Career Recommender System Using Decision Trees

Rachit Jaluthra
Computer Engineering Department,
Delhi Technological University,
Delhi, India
Email : rjrachit005@gmail.com

Massoud Massoudi
Software Engineering Department PhD. Candidate,
Delhi Technological University,
Delhi, India
Email: massoud.massoudi@hotmail.com

Primanshu
Computer Engineering Department,
Delhi Technological University,
Delhi, India
Email : primanshu.dtu@gmail.com

***Abstract*—With so many career opportunities present in the computer science field, it becomes quite tedious for an individual to identify their strengths and goals. The purpose of this study is to form a relation between the experiences of an individual and identify the best suitable career path in the computer science field. Various fields of computer science are classified based on decision trees. Data analysis found some fields to be more likely to be pursued due to similar academic and job interests amongst students. The study observed quality correlation between the experiences and career paths.**

Keywords-component; *decision trees; information gain; gini; entropy;*

I. INTRODUCTION

In today's world people have so many options with products, movies, restaurants and also in their career area. Sometimes an individual may get very confused for which career area to process through. An individual frequently needs to evaluate skills, interests and certifications to decide their career paths. This paper spreads light on a system which can analyze their strengths and help them in pursuing their career. As for an individual looking for guidance in career choosing and for a recruiter recruiting candidates it can help them to analyze their strengths and choose which role is the best option for him. This can reduce workload of recruiters by just entering the details of candidates in the system [1]. Earlier attempts of career trajectory are done by using convolutional neural networks architecture to extract the semantic features from the short sentence in the resume [2]. K-Nearest Neighbour method and Certainty factor are used for predicting student career, based upon students' final exam results and interests [3]. Recommendation systems using collaborative filtering, content based filtering, matrix factorization and a combination of them are widely known t o exist but recommendation using decision trees needs digging. Decision trees are mainly used for classification problems, so it can be used to classify various career streams and recommend the same. Various factors like their strength, interests, skills, communication skills etc can be considered and trained upon. Using the knowledge of how decision trees are formed we can intuitively use them for deciding the career paths. The decision tree in the study is based on answering a set of questions upon which a role is decided and recommended. It analyzes the data and presents it in an easy to understand form of trees. This alternative form of recommendation can form correlation matrix and visual tree structure, giving us more insight of the system[1].

II. ALGORITHM

A. Decision Trees

Trees have a number of implications in real life. One such being used for classification and regression in machine learning. A decision tree has a root node, intermediate nodes and leaf nodes. The starting node is called the root node. decision trees follow a top-down approach with each node splitting the tree further. The leaf node can't be split further and represents the decision or class. For accurate results, it is expected to have greater depth of the tree [4].

The splitting of nodes can be done using a variety of methods such as gini, information gain, variance and chi square. Commonly used decision trees include *Iterative Dichotomiser 3* (ID3), *Classification and Regression trees* (CART) and *C4.5* (an extension of ID3).

B. Information Gain

The randomness in decision trees is measured using the Entropy i.e. it is basically a way of measuring the lack of order that exists in a system [5]. ID3 and C4.5 are the two of many algorithms which use the concept of conditional entropy and information gain for decision trees [6]. Entropy can be calculated as

$$E(X) = -\sum_{i=1}^{n} p_i * log_2(p_i) \quad (1)$$

Here, n is the number of classifications in the selected column and p_i is probability of class i [5].The information gain is then calculated using from (1)

$$IG(X, Y) = E(X) - \sum_{i=1}^{n}((|X_a| / |X|) * E(X_a)) \quad (2)$$

X_a is the set of rows which has value ***a*** in the *target* column. $|X_a|$ and $|X|$ are the number of rows in their corresponding set [6].

Fig. 1. shows a decision tree generated using information gain for splitting nodes. Decision trees in Fig 1. and Fig. 2. are partially shown upto a depth value of 2.

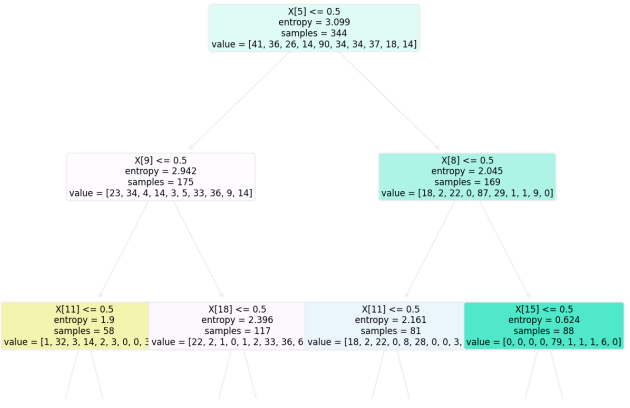

Figure. 1. Decision tree generated via criterion as entropy

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

C. Gini Index

The probability of classifying a certain variable wrong is evaluated using gini measure. CART uses gini measure to design binary splits [7]. Gini impurity :

$$GI = 1 - \sum_{i=1}^{n}(p_i)^2 \quad (3)$$

where p_i is the probability of an object being classified to a particular class. Fig. 2. shows a decision tree generated using the gini index for splitting nodes.

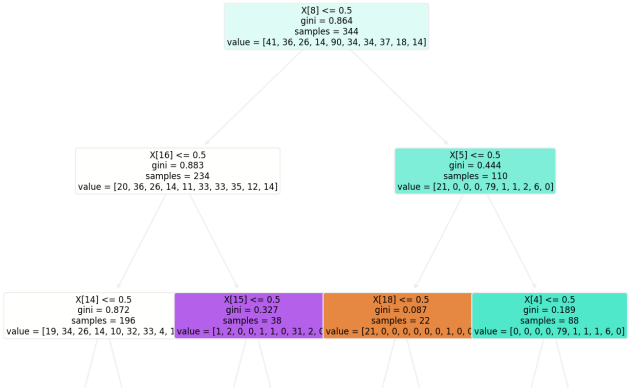

Figure. 2. Decision tree generated using gini index criterion

III. METHODOLOGY

The aim was to address the practical solution for predicting career roles given the traits of an individual. Since aptitude and personality are mostly independent, data of both can not be combined to derive to results [8]. The data needed should contain information about the candidate-skills, certifications, communication skills, habits, it was best covered in qualitative form. It was expected that the majority of the candidates would fall into one or two classes due to similar academics and training, also some of the classes are more prevalent job paths.

The approach used in this paper is highly suitable because there was a correlation between the features of an individual and his job role. For example, a candidate knowing coding, development skills and interest in technology is expected to become a developer and not a marketing or management agent. So it can be practical to use decision trees in which each node asks these questions and then decide the outcome.

A. Data preprocessing and Analysis

The data used was taken from Kaggle. The data contained a set of questions and a job role (*label*) for each row. *524* rows and *20* columns were present, of which 18 were considered as *features* and the last column (*role*) was the *label* column. The feature columns answered the candidates strengths, interests and skills. Label column highlighted which job role is best suited. According to the dataset the label column had 10 classes of roles and all other feature columns had 2 classes except *communication skills* which had 3 classes. Feature columns represented a set of questions like- *If the individual is good at coding? If they have done any machine learning courses? How are their communication skills? If their interest lies in marketing?*

The dataset was first checked for consistency and redundancy. The dataset after checking for duplicate and inconsistent material is then reduced to *492* rows and *20* columns. Since all the information was in qualitative form all the columns were labelled and encoded before processing further. The dataset is then divided into testing and training data in a ratio of *0.30*.

As expected data analysis showed that the majority of the candidates belonged to a particular class i.e. *developer*. The bar graph shown in Fig. 3. is suggestive of the fact that one class of role i.e. developer (*dark blue bar*) contributes for the majority of the job roles.

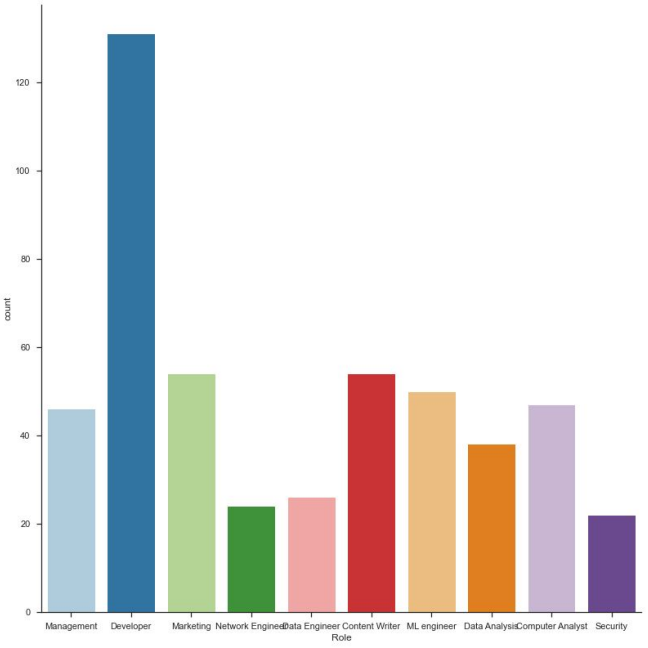

Figure. 3. Number of Candidates versus Role

B. Training and Testing

The dataset was divided into two parts- *feature columns and label column.*

After separating data into training set and test set, the decision tree is then trained. For training and testing decision tree model *DecisionTreeClassifier* is used from library *sklearn*. The data is fitted and tested for the methods gini and entropy. Confusion matrix and accuracy, from a maximum depth of 1 to 10, is evaluated for both gini and entropy criterion. From all the accuracies evaluated the best accuracy is considered and confusion matrix and decision trees are visualized for the same. Further analyses of data are done using support vector machine(SVM) classification and confusion matrix is visualized. During training phase Fig. 1. and Fig. 2. were generated, which gave an in depth visualization of gini based and entropy based decision trees and its associated variables, respectively. Figure 4 marks the comparison of accuracy graphically between SVM, Information gain and gini index method.

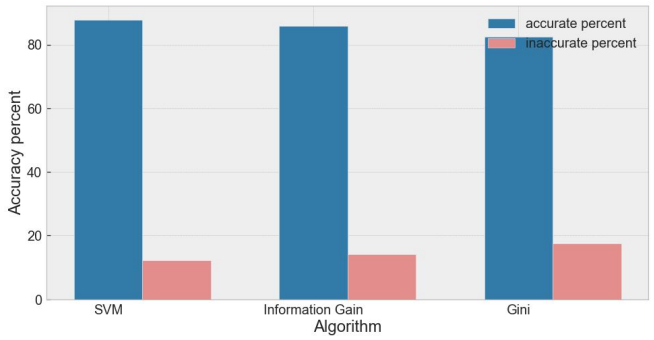

Figure. 4. Accuracy graph.

IV. Conclusion

Decision trees proved to be very handy to use for classification tasks. Decision trees based on gini criterion gave an accuracy of 83.10 percent. Decision trees based on information gain method gave an accuracy of 85.81 percent. SVM gave the highest accuracy of 87.83. It is concluded that decision trees based on information gain, having a tree depth of 6, gave better results compared to the gini measure method.

References

[1] Sripath Roy, K., Roopkanth, K., Uday Teja, V., Bhavana, V., & Priyanka, J. (2018). Student Career Prediction Using Advanced Machine Learning Techniques. International Journal of Engineering & Technology,7(2.20),26-29.doi:http://dx.doi.org/10.14419/ijet.v7i2.20.11738

[2] He, M., Shen, D., Zhu, Y., He, R., Wang, T., & Zhang, Z. (2019). Career Trajectory Prediction based on CNN. 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI).

[3] Nunsina, Tulus, & Situmorang, Z. (2020). Analysis Optimization K-Nearest Neighbor Algorithm with Certainty Factor in Determining Student Career. 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT).

[4] Shankhdhar, A., Agrawal, A., Sharma, D., Chaturvedi, S., & Pushkarna, M. (2020). Intelligent Decision Support System Using Decision Tree Method for Student Career. 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and Its Control (PARC).

[5] Huaining Sun and Xuegang Hu, "An improved learning algorithm of decision tree based on entropy uncertainty deviation," 2012 IEEE 14th International Conference on Communication Technology, Chengdu, 2012, pp. 799-803, doi: 10.1109/ICCT.2012.6511313.

[6] Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. 2011 IEEE Control and System Graduate Research Colloquium.

[7] Sivagama Sundhari, S. (2011). A knowledge discovery using decision tree by Gini coefficient. 2011 International Conference on Business, Engineering and Industrial Applications.

[8] Rangnekar, R. H., Suratwala, K. P., Krishna, S., & Dhage, S. (2018). Career Prediction Model Using Data Mining and Linear Classification. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[9] Ade, R., & Deshmukh, P. R. (2014). An incremental ensemble of classifiers as a technique for prediction of student's career choice. 2014 First International Conference on Networks & Soft Computing (ICNSC2014).

[10] Elayidom, S., Idikkula, S. M., Alexander, J., & Ojha, A. (2009). Applying Data Mining Techniques for Placement Chance Prediction. 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies.

[11] Devasia, T., Vinushree T P, & Hegde, V. (2016). Prediction of students' performance using Educational Data Mining. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).

[12] Supianto, A. A., Julisar Dwitama, A., & Hafis, M. (2018). Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University. 2018 International Conference on Sustainable Information Engineering and Technology (SIET).

[13] 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.

[14] Huang Ming, Niu Wenying, & Liang Xu. (2009). An improved Decision Tree classification algorithm based on ID3 and the application in score analysis. 2009 Chinese Control and Decision Conference.