

An Intelligent Career Guidance System using Machine Learning

Dahanke Ajay-1, Shinde Nilesh-2, Dhagate Anirudh-3, Shaikh Huzaif-4

1-4Student, Computer Engineering, Loknete Gopinathji Munde Institute of Engineering Education & Research, Nashik, Maharashtra.

Abstract - Most of the students across the globe are always in confusion after they complete higher secondary and therefore, the stage where they need to settle on an appropriate career path. The students don't have adequate maturity to accurately understand what a private has got to follow so as to decide on a congenial career path. As we labor under the stages, we realize that each student undergoes a series of doubts or thought processes on what to pursue after 12th, which is that the single tallest question. Then comes the subsequent agony whether or not they have essential skills for the stream they need chosen. Our computerized career counselling system is employed to predict the acceptable department for a personal supported their skills assessed by an objective test which will be based on their interest chosen and their 12th top two subjects marks scored. If one completes their online assessment, which we've created in our system, then automatically they'll find yourself in choosing an appropriate course which is able to also reduce the failure rate by choosing a wrong career path.

Keyword: Career Guidance, Machine Learning Algorithms, OCR (Optical Character Recognition), Image Processing, K-Nearest Neighbor, Career Counselling.

1. INTRODUCTION

When it involves choosing a career, it's not only about what course you decide on, it's quite what you would like to become after your graduation. Counseling is more about knowing and understanding yourself and your capabilities and talents. It's now every student gets plenty of guidance from various circles (parents, teachers, other educational specialists, etc.), and accordingly, the scholar decides about which course they need to hitch. Many times, we've come upon a situation where a student opts for a course/stream and later regrets for having chosen the one. To quote an example, there's a myth that one who does good and scores the best marks in 12th-grade chemistry will tend to decide on chemical engineering because they're good in chemistry, however in point of fact that's not the case. We had multiple rounds of deliberation with students who are currently doing their engineering and students who are currently in 11th and 12th grade. Then we came up with a plan of providing an objective Assessment of one's skill set and their top two marks caliber that recommend the correct stream to settle on and hence we picked this as our problem statement and began thinking through how we are able to help the students in addressing this question.

As a primary step, we came up with broader skill sets that are strongly essential for every department in engineering like technology and engineering, Electronics and Communication Engineering, Electrical and Electronics Engineering, applied science, etc. Depending upon one's mark within the objective assessment we've created, we'll analyze their skill sets and predict which department is suitable for a private. If one uses this functional chart to answer of these questions, the failure rate will drastically reduce in discovering the incorrect choice. Our pointed questions will identify the core strength of student's particular skill sets.

2. LITERATURE SERVEY

A prediction for performance improvement using classification Mohammed M. Abu Tair, Alaa M. El-Halees This paper investigates the educational domain of data mining using a case study from the graduate student's data collected from the college of Science and Technology Khanyounis. The data include fifteen years period [1993- 2007]. It showed what kind of data could be collected, how could we preprocess the data, how to apply data mining methods on the data, and finally how can we benefit from the discovered knowledge. There are many kinds of knowledge can be discovered from the data. In this work we investigated the most common ones which are association rules, classification, and clustering and outlier detection. The Rapid Miner software is used for applying the methods on the graduate student's data set

[1]. Data Mining: Performance Improvement in Education Sector Using Classification and Clustering Algorithm Gadde Shravya Sree DATA mining sometimes is also called knowledge discovery in databases (KDD). Student retention has become an indication of academic performance and enrollment management. Here, potential problem will be identified as earlier. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. One of the most useful data mining techniques for e-learning is classification. Classification maps data into predefined groups of classes. It is often referred to as supervised learning because the classes are determined before examining the data. The prediction of students' performance with high accuracy is more beneficial

for identifying low academic achievements students at the beginning. To improve their performance the teacher will monitor the students' performance carefully

[2] Data Mining: A prediction for performance improvement using classification Brijesh Kumar Bhardwaj, Saurabh Pal The ability to predict a student's performance is very important in educational environments. Students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. A very promising tool to attain this objective is the use of Data Mining. Data mining techniques are used to operate on large amount of data to discover hidden patterns and relationships helpful in decision-making.

[3] An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree Md. Hedayetul Islam Shovon, Md. Hedayetul Islam Shovon GPA still remains the most common factor used by the academic planners to evaluate progression in an academic environment. Many factors could act as barriers to student attaining and maintaining a high GPA that reflects their overall academic performance, during their tenure in university. These factors could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance. With the help of clustering algorithm and decision tree of data mining technique it is possible to discover the key characteristics for future prediction.

[4] Online Career Counseling System using Adaptive e-Learning Kazi Fakir Mohammed, Sushopti Gawade, Vinit Nimkar Adaptive learning can be termed as an educational method that tends to bring interactive teaching devices in the form of computers which in turn accustom the exhibition of educational entities according to the student learning necessities which is depicted by their acknowledgment to tasks and queries. The technology therefore includes concepts which are derived from numerous fields of study including psychology, computer science and education.

2.1 MOTIVATION

The main motive behind developing this project helps out students, recent graduates, and professionals who wish to form a career switch, make the correct decisions that might benefit them without prying the effort of direction. The system may be a cost-cutting, time-efficient method developed to guide people who are unable to urge proper guidance.

2.2 PROBLEM DEFINITION

Students across the planet are always in confusion after they complete higher secondary and also the stage where they need to settle on an appropriate career path. At the age of 18, the scholars don't have adequate maturity to accurately understand what a personal must follow so as to decide on a congenial career path. When it involves choosing a career, it's not only about what course you decide on, it's over what you wish to become after your graduation. Direction is more about knowing and understanding yourself and your capabilities and skills. it's now every student gets plenty of guidance from various circles (parents, teachers, other educational specialists, etc.) to unravel this problem we develop new system Our computerized career counselling system is employed to predict the appropriate department for a personal supported their skills assessed by an objective test. If one completes their online assessment, which we've created in our system, then automatically they're going to find yourself in choosing an appropriate course which is able to also reduce the failure rate by choosing a wrong career path.

3. SOFTWARE REQUIREMENTS SPECIFICATION

3.1 EXISTING SYSTEM

We have an existing Manual Career system with human counsellors to blame, but this technique is out dated and contains a lot of drawbacks. Existing system has the subsequent drawbacks:

1. Focus on one field of interest, without considering relevant skills.
2. Lack of clear goals or ambiguity within the interest of a private.
3. Peer pressure, insistence from family/friends/society.
4. Insufficient/partial knowledge, lack of required skillsets, and fear of indecision.
5. These fundamental drawbacks of existing career guidance systems are important to be addressed and overcome

3.2 PROPOSED SYSTEM

We are developing a web-based application aimed to beat the traditional career guidance processes and methods. Our project uses trending technologies like OCR, and machine learning algorithms to work out the simplest possible career pathway for a personal.

The proposed system has following advantages:

1. Easily accessible and user-friendly web-based interface, which is hassle-free.
2. Saves time and money, as there aren't any physical/financial obstacles in using the system.
3. Improved quality of career guidance methods, and a notch above traditional/in-person counselling practices in efficiency and user- satisfaction.
4. Near-accurate predictions/suggestions supported detailed analysis of the user's performance and skills.
5. Provides real-time suitable career predictions and areas for improvement derived through assessment outcomes.
6. Might help reduce career anxiety among youngsters

4. SYSTEM ARCHITECTURE

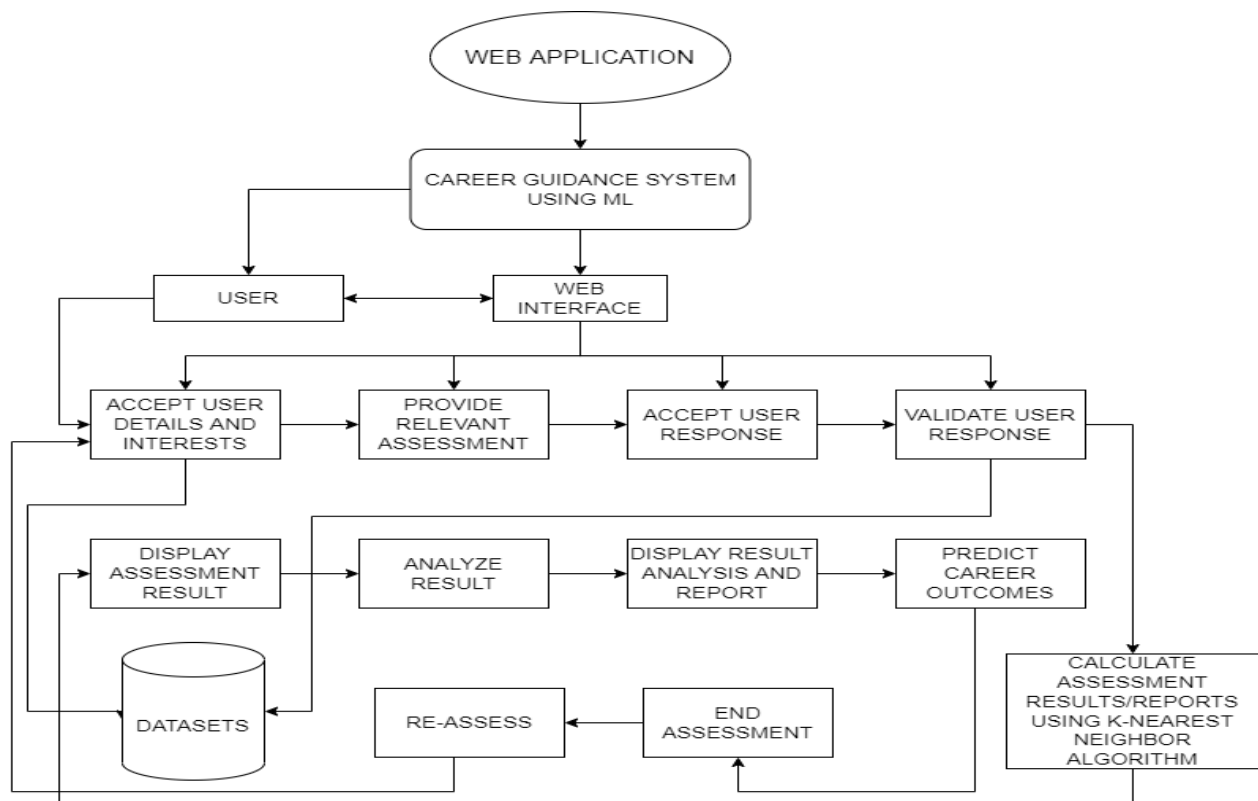


Fig 1: System Architecture

5. MODULES

Our system consists of three modules.

- 5.1. Skill Set Assessment Module.
- 5.2. Predication Module.
- 5.3. Result Analysis

5.1 SKILL SET ASSESSMENT MODULE.

In this module the candidate take up an assessment which having combination of psychological and core skill-oriented questions. This module is designed and developed with the help of various web technologies such as HTML 5, CSS 3 and JavaScript. Hyper Text Markup Language 5 (HTML version 5) is a markup language used to design documents to get displayed in a web browser. It gives a skeletal structure to the document and so the document will be static if we only use HTML. Cascading Style Sheets 3 (CSS version 3) is a style sheet language, which is especially used for adding styles and designs to the HTML document, so that the representation becomes better. JavaScript is a client-side scripting language that is mainly used for adding interactivity to the HTML web page. This makes the HTML page more of a dynamic page.

HTML 5 and CSS 3 plays a vital role in the front-end development and JavaScript plays a vital role in back-end development. Each question will be displayed separately with multiple choices. The validation part will be done with the help of JavaScript where each choice in a question will be having different weightage according to the best suitable answer. The validation will be done in a skill-wise manner where the final result will be displayed in a skill-wise manner respectively.

5.2 PREDICATION MODULE.

The prediction module is the primary and core module among all the other modules. This module is built with various technologies such as machine learning algorithms, API and datasets for training the machine learning model. All the implementations are done with the help of python programming language. Python is a general-purpose programming language which can be used in almost every part of the implementation. The framework is predominantly developed with the help of python in the prediction module.

K-Nearest Neighbors is the machine learning algorithm used for classification purposes. K-Nearest Neighbors is a supervised machine learning algorithm used to classify the target values with the help of determining the distance between each neighbor using any formulae like Euclidean distance, Minkowsky, cosine similarity measure, chi square and correlation. K-Nearest Neighbor is well suited for classification problems. Even though there are many classification algorithms like Support Vector Machine, Random Forest, and Navies Bayes etc. K-Nearest Neighbor is the one and only algorithm which has an accuracy of more than 90% with the dataset we have created of our own through various methodologies.

K-Means Clustering is the machine learning algorithm used for clustering pumices. K-Means Clustering is an unsupervised machine learning algorithm used to partition n observations into K clusters in which each value combines with the cluster with the nearest mean. In this framework. K-Means Clustering is specifically used to group the departments which are most appropriately mapped with the candidate's performance and to provide secondary recommendations.

Flask API (Application Programming Interface) is used in the framework for having a communication between the front end and back end. An application programming interface is used to send requests and receive responses to and from the application. In our application. The scores obtained from the states assessment module is passed to the machine teaming model via the flask API. The request in this scenario is to receive a suitable department with respect to the candidate's performance. The response in this scenario is to provide the predicted department with respect to the candidate's performance. The implementation of this application programming interface part is done with the help of python. The dataset used for the machine learning model is developed manually as there was no appropriate data available related to the core concept of this application. All the values present in the dataset are of only numerical values. The dataset basically contains more than 500 rows which means 500 unique values with several features and target variables. They are seven different features available in the dataset which consists of core skills and sub skills such as analytical skills, logical reasoning skills. Mathematical skills. Problem solving skills. Programming skills. Creativity skills and hardware skills. Among these seven skills. Some are considered as core skills for a specific department while those same skills are also considered as sub skills for a different department. As this is a multi-class problem. We have used multi class classification technique while developing the K-Nearest Neighbor model. Hence. There will be several target labels present in the damsel. In this dataset. There are five different target labels available each representing a specific department. For the machine learning model. 80% of the data present in the dataset is taken for training purposes and 20%of die data present in the damsel is taken for validation and testing purposes.

5.3 RESULT ANALYSIS

The performance of the machine learning model can be determined by a concept called confusion matrix. A confusion matrix is represented in the form of a table in which there will be four different values present in the matrix such as true

positive. True negative. False positive and false negative. This can be used on the test dataset for which the true values are already known.

1. True positives (TP) are the value in which the examples are correctly determined as positive.
2. False positives (FP) are the value in which the examples are negative but are actually determined as positive.
3. True negatives (TN) are the value in which the examples are correctly determined as negative.
4. False negatives (FN) are the value in which the examples are positive but are actually determined as negative.

The confusion matrix determines the performance of the classification model by calculating precision, recall, accuracy, F-measure and error-rate. The following are the formula for each of the above-mentioned performance measures:

$$TP = TP / (TP+FN)$$

$$FP = FP / (FP+TN)$$

$$\text{Precision} = TP / (TP+FP)$$

$$\text{Recall} = TP / (TP+FN)$$

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN)$$

$$\text{F-measure} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

$$\text{Error-rate} = 1 - \text{accuracy}$$

6. PERFORMANCE AND MEASURES FOR CLASSIFICATION TECHNIQUES.

A performance measure is a numeric description of an agency's work and the results of that work. performance measures are based on data, and tell a story about whether an agency or activity is achieving its objectives and if process is being made towards attaining policy or organizational goals productivity, profit margin, scope and cost are some examples of performance metrics that a business can track to determine if target objectives and goals are being met. There are different areas of a business and each area will have its own key performance.

Classification is referring to a predictive modelling problem where a class label is predicted for a given examples of input data examples of classification is problem include given an example, classify if its spam or not. Given a handwritten character, classify it as one of the known characters.

Following figure shows performance for classification Techniques Classification of techniques using classification of machine learning algorithms like KNN (K-Nearest Neighbor) is one of the simple machine learning algorithms based on Supervised Learning techniques

TABLE I
 PERFORMANCE MEASURES FOR CLASSIFICATION TECHNIQUES

Classification Technique	Accuracy	F - measure	Error - rate
KNN	0.9410	0.9213	0.01964
SVM	0.8632	0.9018	0.03154
Naive Bayes	0.8714	0.8835	0.06127

Fig 2: Performance Measure for Classification

6.1 CLASS DIAGRM IN UML

Shows a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects. First block is representing the generated skill assessment test, which is then validated and sent to career guidance recommendation engine. Next block does skillset analysis and the system proceeds to recommendation, and further decision depends on user

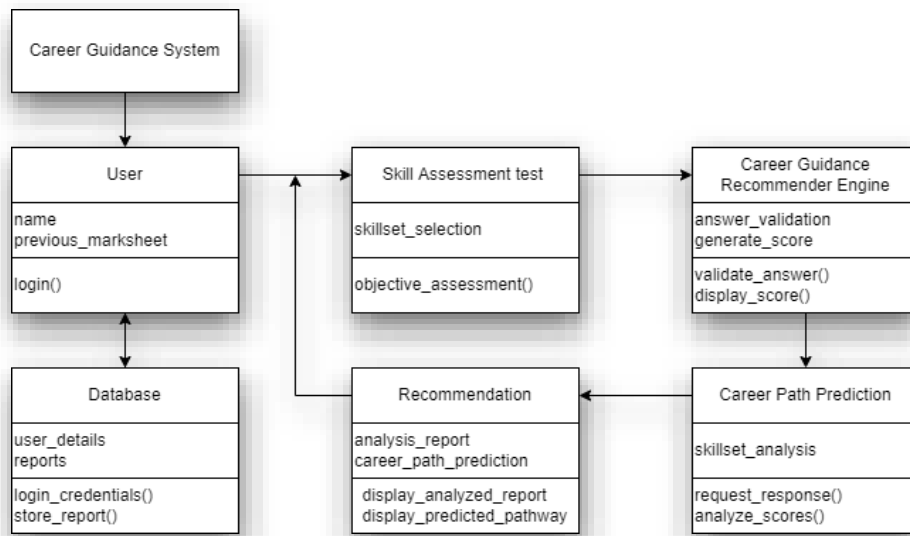


Fig 3: Class Diagram in UML

7. ALGORITHM

Step1: Start.

Step2: OCR: for extracting text of subject respect to marks from result (image/PDF)

Step3: The candidate takes up an assessment, which will be having a combination of psychological and core skills oriented questions.

Step4: Validating candidate answers and displaying skill set wise marks.

Step5: Requesting a response (suitable department) by sending candidates score in each skill set.

Step6: Take input as:

- 1) What, the efficiency of the clustering can be determined by Within Cluster Sum of Squares (WCSS). WCSS can be calculated using the formula given above
- 2) when/Where the confusion matrix determines the performance of the classification model by calculating precision, recall, accuracy, f-measure and error-rate. The following are the formula for each of the above- mentioned performance measures:
- 3) Who, the performance of the machine-learning model can be determined by a concept called confusion matrix. A confusion matrix is represented in the form of a table in which there will be four different values present in the matrix such as true positive, true negative, false positive and false negative. This can be used on the test dataset for which the true values are already known.
- 4) Why, True positives (TP): Are the value in which the examples are correctly determined as positive.

False positives (FP): Are the value in which the examples are negative but are actually determined as Positive.

True negatives (TN): Are the value in which the examples are correctly determined as negative.

False negatives (FN): Are the value in which the examples are positive but are actually determined as Negative.

Step7: Providing a detailed analysis of the predicated department with candidate's performance.

Step8: Stop





8. CONCLUSION

In the system, we have designed and developed a web-based application for a career guidance system which provides suitable recommendations for a candidate in choosing an appropriate department. The recommendation provided in the proposed system is more accurate than the existing career guidance system. We have used the K-Nearest Neighbor algorithm to classify the skill sets of the candidate and predict a suitable department with respect to the performance of the candidate and we have also used K-Means Clustering algorithm and the clusters formed is by splitting the students' scores of the particular skill set and determining the rate of success for various departments in every cluster.

9. REFERENCES

- [1]. A.M. El-Halee's, and M.M. Abu Tair, "Mining educational data to improve students' performance: A case study," International Journal of Information and Communication Technology Research, 2011, pp. 140-146. M. tech Er. Rimmy Chuchra. "Use of data mining techniques for the evaluation of student performance: a case study". International Journal of Computer Science and
- [2]. Gadde Shrivya Sree and Ch. Rupa. "Data Mining: Performance Improvement In Education Sector Using Classification And Clustering Algorithm" IV/IV B.Tech, Dept. of CSE, VVIT, Nambur, Guntur dist, A.P, India.
- [3]. Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining:" A prediction for performance improvement using classification & (IJCSIS) International Journal of Computer Science Information Security, 9(4), April, 2011
- [4]. Md. Hedayetul Islam Shovon and Mahfuza Haque." An approach of improving student's academic performance by using k-means clustering algorithm and decision tree". (IJACSA) International Journal of Advanced Computer Science and Applications, 3(8):146-149, August, 2012.
- [5]. Kazi Fakir Mohammed and Sushopti Gawade and Vinit Nimkar "Proceeding International conference on Recent Innovations in Signal Processing and Embedded Systems (RISE -2017) 27-29 October, 2017"
- [6]. Md. Yeasin Arafath, Mohd. Saifuzzaman, Sumaiya Ahmed and Syed Akhter Hossain "2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018"
- [7]. Bharat Patel, Varun Kakuste, Magdalini Eirinaki "2017 IEEE Third International Conference on Big Data Computing Service and Applications"
- [8]. Khairunnas Jamal, Rahmad Kurniawan, Ilyas Husti, Zailani, Mohd Zakree Ahmad Nazri and Johar Arifin Predicting Career Decisions Among Graduates of Tafseer and Hadith
- [9]. Kasem Seng and Akram M.Zeki "2014 3rd International Conference on Advanced Computer Science Applications and Technologies"

10. BIOGRAPHIES

	<p>Dahanke Ajay is a final year Computer Engineering student at Savitribai Phule Pune University. His current interests include Software Development, Artificial Intelligence, Machine Learning, Data Science, and Cloud Computing.</p>
	<p>Shinde Nilesh is a final year Computer Engineering student at Savitribai Phule Pune University. His current interests include Artificial Intelligence, Machine Learning, Cyber Security, Ethical Hacking and Penetration Testing, Cloud Computing & Security.</p>
	<p>Dhagate Anirudh is a final year Computer Engineering student at Savitribai Phule Pune University. His current interests include Cryptography, Data Analytics, NFT and Data Mining.</p>
	<p>Shaikh Huzaif is a final year Computer Engineering student at Savitribai Phule Pune University. His current interests include Artificial Intelligence, Machine Learning, Animation, Cryptography and Robotics.</p>