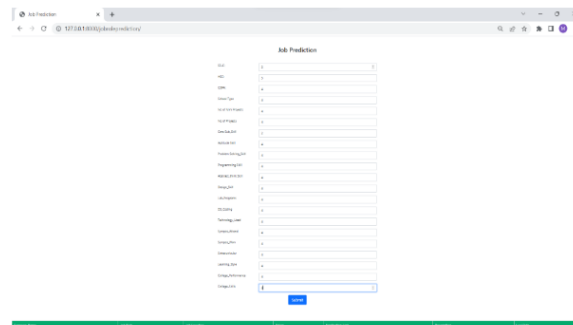
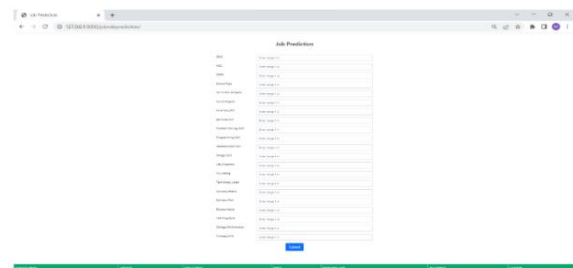
View: The View is the user interface — what you see in your browser when you render a website. It is represented by HTML/CSS/Javascript and Jinja files. To check more, visit – Django Views.

Template: A template consists of static parts of the desired HTML output as well as some special syntax describing how dynamic content will be inserted. To check more, visit – Django Templates

1.Home Page



3. Admin Home page



7. OUTPUT RESULTS

8. CONCLUSION

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.

9. FUTURE SCOPE

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes: 1. Increase the accuracy of the algorithms. 2. Improvising the algorithms to add more efficiency of the system and enhance its working. 3. Working on some more attributes so to tackle diabetes even more.

Future Scope 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, psychometry 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.

10. 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, Sayma Sultana, 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] 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.