Student's Career Interest Prediction using Machine Learning

[1]Prof. Priyanka Shahane, [2]Prutha Rinke, [3]Taniksha Datar, [4]Soham Badjate

[1]Assistant Professor, SCTR's Pune Institue of Computer Technology,411043
[2,3,4]Student, SCTR's Pune Institue of Computer Technology, 411043

Abstract - *Many students are known to encounter uncertainty about their career choices and the marketability of the interests they own and wish to pursue. Due to the availability of many domains in the evolving world, the complexity of this issue is increasing. So, in this competitive world with so many domains and skills to work on, it is nearly impossible for a human to explore all the domains and get knowledge of the skills in such a short period so as to acquire an idea about what domains and skills would interest the individual, which can help in building the overall career path. It can be predicted based on an individual's academic and non-academic history and also keeping in mind the individual's interests, school performances, and perspective on future careers. It is a critical data set for research, as people tend to explore many areas along the way. This research aims at studying various machine learning based techniques for students' career interest prediction.*

Key Words: **(SVM, OneHot Encoding, Decision Tree, XG Boost, K-Nearest Neighbor, Educational Data Mining, Regression Models, Supervised Learning, Unsupervised Learning)**

1. INTRODUCTION

In today's competitive world it is very difficult for students to understand their interests and find their suitable career. There are numerous fields to choose from. Choosing from this huge plethora of career options is a real challenge before the individuals today. To compete and reach the goals of the students it is very important for them to plan and organize from the initial stages of their lives. For this it is necessary to constantly evaluate their performance, identify interests and keep track of how close they are to their goals and also assess if they are on the right track towards their target. This helps them to improve themselves and also pre evaluate their capabilities before going to the career peak point.

Recruiters while recruiting people into their companies evaluate candidates on different parameters and then draw a final conclusion to select an employee or not. If selected they have to find the best suited role and career area for the selected employees. Various career recommendation systems, job role recommendation and prediction systems are being used in different private performance evaluation portals like Co-Cubes, AMCAT. These portals evaluate the students technically and suggest the students and companies job roles best suited for them based on their performance. Various factors including abilities of students in sports, academics, extracurriculars and also their hobbies, interests, skills are also taken into consideration.

We have used advanced machine learning algorithms like SVM, Random Forest decision tree, XG boost for predictions. After training and testing the data with these algorithms we take the most accurate results into consideration for further processing. This paper deals with various advanced machine learning algorithms that involve classification and predictions, which are used to improve the accuracy for better prediction and also analyzing these algorithm's performances.

1.1 Problem Definition and Scope

A problem that never gets resolved in a student's life are their career choices. With various career paths as options and choosing out of these is a difficult task for them. It is impossible for them to explore each and every domain present in the word and choose the most suitable of these. And thus, a model is developed based on the student's performance and interests and is implemented to find out the most suitable career option for them.

This consists of excellent potential as it is seen as a benefit to the younger generation for their career choices and preferences. An accurate prediction model can help not just young age but also retired adults to work on the preferable career choice suggested to them and contribute to society for their social well-being. This model would help people to dig out their choices from a valley of career paths and provide them with suitable jobs which would help them support themselves and also increase the global economy.

2. LITERATURE SURVEY

[1] Personalized Career Path Recommender System for Engineering Students Opting for a university specialization is a grueling decision for all academy scholars. Due to the lack of guidance and limited online coffers, scholars' opinions depend on the private comprehension of family and friends. This increases the threat of high university dropout rates, and students changing their university disciplines and choosing an irrelevant career path.

To address the downsides mentioned, this exploration paper presents a Personalized Career- path Recommender System(PCRS) to give guidance, and the right choice, and help all academy scholars choose their separate disciplines. The main idea of PCRS is to mimic the part of professional counsels who help scholars make this hard decision by assaying their academic and particular interests. The design of PCRS is grounded on a fuzzy-sense intelligence with two main input parameters; academic performance and specific profile.

[2] Prediction of student's academic performance using machine learning algorithms Learning analytics and supportive learning are known to be emerging research areas in today's era of big data, data mining, machine learning, and artificial intelligence to facilitate students' learning. Student education is crucial to the sustainable development of society as students learn knowledge from schools and through extracurricular activities and create abilities to contribute to the community. There are many students who have progressed to higher levels of education and earned degrees like Ph.D. while many graduate every year. However, there are some students who marginally pass the course and some who fail the course as well who are required to have a compulsory retake the same course. This paper has been proposed as an improvised conditional network-based deep support vector machine (ICGAN- DSVM) algorithm.ICGAN focuses on addressing the issue of low data volume that is using fewer datasets by mimicking new training datasets whereas DSVM extends SVM from shallow learning to deep learning. DSVM takes the advantage of a small dataset, as a key difference in comparing with the traditional deep neural network which makes it more efficient to work and execute.

[3] A Machine Learning Based Approach for Recommending Courses at Graduate Level Students often face confusion during their academic career regarding the choice of the courses they make and their future scope, hence they need proper guidance for the same which not everyone is accessible. This paper is used to propose a system that can be used for recommending courses to students according to their benefits and interests. The paper focuses on a higher-level or graduate level of study. It utilizes techniques of Data Mining and Machine Learning in order to give accurate results to the students. The factors taken into account are based on the student's performance and their preferred interest and skills. Machine learning methods like neural networks and various learning algorithms can help in a student's career and their best interest.

[4] Analysis Optimization K-Nearest Neighbor Algorithm with Certainty Factor in Determining Student Career A career is a series of development or progress in their respective professions experienced by every human being. These careers are manifested as a job title and job work and something related to the professional world and academic growth. The crucial element of a career is the development or progress on every level of life experienced by humans. K-Nearest Neighbour is a method for data. Classification and organization while the certainty factor is an uncertain decision-making method depicted in this paper. This study uses datasets like students' interests, talents and exam scores to predict a career-appropriate decision for each student. The Student career prediction system was formed by combining two methods which are K-Nearest Neighbor and the Certainty Factor. It was expected that the two-way analysis might provide a piece of better formation for students in determining their careers with accurate results. The K-Nearest Neighbor method received a value derived from the Certainty Factor which is apparently beneficial in predicting career prediction. A system was proposed that uses the KNN-certainty factor method to predict the student's career with accuracy.

[5] Analyzing the learning of individual and suggesting field of study using Machine Learning. The learning style can be defined as the way a student prefers to acquire, process, and retain the knowledge received from external sources. The prominent learning style classification model is called the VAK model. According to this theory visual, kinesthetic and auditory are the three major kinds of learning styles. Many kinds of research have shown that people prefer more than one way of learning and memorizing, hence categorizing a person as one of the above types as done in traditional methods is not reliable. A method that is used to identify learning styles more accurately is required. Machine learning is applied to achieve our aim in the most efficient way. Once we have accurate information about learning styles and methodologies, we can use it to suggest career options and predict outcomes. This research mainly aims at predicting the learning style combinations of students and suggesting a field of a domain using algorithms like k-means, SVM, and decision tree.

3. COMPARATIVE ANALYSIS

Sr. No.	Paper	Best classification technique	Other techniques tested	Dataset	Performance
1	Personalized Career Path Recommender System for Engineering Stu- dents.[1]	PCRS technique	Fuzzy-Logic System	Personal Information, Academic Performance and Personality Type	91% specificity and 90% average accuracy
2	Predicting Students' Performance using Generative Adversarial and Deep Support Vector Machine [2]	ICGAN-DSVM algorithm	Methodology of ICGAN DSVM	Student Performance, IGAN and CGAN datasets	Accuracy not mentioned
3	Prediction of students' academic performance using machine learning algorithms [3]	RF, NN and SVM algorithms	DM Model, Predictive Model, Descriptive model	Student Information System (SIS) of a State University, Turkey	Accuracy and susceptibility
4	Predicting Stu- dents' Academic Performance at Secondary and Intermediate Lev- el Using Machine Learning [4]	Recurrent neural net-works	Algorithms of Machine Learning, classifiers based on Deep-learning and regression models	BISE Peshawar, of SSC & HSSC were acquired from 6 different Boards	Accuracy and susceptibility
5	A Machine Learning Based Approach for Recommending Courses at Graduate Level [5]	Xavier initialization	ML algorithms	From 200 students engineering colleges with students' recent semesters' performances, interests	73% accuracy
6	Career Recommendation Systems using Content based Filtering [6]	Career Recommendation frameworks	Content based filtering algorithm, Natural Language Processing	Percentage in OS, Algorithms and design, Average percentage in programming, software engineering, electronics, computer architecture, mathematics.	No accurate accuracy
7	An Intelligent Ca- reer Guidance System using Ma- chine Learning [7]	K-Nearest Neighbor	Confusion Matrix, SVM, Naïve Bayes	500 unique values with skillset of candidates using multitask classification technique	No accurate accuracy
8	Analysis Optimization K-Nearest Neighbor Algorithm with Certainty Factor in Determining Student Career.[8]	K- Nearest Neighbor	Certainty Factor method	Three datasets Computer Engineering network, software Engineering, Multimedia are used	93.83%
9	Career Trajectory Prediction based on CNN [9]	CNN Model Architecture	RNN model, DCNN model	Data Castle China game dataset	Accuracy and susceptibility
10	Analyzing the learning of individual and suggesting field of study using Ma- chine Learning.[10]	SVM and Decision tree	SVM and Decision tree	Student Survey conducted in Amrita Vishwa Vidyapeetham University	92.8% accuracy

4. DIFFERENT ALGORITHMS USED

4.1 Decision trees

Tree based learning calculations are considered to be one of the leading and generally utilized directed learning strategies. Tree based strategies enable prescient models with tall precision, steadiness and ease of elucidation. Not at all like straight models, they outline nonlinear connections very well. They are versatile at tackling any kind of issue at hand (classification or relapse). Strategies like choice trees, irregular timberland, angle boosting are being prevalently utilized in all sorts of information science problems. Hence, for each examiner it's critical to memorize these calculations and utilize them for demonstrating. Choice tree may be a sort of administered learning calculation (having a pre-defined target variable) that's for the most part utilized in classification issues. It works for both categorical and nonstop input and yield factors. In this procedure, we part the populace or test into two or more homogeneous sets (or sub-populations) based on most critical splitter / differentiator in input variables

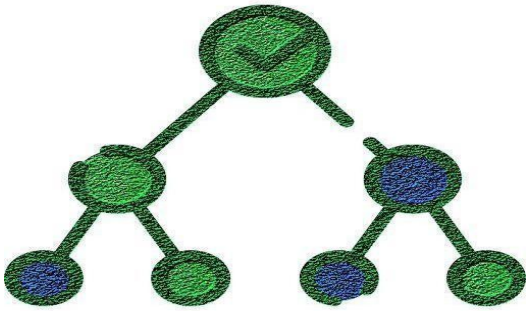

Fig -1: Decision tree

Types of choice tree is based on the type of target variable we have. These are of two types:

1. Categorical Variable Choice Tree: Choice Tree which has categorical target variable at that point it called as categorical variable choice tree. Illustration:- In over situation of understudy issue, where the target variable was "Student will play cricket or not" i.e. YES or NO.

2. Continuous Variable Choice Tree: Choice Tree has persistent target variable at that point it is called as Continuous Variable Choice Tree .

4.2 Support Vector Machine

Support Vector Machine (SVM) may be a directed machine learning calculation which can be utilized for both classification or relapse challenges. In any case, it is generally utilized in classification issues. In this calculation, we plot each information thing as a point in n dimensional space (where n is the number of highlights you have) with the esteem of each highlight being the esteem of a specific facilitator. At that point, we perform classification by finding the hyper-plane that separates the two classes very well. Support Vectors are essentially the coordinates of person perception. Back Vector Machine could be a wilderness which best isolates the two classes (hyper-plane/ line). In SVM, it is easy to have a straight hyper-plane between these two classes. But, another burning address which emerges is, ought to we ought to add this include physically to have a hyper-plane. No, SVM features a procedure called the kernel trick. These are capacities which takes moo dimensional input space and change it to the next dimensions.

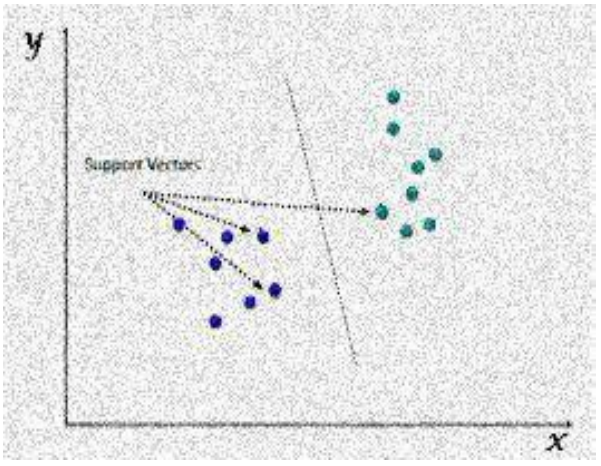

Fig -2: Support Vector Machine

4.3 One Hot Encoding

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 isnot 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 < ret < fight. As this is ordinally characteristic and is usually not required an desired and so OneHot encoding is required for correct representation of distinct elements of a variable. It makes representation of categorial variables to be more expressive.

4.4 XG Boost

XGBoost is an open-source program library which gives the angle boosting system for C++, Java, Python, R, and Julia. It works on Linux, Windows, and macOS. From the venture depiction, it points to supply a "Adaptable, Convenient and Disseminated Slope Boosting (GBM, GBRT, GBDT) Library". Other than running on a single machine, it too bolsters the disseminated preparing systems Apache Hadoop, Apache Start, and Apache Flink. It has picked up much popularity and consideration as of late because it was the calculation of choice for numerous winning groups of a number of machine learning competitions XGBoost at first begun as a investigate extend by Tianqi Chen as portion of the Dispersed (Profound) Machine Learning Community (DMLC) bunch. At first, it started as a terminal application which may be designed employing a libsvm arrangement file. After winning the Higgs Machine Learning Challenge, it got to be well known within the ML competition circles.

Before long after, the Python and R bundles were built and presently it has bundles for numerous other dialects like Julia, Scala, Java, etc. This brought the library to more designers and got to be prevalent among the Kaggle community where it has been utilized for a expansive number of competitions. It before long got to be utilized with numerous other bundles making it easier to utilize within the particular communities It presently has integrative with scikit - learn for Python users, conjointly with the caret bundle for R clients. It can too be coordinates into Information Stream systems like Apache Start, Apache Hadoop, and Apache Flink utilizing the dreamy Rabit and XGBoost4J. The working of 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 Page | 31 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 scikitlearn 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 best highlights that the execution of the calculation gives are: Programmed dealing with of lost values with meager mindful usage, and it gives square structure to advance parallel development of tree and proceeded preparing which underpins advance boost an as of now fitted demonstrate on the fresh data. Gradient boosting could be a method where unused models are made that can anticipate the errors or remains of past models and after that included together to create the ultimate expectation. they utilize angle descendent calculations to decrease misfortune amid including of unused models. They back both classification and relapse sort of challenges. Within the preparing portion by and large an objective work is characterized. Characterize an objective work and attempt to optimize it. obj=i=1 Σl(yi,yi^(t))+ i=1 Σ Ω(fi

5. METHODOLOGY

5.1 PROCESS FLOW

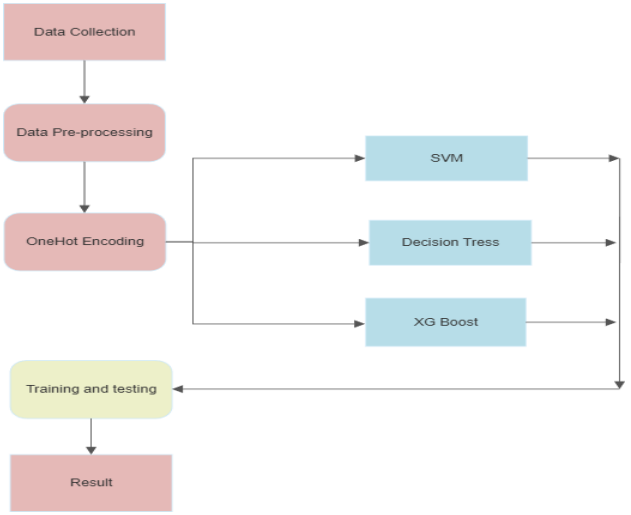

Chart -1: System Architecture

5.2 DATA COLLECTION

Collection of data is the foremost and most important task of any machine learning project. This is the data that we tend to feed the algorithms to obtain the required predictions and results. The algorithm's efficiency and accuracy depends upon the correctness and the quality of the data that is feed-ed. For student career prediction several parameters are taken into consideration like students educational scores in varied subjects, specializations, programming and analytical capabilities,

memory, personal details like hobbies, interests, sports, competitions, hackathons, workshops, certifications, books interested and many more. As a lot of these factors play important role to decide students progress towards career track, all of these are taken into thought. Data is collected from different sources. 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 some from college alumni databases.

5.3 DATA PRE-PROCESSING

Collecting the data is a necessary task, but making that data insightful is also necessary. Data will be collected from various sources, and there may be a lot of invalid data values and unwanted data. Cleaning all these data and replacing them with appropriate or approximate data and removing null and missing data are the basic steps in preprocessing data. Even the collected data may contain complete garbage values. It may not be in the format or in the way it should 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 specific format. All of these things are taken into consideration while data pre-processing.

5.4 APPLICATION OF ALGORITHMS

The next step after data pre-processing is applying the algorithms to the data and obtaining the results and observations. After applying the algorithms we need to analyze and try to improve the accuracy of the algorithm.

5.5 TRAINING AND TESTING

Finally, after data processing and training, the next task is obviously testing. This is where performance, quality of data, and required output all appear. Of the massive data set collected, 80 percent of the data is used for training and 20 percent of the data is reserved for testing. Training the machine is the process of making it learn and giving it the capability to make further predictions based on the training it took. Testing means having a predefined data set and output also previously labeled and the model is tested whether it is working properly or not. If the maximum number of predictions are correct, the model will have a good accuracy percent- age and be reliable to continue using otherwise better to change the model.

6.CONCLUSIONS

Competition in today's world is heavily propagating day by day. Especially since it's too common in the current day's specialized world. So as to compete and achieve the ambition of children. It needs to be mapped and organized from the original stages of their education. So it's certifiably important to constantly estimate their performance, identify their interests, and estimate and mark them to see how close they are to their goal and whether they are on the correct path for their target. This assists them with self-improvement and motivation and makes the right choice. It enables them to choose a better career path even if their qualifications are insufficient. It also helps in pre-evaluating themselves before reaching the career peak point. Not only that, but recruiters consider aspirants when hiring new employees and estimate their diverse parameters and draw a final conclusion to take an employee or not and if selected, dig up the best-suited role and career field for them. After reckoning all the factors, the total number of parameters that were taken into consideration to solidify the model.

An important web operation can be developed where inputs aren't given directly rather student parameters are taken by assessing them through different assessments and quizzing. Specialized, analytical, logical, memory-based, psychometry and common awareness, interests, and skill-based tests can be aimed and parameters are composed through them so that effects will be indisputably on-target and the system will be more dependable to exploit.

Also, decision trees have many confines like overfitting, no pruning, and lack of capability to deal with null and missing values, and many algorithms have problems with a huge number of values. All these can be taken into consideration and indeed more dependable and more accurate algorithms can be used. Also, the design will be more important to depend upon and indeed more effective to depend upon.

REFERENCES

[1] Kwok Tai Chui, Ryan wen liu, Mingbo Zhao, "Predicting Students' Performance With School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine".

[2] Nunsina, Tulus, Zakarias Situmorang," 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).

[3] Vignesh S, Shivani Priyanka C, Shree Manju H, Mythili, "An Intelligent Career Guidance System using Machine Learning", 2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS).

[4] Namratha M, Tanya V Yadalam, Vaishnavi M Gowda, Vanditha Shiva Kumar, Disha Girish," based Filtering", Proceedings of the Fifth International Career Recommendation Systems using Content Conference on Communication and Electronics Systems (ICCES 2020) IEEE Conference Paper.

[5] John Britto , Sagar Prabhu, Abhishek Gawali and Yogesh Jadhav, "A Machine Learning Based Approach for Recommending Courses at Graduate Level", Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019) IEEE Xplore Part Number:CFP19P17-ART;ISBN:978-1-7281-2119-2.

[6] Manar Qamhieh, (Member, Ieee), Haya Sammaneh, And Mona Nabil Demaidi, (Senior Member, IEEE), "Personalized Career-Path Recommender System for Engineering Students", M. Qamhieh et al.: PCRS for Engineering Students.

[7] Mustafa Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms", Yağcı Smart LearningEnvironments(2022),https://doi.org/10.1186/s40561-022-00192-z.

[8] Shah Hussain1, Muhammad Qasim Khan, "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning", Received: 22 June 2020 / Revised: 12 April 2021 / Accepted: 21 April 2021 © The Author(s), under exclusive license to Springer-Verlag GmbH Germany, part of Springer Nature 2021.

[9] Miao He, Dayong Shen, Yuanyuan Zhu, Renjie He, Tao Wang, Zhongshan Zhang, "Career Trajectory Prediction based on CNN", Authorized licensed use limited to : University of Glasgow. Downloaded on June 02,2020 at 06:05:11 UTC from IEEE Xplore.

[10] Ananthu S Kuttattu, Gokul G S, Hari Prasad, Jijith Murali, Lekshmi S Nair," Analysing the learning style of an individual and suggesting field of study using Machine learning techniques", doi 10.1109/ICCES45898.2019.9002051.

[11] Priyanka Shahane, Deipali Gore "A Survey on Classification Techniques to Determine Fake vs. Real Identities on Social Media Platforms," IJRDT, 2018.

[12] Priyanka Shahane, Sneha Kasbe, Rohini Kasar,M.P. Navle "Ontology Based Information Retrieval System Using Multiple Queries For Academic Library," IRJET, 2018

[13] P. Shahane, "Campus Placements Prediction & Analysis using Machine Learning," 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), 2022, pp. 1-5, doi: 10.1109/ESCI53509.2022.9758214.

[14] P. Shahane, "Detection of Fake Profiles on Twitter using Random Forest & Deep Convolutional Neural Network," International Journal of Management, Technology And Engineering, Volume IX, Issue VI, JUNE/2019.