Student Skill Identification and Career Guidance System Based on Machine Learning.

Uditha Janadara
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
udithajandara.h@gmail.com

Dinindu Koliya Harshanath Webadu Wedanage
Smart Infrastructure Facility
University of Wollongong
Wollongong, Australia
dkhww937@uowmail.edu.au

Samantha Thelijjagoda
SLIIT Business School
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
samantha.t@sliit.lk

Archchana Kugathasan
Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
archchana.k@sliit.lk

***Abstract*—Information technology is a massive industry consisting of a wide range of tech stacks and professional experts. Various career paths such as full stack developer, backend developer, front end developer, and many more careers are observed in the information technology industry. However, a lack of awareness of own talent may lead to poor decisions. On the other hand, Employees change their careers regularly, seeking for the most suitable and comfortable career for them, specially in IT industry. Therefore, identifying the capabilities and providing effective guidance to newcomers is really important.This research focuses on implementing a career guidance system, which consists of career path recommendations and career opportunities recommendations purposing of assist fresh IT undergraduates to by providing a effective career guidance. For this purpose, data were collected from IT industry related job postings using some keyword extraction methods. Several models were implemented such as Support vector machine classification, Logistic regression, Naive bayes and their performance were compared using model accuracy to select the accurate model.For the prediction of most suitable career path, Support vector machine with Rbf kernel resulted with highest accuracy of 83%. The ultimate goal of this research is to help IT undergraduates to idntify their skills and provide a stable start for their career life. Keywords: career recommendation, career guidance, classification, skill identification**

I. Introduction

The IT industry has shown rapid growth over the last few decades with a number of growing career opportunities. Since the IT industry provides a wide range of career paths, each career requires a unique combination of technical and soft skills.According to the[ref], their is a chance of getting wrong decision on career path selection due to some external factors such as family pressure, friend's career and social pressure.Therefore having a clear understanding about their own capabilities and personal interest is important while choosing a career path as a fresher. Many Students start their first career after they completed the higher education or during the higher education. This stage can be considered as a key point of the person's career life, since it's the first time they take the decision of their career path.Therefore, it is important to provide appropriate guidance to the students prior to their entry into the industry. Number of research have been conducted on focusing on career path recommendation and career guidance.

Vignesh S, Shivani Priyanka and Shree Mangju have come up with a career guidance system for the engineering students based on the their skills.[ref]. They have conducted an assessment to evaluate the student skills, which includes psychological and the core-skill oriented questions. Students have been clustered into different departments (computer science engineering, electric and electronic engineering, electronic and communication engineering, and mechanical engineering) based on the identified skills with the help of the K-means clustering algorithm. Tajul Rosli Razak, Muhamad Arif Hashim and Noorfaizalfarid Mohd Noor have implemented a career path recommendation system using fuzzy logic. Their study was focused on the computer and the mathematical students.In addition to the student technical skills, they have considered the student personality as well. Personality and skills data were collected through a series of interviews. Skills have been labeled with three linguistic variables which are "Good", "Medium", and "Weak".Considered careers are also labeled, respective to the student, with another three linguistic variables which are "Yes", "No", and "maybe.

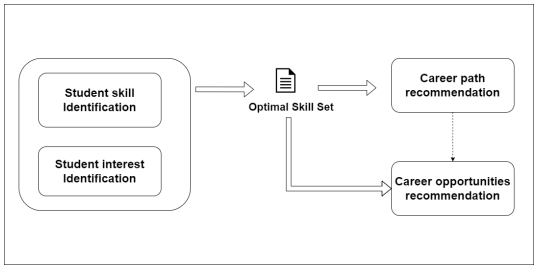
Ashutosh Shankhdhar, Akash Agrawal and Deepak Sharma have come up with an intelligent decision support system using the decision tree algorithm. They have collected student academic performance through a survey and some quizzes to collect data about student personality. In this study also considered about student personality addition to the technical skills. Decision tree algorithm have been used to predict the most suited career for the student. Bharat Patel, Varun Kakuste, and Magdalini Eirinaki carried out a career path recommendation framework, based on the skills extracted from the resume.They have followed a hybrid approach with content based with

collaborative filtering to avoid the cold start. In terms of candidate profile classification, there are numerous research have been conducted using text classification techniques. Tere Gonzalez, Pano Santos, and Fernando Orozco have come up with a Adaptive Employee Profile Classification system which classify the employees according to their skills and experience by analyzing their resume. Another resume classification system have implemented which use the LinkedIn profile descriptions to classify the candid ate profile. Razkeen Shaikh, Nikita Phulkar, and Harsha Bhute have implemented a system to classify the candidate by analyzing the profile using text categorization and semantic analysis. Recommendation have been given by calculating the similarity between candidate skills and the required skill set.
By examining existing career guidance systems, the student skill identification phase is usually performed by analyzing student academic data. When the student grades are evaluated module-wise, in most cases it won't able to get accurate details on individual skills of the student. To overcome this problem, this study have proposed a content tagging concept for the examinations which can evaluate each skill individually. It will help to provide a efficient career guidance for the student based on more reliable data of student performance. Most of the previous studies have paid attention to providing career path recommendations covering a wide range of industries and careers. Providing efficient career guidance considering a specific industry is quite challenging. To address that challenge, this study proposes a machine learning approach for career path recommendation based on the identification of the student skills and the IT industry-specific requirements.

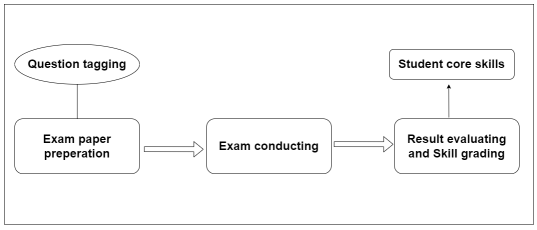
II. Methodology

This section describes the implementation of the career guidance system which consists of three main modules.

- Student skills and personal interest identification
- Career path recommendation
- Career opportunities recommendation



A. Student Skills and Personal Interest Identification



In the initial stage, which is the preparation phase of the exam paper, all the questions in the exam paper will be tagged according to the required skills. Some existing studies have proposed this content tagging concept [ref] for student skill identification, which can be used to address the above-mentioned problem, which is the difficulty of getting accurate data on student performance of individual skills. These skills can be technical skills such as programming languages, frameworks, databases, or soft skills such as problem solving skills, analytical skills, critical thinking skills. One question can be tagged with one or multiple tags. The tags are predefined and available in the preparation of the exam paper.
Below figure demonstrate the structure of the tags.

Skill	Category	type
Java	Language	Technical skill
React Js	Framework	Technical skill
Problem solving		Soft Skill

Second phase, examinations will be conducted through a online examination platform and obtained marks for each and every question will be recorded.
In the 3rd phase, examination results will be evaluated and average mark for each and every skill will be calculated. Based on the average mark, skills will be graded. If the skill have expected grade level or above, it will identified as a core skill of the student. Another important factor mentioned above is the student's personal interest. Even if the student showed better performance in different subject areas, it can't assume that the student is preferred in all of them. This system has implemented a feedback mechanism to collect the student's personal preferences in covered subject areas and any other related subject.Student can rate their experience on covered learning areas or input any other personal interests related to the IT industry. Those ratings will be saved along with previously identified skills. Then, the intersection of the previously identified core skill set and the student's most preferred skill set will be extracted as the optimal skill set for the career recommendation.

to be added - figure - grade table

B. Career Path Recommendation

Career recommendation model has been implemented as a multi-class text classification model to take the student profile as the input and to recommend the most suitable career as the output. Student profile is maintained by the system under four aspects. Skills identified by the system, Skills determined by the student, student career related personal interests, and the personal skills of the student. Here, students are allowed to add skills to their profile since a student can have some skills in addition to what they learn in the academic subject stream. Below table showcase an example of a prepared student profile by the system. Those profile details will be provided to the model as a descriptive input.

Identified Technical Skills	JavaScript, Java, jQuery, React, Node JS, MongoDB, MySql, Django, AWS, Docker, Flutter, Git, Python
Personal Skills	Problem solving, Analytical, Creative thinking
Personal Interest	AI, Machine learning
Technical skills declared by the student	Azure, Datagran

The model was deployed on a django application by implementing the Representational State Transfer Protocol (REST) Application Programming Interface (API) which receives the input from the user and provides the recommendation as the output.In the initial phase, since student skills are not recorded in the system, it leads to a cold start. To avoid that, students will ask to provide their fields of interest or their own skills to the system. Later, they will be fine-tuned along with the learning and evaluation process.

C. Career opportunities recommendation

Career opportunities recommendation module was implemented that focuses on recommending career opportunities in the IT industry by matching the set of identified student skills with the skills extracted from the job postings. Here the predicted career path in the second step will be a optional parameter to recommend the career opportunities since career opportunities recommendation was focused on recommendation of job opportunities based on student skills and the experience level. Student can determine their experience level in the system. Job postings will be categorized in to three categories based on the required experience level.

Job position title/ Description	**Experience level**
Intern/Trainee	Entry level
Associate or 1-5 years exp	Mid-level
Senior/Lead or 5 years < exp	Senior level

Student will have more choices based on the identified skills and if the student decided to go with the recommended career, system will filter out the job posting related to the recommended career. Current job postings were accessed through a API and required skills were extracted. To extract skills from the job postings, resumeparse library have used here. Here, since the system has its own skill pool [table figure], skills were extracted according to that skills and the student defined skills.[ref?]. The skills extracted were saved in each job posting as a JSON value. Then the similarity between the skills extracted from the job posting and the student skill set was calculated using the Jaccard similarity index.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A - Set 1

B - Set 2

J - Jaccard Distance

The Jaccard similarity index compares values for two sets to identify which values are shared and which are distinct. The similarity will be given in range of 0% to 100%. Based on the calculated similarity, most suitable job opportunities will recommend for the student.

To be added: Plot of similarity calculation

D. Data Collection and Preparation

In the data collection phase, it has considered gathering data of various careers in the IT industry and their requirements.In this research, it has been narrowed down to the below careers which are demanding in the computer science.

- Backend Developer
- Cloud Engineer
- Database Administrator
- Data Scientist
- Database Administrator
- Devops Engineer
- Frontend Developer
- Fullstack Developer
- Mobile Application Developer
- Network Engineer
- Systems Analyst
- System Architect
- Software Quality Assurance Engineer
- Security Analyst
- UI/UX Designer

Data set has prepared with the existing data sets available to use which contains job postings in last few years From the selected data sets, IT industry related job postings were separated and since some job postings contains entire job description, skills were extracted using keyword extraction techniques.Gathered data was arranged as required skills along with the career path.Data points which contains null value or insufficient job description has removed. Then, extracted text data were converted to a matrix of token counts using Countvectorizer.Finally the count matrix was transformed to a normalized tf-idf representation using TfidfTransformer. The table below describes the structure of the data set.

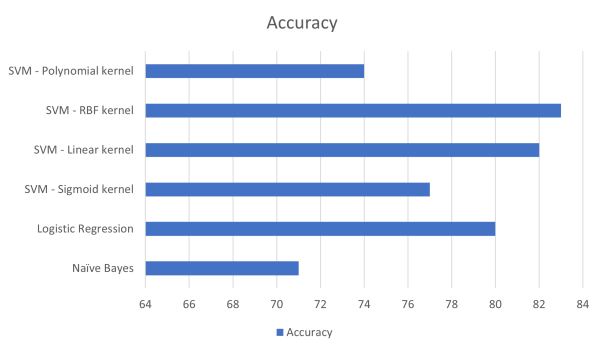
Career	Skills
Frontend Developer	JavaScript, React, HTML, CSS, SCSS, Designing, CI/CD, REST API, Git
Backend Developer	Java, MySql, Spring boot, AWS,Git, problem solving skills, Analytical skills
Mobile Application Developer	Flutter, Firebase, NodeJs, Swift, Cordova, Analytical skills

E. Model Selection and Training

Three machine learning algorithms have been considered in the model selecting phase.

- Naive Bayes
- support vector machine
- Logistic regression

Figures below shows the comparison between different algorithms in terms of accuracy.



In the model selection phase, three main supervised machine learning algorithms were considered for the classification model, which are Naive bayes, Logistic regression and Support vector machine. Support vector machine (SVM) algorithm was tested out with four kernels which are linear, polynomial, sigmoid and rbf. As described in above figure, support vector machine algorithm with the Rbf kernel, shown the highest accuracy, which is 83%. Support vector machine is a supervised machine learning algorithm used for both classification and regression problems.A particular type of Gaussian kernel called an RBF kernel projects high-dimensional data and searches a linear separation for it. Based on the evaluation phase results, SVM with an RBF kernel was chosen as it has shown the highest accuracy and moreover it has shown better performance in previous studies in multiclass text classification [ref].

The above resulted data set has divided as the training and the testing data set, where 70% of the data set was used as the training data set to train the model and 30% of the initial data set was used to test the trained model.While training the model a pipeline has used to work with the vectorizer, transformer and the classifier. CountVectorizer was used to convert the extracted keywords into numeric values. TfidfTransformer was used to convert the values into a tf-idf representation which helps to determine how relevant each word is in the input phrase. Finally, the most suited career path was recommended, using the SVM as the classifier.

III. Conclusion

The present study on career guidance system focusing on IT industry was conducted using machine learning approach to provide a better and efficient guidance for IT undergraduate student and help them to get a better understanding on their skills.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.