A

Project Report on

**ML BASED CAREER GUIDANCE TOOL**

Presented by

**Ankush R. Sharma**

**Pranav D. Joshi**

B. Tech CSE – II

Computer Science and Engineering

2021 – 2022

Guided by

**Mr. Mohseen Ahmed**

Submitted to

A picture containing object, clock, light

Description automatically generated

**MGM’s College of Engineering, Nanded**

Under

**Dr. Babasaheb Ambedkar Technological University**

Lonere-402103, Tal-Mangaon, Dist.-Raigad

**CERTIFICATE**

This is to certify that the report entitled

**“ML Based Career Guidance Tool”**

Submitted by

**Ankush R. Sharma**

**Pranav D. Joshi**

In satisfactory manner as a partial fulfillment of

B. Tech CSE-II in Computer Science and Engineering

To



**MGM’s College of Engineering, Nanded**

Under

**Dr. Babasaheb Ambedkar Technological University,**

Lonere-402103, Tal-Mangaon, Dist.-Raigad

Has been carried out under my guidance,

**Mr. Mohseen Ahmed**

Guide

**Dr. Mrs. G. S. Lathkar**

**Director**

MGM’s College of Engineering, Nanded

**Dr. Mrs. A. M. Rajurkar**

**H.O.D.**

Computer Science and Engineering

**ACKNOWLEDGEMENT**

We are greatly indebted to our project guide **Mr. Mohseen Ahmed** for his able guidance throughout this work. It has been an altogether different experience to work with him and we would like to thank him for his help, suggestions, and numerous discussions.

We gladly take this opportunity to thank **Dr. Rajurkar A. M.** (Head of Computer Science & Engineering, MGM’s College of Engineering, Nanded).

We are heartily thankful to **Dr. Mrs. Lathkar G. S.** (Director, MGM’s College of Engineering, Nanded) for providing facility during progress of project, also for her kindly help, guidance, and inspiration.

Finally, we are also thankful to all those who help directly or indirectly to develop this project and complete it successfully.

With Deep Reverence,

Ankush R. Sharma (226)

Pranav D. Joshi (215)

B. Tech CSE- II

# ABSTRACT

Career choice has a pivotal role in college students’ life planning. In today’s world choosing the right career is the toughest decision. Today many students are confused about their future. They do possess some skills but they are not able to identify their abilities and a proper domain. Different people suggest different career options but at last, the student has to select their career. In this project, we have focused on this problem of the student using machine learning. With the help of machine learning, we will help the student to decide which is the best career option and domain for them using different machine learning techniques. The career is decided based on academic information filled by the student. This project will help the student to get directed towards a specific domain as per their skills.

Over the past few years several systems have been built to help students select the right career path by predicting the best career option based on their academic factors. However, academic factors are not the only relevant factor; we do need to consider one's cognitive abilities and psychometric factors too; such as, speed, learning capacity, endurance and memory to achieve the best career outcomes.

**Technologies used in this project: -**

* **Languages**: Python, HTML, CSS, PHP, JavaScript, Bootstrap. ·
* **Platform**:  VS-Code. ·
* **Algorithms**: K- Nearest Neighbour, Logistic regression. ·
* **Libraries**:
  + Pandas:  used for data manipulation and analysis.
  + NumPy:  used for working with arrays.
  + MatplotLib and Seaborn:  Python data visualization library.
  + Sklearn:  machine learning library.
* **Database**:  MySQL
* **Framework**:  Flask

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter no.** | **Title** | **Page no.** |
|  | **Acknowledgement** | **I** |
|  | **Abstract** | **II** |
|  | **Contents** | **III** |
|  | **List of figures** | **VII** |
|  | **List of tables** | **X** |
| **1.** | Introduction | **1** |
| **1.1** | Overview | **1** |
| **1.2** | Machine Learning | **2** |
| **1.3** | Supervised Learning | **3** |
| **1.4** | Classification | **3** |
| **1.5** | Frontend Development | **4** |
| **1.5.1** | HTML | **4** |
| **1.5.2** | CSS | **4** |
| **1.5.3** | JavaScript | **4** |
| **1.5.4** | PHP | **5** |
| **1.6** | Flask Framework | **6** |
| **1.7** | MySQL Database | **6** |
| **1.8** | Business problem | **7** |
| **1.9** | Motivation and Benefits | **7** |
| **2.** | Dataset | **8** |
| **2.1** | Collecting data | **8** |
| **2.2** | Statistics | **8** |
| **2.3** | Attributes | **9** |
| **2.4** | Prepare the data | **9** |
| **3.** | Exploratory data analysis | **10** |
| **3.1** | Overview | **10** |
| **3.2** | Tools and techniques | **10** |
| **3.3** | EDA on dataset | **11** |
| **3.4** | Data visualization | **15** |
| **4.** | Data preparation techniques | **16** |
| **4.1** | Feature engineering | **16** |
| **4.1.1** | What is feature engineering? | **16** |
| **4.1.2** | Importance of feature engineering | **16** |
| **4.1.3** | Feature engineering process | **17** |
| **4.1.4** | Feature engineering methods used in the project | **17** |
| **4.2** | Data pre-processing | **18** |
| **4.2.1** | What is data pre-processing | **18** |
| **4.2.2** | Data pre-processing steps | **18** |
| **4.2.3** | Data pre-processing methods used in the project | **20** |
| **5.** | Model building and implementation | **21** |
| **5.1** | Logistic regression | **21** |
| **5.1.1** | Types of logistic regression | **21** |
| **5.1.2** | Why logistic regression | **21** |
| **5.1.3** | Sigmoid function and hypothesis representation | **22** |
| **5.2** | K – nearest neighbour | **23** |
| **5.2.1** | Why do we need KNN algorithm? | **23** |
| **5.2.2** | How does KNN work? | **24** |
| **5.2.3** | How to select the value of K in KNN algorithm? | **26** |
| **5.3** | Advantages and disadvantages | **26** |
| **6.** | Frontend | **28** |
| **6.1** | HTML | **28** |
| **6.1.1** | Introduction | **28** |
| **6.1.2** | Basic tags | **29** |
| **6.1.3** | Attributes | **30** |
| **6.2** | CSS | **33** |
| **6.3** | JavaScript | **34** |
| **6.3.1** | Keywords | **35** |
| **6.3.2** | Data types | **35** |
| **6.3.3** | Operators | **36** |
| **6.4** | PHP | **36** |
| **6.4.1** | What is php? | **36** |
| **6.4.2** | What is php file? | **36** |
| **6.4.3** | What can php do? | **36** |
| **6.4.4** | Why php? | **37** |
| **7.** | System design and project code | **38** |
| **7.1** | System design | **38** |
| **7.2** | Website and project code | **40** |
|  | Conclusion | **52** |
|  | Reference | **53** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure no.** | **Title** | **Page no.** |
| **2.1** | Dataset | **8** |
| **3.1** | Mean of marks scored by students in different subjects | **12** |
| **3.2** | Different workshops attended by students | **12** |
| **3.3** | Count plot of management or technical column | **13** |
| **3.4** | Count plot of behaviour column | **13** |
| **3.5** | Count plot of introvert column | **13** |
| **3.6** | Count plot of teamwork column | **14** |
| **3.7** | Count plot of hackathon column | **14** |
| **3.8** | Count plot of new column Olympiads test | **14** |
| **3.9** | Count plot of new column self-learning | **15** |
| **5.1** | Sigmoid function | **22** |
| **5.2** | KNN example | **24** |
| **5.3** | New point classification (KNN) | **25** |
| **5.4** | Euclidean distance | **25** |
| **5.5** | New point classified | **26** |
| **6.1** | HTML basic structure | **29** |
| **6.2** | Img tag | **32** |
| **6.3** | Lists in HTML | **33** |
| **7.1** | System design flowchart | **39** |
| **7.2** | Login page | **40** |
| **7.3** | Registration page | **40** |
| **7.4** | Home page (1) | **41** |
| **7.5** | Home page (2) | **41** |
| **7.6** | Home page (3) | **42** |
| **7.7** | Home page (4) | **42** |
| **7.8** | Form page | **43** |
| **7.9** | Result page | **44** |
| **7.10** | About us page | **44** |
| **7.11** | Contact us page | **45** |
| **7.12** | Courses page (1) | **45** |
| **7.13** | Courses page (2) | **46** |
| **7.14** | Courses page (3) | **46** |
| **7.15** | Courses page (4) | **46** |
| **7.16** | Knowledge network page (1) | **47** |
| **7.17** | Knowledge network page (2) | **47** |
| **7.18** | Knowledge network page (3) | **47** |
| **7.19** | Knowledge network page (4) | **48** |
| **7.20** | Knowledge network page (5) | **48** |
| **7.21** | Knowledge network page (6) | **48** |
| **7.22** | Job role page (1) | **49** |
| **7.23** | Job role page (2) | **49** |
| **7.24** | Job role page (3) | **50** |
| **7.25** | Flask code (1) | **50** |
| **7.26** | Flask code (2) | **50** |
| **7.27** | Flask code (3) | **51** |
| **7.28** | Database | **51** |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table no.** | **Title** | **Page no.** |
| **3.1** | Descriptive statistics | **11** |
| **5.1** | Advantages and disadvantages | **27** |
| **6.1** | HTML basic tags | **29** |
| **6.2** | Core attributes | **30** |
| **6.3** | Table tag attributes | **30** |
| **6.4** | Text formatting | **31** |
| **6.5** | Form tag attributes | **32** |
| **6.6** | List of keywords in JavaScript | **35** |

*Chapter 1*

# introduction

* 1. **Overview**

Career guidance can be described as a process through which students become familiar with various career options, job opportunities and are prepared for those opportunities. Career counselling is the approach that will allow the student to understand his options, find his best skills and get acquainted with the world of work in order to make choices about employment, education and life.

Competition in today’s society is heavily multiplying day by day. It is too hard in the present day to face the technical world. So as to compete and reach the goal of students, they need to be planned and organized from the initial and final stages of their education. So it's important to perpetually assess their performance, establish their interests and assess how close they’re to their goal and assess whether or not they are within the right path that directs towards their target. 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 peak 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 relevant stream and career area to student. 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 candidates' performance in skills, talents and interests and place the candidate in the right job role suited for them. These kinds of prediction systems make their recruitment tasks very easy because as the inputs are given, recommendation is done based on inputs. Though the career counsellors may assist the students many times it would be difficult for them to completely understand the inclination of the students, academics and thus the counselling process may be limited. Also, not all students would be privileged to avail of such facilities. Globally there are some attempts in this area, but we need to work on this area for our students. Hence, we would be working on the web-based application, henceforth referred to as “Intelligent Career Planning & Guidance Assistant”

Intelligent Career Planning & Guidance Assistant is a computer program built with the help of experts where the details of the students and their aptitudes help finding the right course for their future. Choosing the right field after engineering is a very important life decision. Many Machine learning techniques have been applied to develop student performance prediction algorithms.

The proposed solution is a web-based application for engineering students early enough to:

* Understand their inclination
* Enhance understanding of their personality types
* Educate on the various options
* Enable them for their career planning, development, and guidance.
* Provide guidance on a continuous basis
* Make information available on career, education, etc. through sources
* Assist from choosing wrong options h. Be a partner in the overall journey
  1. **Machine Learning**

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

The term Machine Learning was coined by **Arthur Samuel** in **1959**, an American pioneer in the field of computer gaming and artificial intelligence, and stated that “it gives computers the ability to learn without being explicitly programmed”.  And in **1997, Tom Mitchell** gave a “well-posed” mathematical and relational definition that **“**A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.**”**

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods is **supervised learning** which trains algorithms based on example input and output data that is labelled by humans, and **unsupervised learning** which provides the algorithm with no labelled data in order to allow it to find structure within its input data

* 1. **Supervised Learning**

In supervised learning, the computer is provided with example inputs that are labelled with their desired outputs. The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples.

* 1. **Classification**

When inputs are divided into two or more classes and the learner must produce a model that assigns unseen for these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

In this project, we have used two classification algorithms:

1. Logistic Regression
2. K-Nearest Neighbours
   1. **Frontend Development**
      1. **HTML**

**HTML** stands for Hypertext Markup Language. It is used to design web pages using a markup language. HTML is the combination of Hypertext and Markup language. Hypertext defines the link between the web pages. A markup language is used to define the text document within tag which defines the structure of web pages. This language is used to annotate (make notes for the computer) text so that a machine can understand it and manipulate text accordingly. Most markup languages (e.g. HTML) are human-readable. The language uses tags to define what manipulation has to be done on the text.

HTML is a markup language used by the browser to manipulate text, images, and other content, in order to display it in the required format. HTML was created by Tim Berners-Lee in 1991. The first-ever version of HTML was HTML 1.0, but the first standard version was HTML 2.0, published in 1995.

* + 1. **CSS**

**C**ascading **S**tyle **S**heets, fondly referred to as CSS, is a simple design language intended to simplify the process of making web pages presentable.

CSS handles the look and feel part of a web page. Using CSS, you can control the colour of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colours are used, layout designs, and variations in display for different devices and screen sizes as well as a variety of other effects.

CSS is easy to learn and understand but it provides powerful control over the presentation of an HTML document. Most commonly, CSS is combined with the markup languages HTML or XHTML.

* + 1. **JavaScript**

JavaScript is a cross-platform, object-oriented scripting language used to make WebPages interactive (e.g., having complex animations, clickable buttons, popup menus, etc.). There are also more advanced server-side versions of JavaScript such as Node.js, which allow you to add more functionality to a website than downloading files (such as real-time collaboration between multiple computers). Inside a host environment (for example, a web browser), JavaScript can be connected to the objects of its environment to provide programmatic control over them. JavaScript contains a standard library of objects, such as Array, Date, and Math, and a core set of language elements such as operators, control structures, and statements. Core JavaScript can be extended for a variety of purposes by supplementing it with additional objects.

* + 1. **PHP**

PHP started out as a small open-source project that evolved as more and more people found out how useful it was. Rasmus Lerdorf unleashed the first version of PHP way back in 1994. PHP is a recursive acronym for "PHP: Hypertext Pre-processor". ·

* PHP is a server-side scripting language that is embedded in HTML. It is used to manage dynamic content, databases, session tracking, even build entire e-commerce sites. ·
* It is integrated with a number of popular databases, including MySQL, PostgreSQL, Oracle, Sybase, Informix, and Microsoft SQL Server. ·
* PHP is pleasingly zippy in its execution, especially when compiled as an Apache module on the UNIX side. The MySQL server, once started, executes even very complex queries with huge result sets in record-setting time. ·
* PHP supports a large number of major protocols such as POP3, IMAP, and LDAP. PHP4 added support for Java and distributed object architectures (COM and CORBA), making n-tier development a possibility for the first time. ·
* PHP is forgiving: PHP language tries to be as forgiving as possible. ·
* PHP Syntax is C-Like.

**Common Uses of PHP:**

* PHP performs system functions, i.e., from files on a system it can create, open, read, write, and close them. ·
* PHP can handle forms, i.e., gather data from files, save data to a file, through email you can send data, return data to the user. ·
* You add, delete, and modify elements within your database through PHP. ·
* Access cookies variables and set cookies. ·
* Using PHP, you can restrict users to access some pages of your website. ·
* It can encrypt data.
  1. **Flask Framework**

Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask’s framework is more explicit than Django’s framework and is also easier to learn because it has less base code to implement a simple web-Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on WSGI (Web Server Gateway Interface) toolkit and Jinja2 template engine.

**Advantages of Flask:**

* Higher flexibility.
* Higher compatibility with latest technologies.
* High scalability for simple web applications.
* Technical experimentation.
* Customization.
* Slightly higher framework performance.
* Easier to use for simple cases.
* Smaller size of the code base.
  1. **MySQL Database**

MySQL is a fast, easy-to-use RDBMS being used for many small and big businesses. MySQL is developed, marketed and supported by MySQL AB, which is a Swedish company. MySQL is becoming so popular because of many good reasons −

* MySQL is released under an open-source license. So you have nothing to pay to use it.
* MySQL is a very powerful program in its own right. It handles a large subset of the functionality of the most expensive and powerful database packages.
* MySQL uses a standard form of the well-known SQL data language.
* MySQL works on many operating systems and with many languages including PHP, PERL, C, C++, JAVA, etc.
* MySQL works very quickly and works well even with large data sets.
* MySQL is very friendly to PHP, the most appreciated language for web development.
* MySQL supports large databases, up to 50 million rows or more in a table. The default file size limit for a table is 4GB, but you can increase this (if your operating system can handle it) to a theoretical limit of 8 million terabytes (TB).
* MySQL is customizable. The open-source GPL license allows programmers to modify the MySQL software to fit their own specific environments.

* 1. **Business Problem**

In this Machine Learning end to end project, we are working on students’ data containing the proficiency of the students in various subjects. According to this data we have to predict the suitable job roles for them. The data contain the students’ behaviour and our job to find the insights from it.

This project will be helpful for the two entities: students and organizations. Where students can get recommended job roles according to their proficiency in each subject; and for the organizations who hire fresher’s for generic role. They can provide this tool to the students who got training in various domains. So that the organization will get to know in which domain a particular student should be put in.

* 1. **Motivation and benefits**

**Motivation**

* **Dataset**: Dataset has categorical variables which made dataset challenging and interesting.
* **Critical thinking**: In this project, I was able to think like a professional developing a solution from a situation.
* **In-depth exposure of Machine Learning:** In order to solve this problem, we had to refer to external sources instead of relying on academic knowledge; which provided exposure to different logics.

**Benefits**

* Improve myself in the field of data science and development.
* By predicting the suggested job roles for the students, it will help them in pursuing better career options.

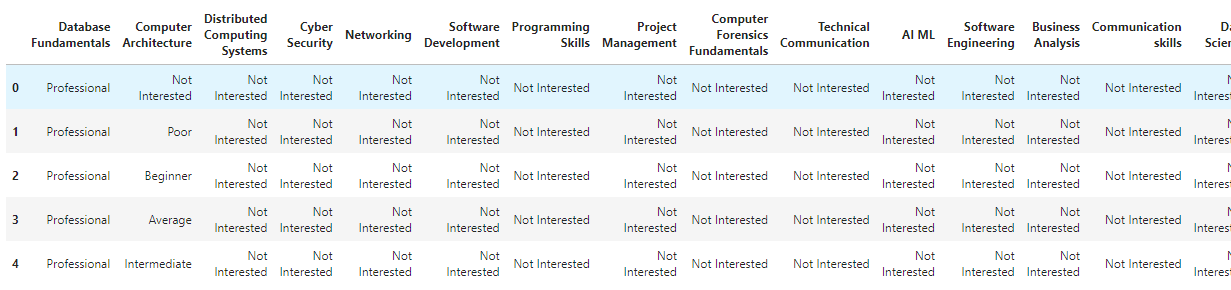
*Chapter 2*

# DATASET

**2.1 Collecting data**

Given the problem you want to solve, you will have to investigate and obtain data that you will use to feed your machine. The quality and quantity of information you get are very important since it will directly impact how well or badly your model will work. You may have the information in an existing database or you must create it from scratch. If it is a small project, you can create a spreadsheet that will later be easily exported as a CSV file.

* **Data source:** The source of the data is provided in from of the link <https://drive.google.com/file/d/1vFeWC0EUNj5TqEC8_SFDvMotbYDxMUZz/view?usp=sharing>
* **Data set:** Dataset has 9179 records and 18 columns/feature that describe nature of user, as shown in the fig. 2.1
  1. **Statistics**
* Number of columns: 18
* Number of entries: 9179
* Missing values: 1
* Missing values %: 0.011
* Duplicate rows: 1426
* Duplicate rows %: 15.535
* Variable types: (object – 18)



**Fig. 2.1** Dataset

* 1. **Attributes**
* **Database Fundamentals** – Proficiency of a student in Database Fundamentals
* **Computer Architecture** – Proficiency of a student in Computer Architecture
* **Distributed Computing Systems** – Proficiency of a student in Distributed Computing Systems
* **Cyber Security** – Proficiency of a student in Cyber Security
* **Networking** – Proficiency of a student in Networking
* **Software Development** – Proficiency of a student in Software development
* **Programming Skills** – Proficiency of a student in Programming Skills
* **Project Management** – Proficiency of a student in Project Management
* **Computer Forensics Fundamentals** – Proficiency of a student in Computer Forensics Fundamentals
* **Technical Communication** – Proficiency of a student in Technical Communication
* **AI ML** – Proficiency of a student in Artificial Intelligence and Machine Learning
* **Software Engineering** – Proficiency of a student in Software Engineering
* **Business Analysis** – Proficiency of a student in Business Analysis
* **Communication Skills** – Proficiency of a student in Communication Skills
* **Data Science** – Proficiency of a student in Data Science
* **Troubleshooting Skills** – Proficiency of a student in Troubleshooting Skills
* **Graphics Designing** – Proficiency of a student in Graphics Designing
* **Role** – Suggested roles.
  1. **Prepare the data**

This is a good time to [visualize your data](https://analyticsindiamag.com/how-to-get-started-with-visual-ai-the-new-automl-solution-by-datarobot/) and check if there are correlations between the different characteristics that we obtained. It will be necessary to make a selection of characteristics since the ones you choose will directly impact the execution times and the results. Additionally, you must balance the amount of data we have for each result -class- so that it is significant as the learning may be biased towards a type of response and when your model tries to generalize knowledge it will fail. You must also separate the data into two groups: one for training and the other for model evaluation which can be divided approximately in a ratio of 80/20 but it can vary depending on the case and the volume of data we have.

*Chapter 3*

**EXPLORATORY DATA ANALYSIS**

### 3.1 Overview

Exploratory Data Analysis (EDA) is the first step in your data analysis process. Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the answers you need.

Exploratory data analytics refers to the various ways to explore data. When you have a raw data set, it won’t provide any insight until you start to organize it. This can be some kind of readable format like an excel spreadsheet or, depending on your data, a complex visual model that visualizes data points. An in-depth exploration means structuring and visualizing data so you can identify patterns, outliers, anomalies, and other factors.

Developed by John Tukey in the 1970s, exploratory analysis is often described as a philosophy, and there are no hard-and-fast rules for how you approach it. That said, it also gave rise to a whole family of statistical-computing environments both used to help define, “What is EDA?” and to tackle specific tasks such as:

* Spotting mistakes and missing data;
* Mapping out the underlying structure of the data;
* Identifying the most important variables;
* Listing anomalies and outliers;
* Testing a hypotheses / checking assumptions related to a specific model;
* Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables);
* Estimating parameters and figuring out the associated confidence intervals or margins of error.

### 3.2 Tools and Techniques

Among the most important statistical programming packages used to conduct exploratory data analysis are S-Plus and R. The latter is a powerful, versatile, open-source programming language that can be integrated with many BI platforms.

Specific statistical functions and techniques you can perform with these tools include:

* Clustering and dimension reduction techniques, which help you to create graphical displays of high-dimensional data containing many variables;
* Univariate visualization of each field in the raw dataset, with summary statistics;
* Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you’re looking at;
* Multivariate visualizations, for mapping and understanding interactions between different fields in the data;
* K-Means Clustering (creating “centres” for each cluster, based on the nearest mean);
* Predictive models, e.g., linear regression.

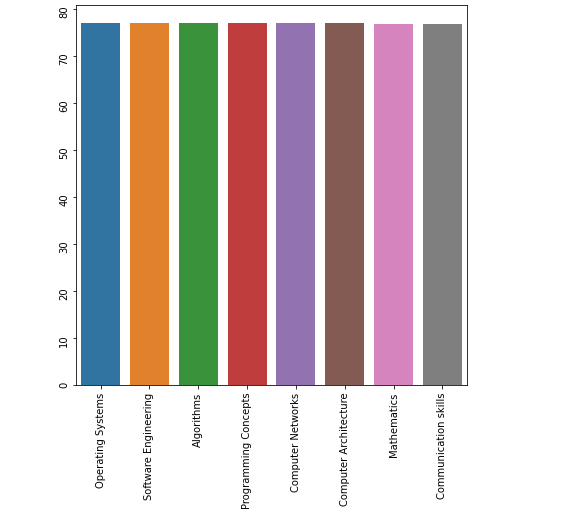
### 3.3 EDA on dataset

The table 3.1 shows the descriptive analysis on the dataset:

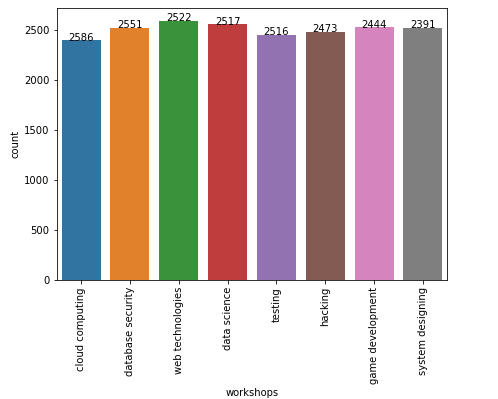
|  |  |
| --- | --- |
| NUMBER OF COLUMNS | 18 |
| NUMBER OF ROWS | **9179** |
| MISSING VALUES | **1** |
| dUPLICATE ROWS | **1476** |
| VARIABLE tYPES | **Int64 - 0**  **Object - 18** |

**Table no. 3.1** Descriptive Statistics

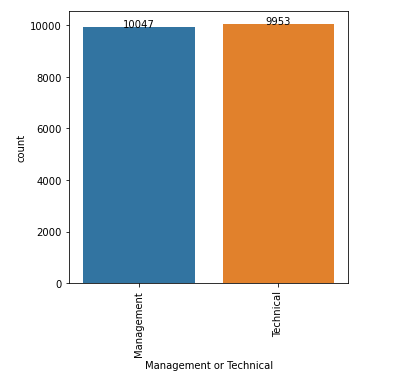
The graphs representing fig. 3.1 – fig. 3.9 explain the EDA performed on the datasets and insights derived from it.

****

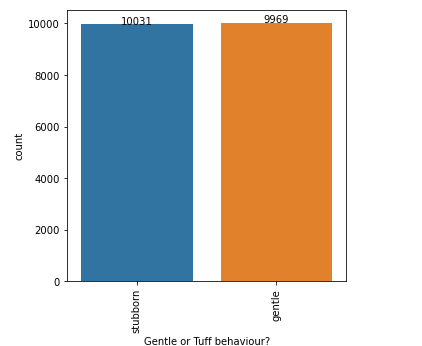
**Fig. 3.1** Mean of marks scored by students in different subjects

****

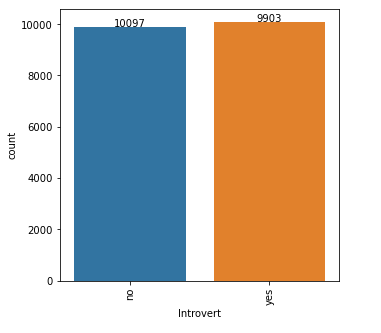
**Fig. 3.2** Different workshops attended by students

****

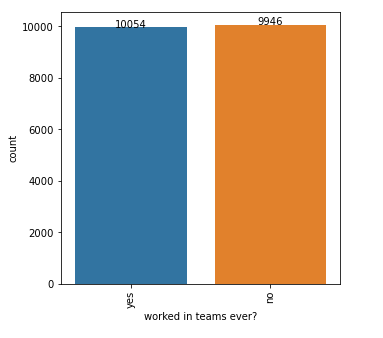
**Fig. 3.3** Count plot of ‘Management or Technical’ column

****

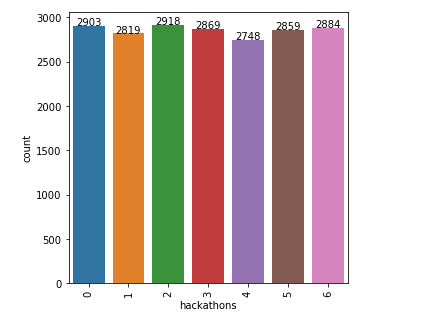
**Fig. 3.4** Count plot of Behaviour column

****

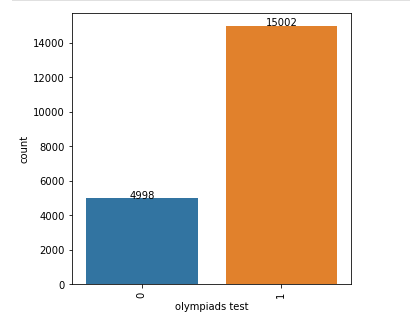
**Fig. 3.5** Count plot of introvert column

****

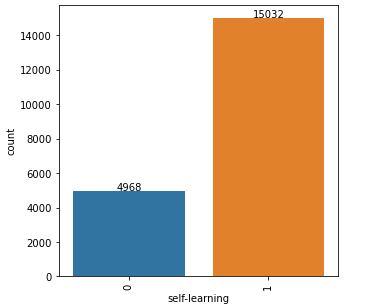
**Fig. 3.6** Count plot of teamwork column

****

**Fig. 3.7** Count plot of hackathon column

****

**Fig. 3.8** Count plot of new column Olympiads test

****

**Fig. 3.9** Count plot of new column self-learning

### 3.4 Important points

Important points noted after data visualizations are:

* Mean of marks as shown in fig. 3.1 scored by students in different subjects is almost same in this dataset.
* Count of different hackathons and workshop attend by student as shown fig. 3.2 and fig. 3.7 is almost same.
* Two columns test and Olympiads are combined. Its shows there are the greatest number of students who gave different external exams as shown in fig 3.8.
* Fig. 3.9 shows that a greater number of students endures self-learning.
* In fig. 3.6, teamwork column is equally divided in this dataset who worked on teams.

*Chapter 4*

**DATA PREPARATION TECHNIQUES**

**4.1 FEATURE ENINEERING**

**4.1.1 What is Feature Engineering?**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model — it could be the colour of an object or the sound of someone’s voice. Feature engineering, **in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.** It can produce new feature for both supervised and unsupervised learning, with the goal **of simplifying and speeding up data transformation while also enhancing model accuracy.** Feature engineering is required when working with the machine learning models. Regardless of the data or architecture, a terrible feature will have a direct impact on your model.

**4.1.2 Importance of Feature engineering**

Feature Engineering is a very important step in machine learning. Feature engineering refers to the process of designing artificial features into an algorithm. These artificial features are then used by that algorithm in order to improve its performance, or in other words reap better results. Data scientists spend most of their time with data, and it becomes important to make models accurate. When feature engineering activities are done correctly, the resulting dataset is optimal and contains all of the important factors that affect the business problem. As a result of these datasets, the most accurate predictive models and the most useful insights are produced. At the start of every machine learning project the raw data will be inevitably messy and unsuitable for training a model. The first step is always data exploration and cleaning, which involves changing data types and removing or imputing missing values. With an understanding of the data gained through exploration, it can be prepared in such a way that it is useful for the model.

**4.1.3 Feature engineering process**

* **Feature Creation**: Creating features involves creating new variables which will be most helpful for our model. This can be adding or removing some features. As we saw above, the cost per sq. ft column was a feature creation.
* **Transformations**: Feature transformation is simply a function that transforms features from one representation to another. The goal here is to plot and visualise data, if something is not adding up with the new features, we can reduce the number of features used, speed up training, or increase the accuracy of a certain model.
* **Feature Extraction**: Feature extraction is the process of extracting features from a data set to identify useful information. Without distorting the original relationships or significant information, this compresses the amount of data into manageable quantities for algorithms to process.
* **Benchmark:** A Benchmark Model is the most user-friendly, dependable, transparent, and interpretable model against which you can measure your own. It’s a good idea to run test datasets to see if your new machine learning model outperforms a recognised benchmark. These benchmarks are often used as measures for comparing the performance between different machine learning models like neural networks and support vector machines, linear and non-linear classifiers, or different approaches like bagging and boosting.

**4.1.4 Feature engineering methods**

In the development of our project, we have used the following feature engineering methods:

1. **One hot Encoding:**
   1. **One-hot encoding** is one of the techniques used to perform this conversion. One-hot encoding is essentially **the representation of categorical variables as binary vectors**. These categorical values are first mapped to integer values. Each integer value is then represented as a binary vector that is all 0s (except the index of the integer which is marked as 1.
2. **Label encoding:**
   1. **Label Encoding** refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

**4.2 DATA PREPROCESSING**

**4.2.1 What is data pre-processing?**

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. It can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects.

Machines like to process nice and tidy information – they read data as 1s and 0s. So, calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis. When using data sets to train machine learning models, you’ll often hear the phrase “garbage in, garbage out” This means that if you use bad or “dirty” data to train your model, you’ll end up with a bad, improperly trained model that won’t actually be relevant to your analysis.

Good, pre-processed data is even more important than the most powerful algorithms, to the point that machine learning models trained with bad data could actually be harmful to the analysis you’re trying to do – giving you “garbage” results. Depending on your data gathering techniques and sources, you may end up with data that’s out of range or includes an incorrect feature, like household income below zero or an image from a set of “zoo animals” that is actually a tree

**4.2.2 Data pre-processing steps**

1. **Data quality assessment**

Take a good look at your data and get an idea of its overall quality, relevance to your project, and consistency. There are a number of data anomalies and inherent problems to look out for in almost any data set

1. **Data cleaning**

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Dating cleaning is the most important step of pre-processing because it will ensure that your data is ready to go for your downstream needs.

* **Missing data**
* **Noisy data**
* **Binning:** Binning sorts data of a wide data set into smaller groups of more similar data. It’s often used when analysing demographics. Income, for example, could be grouped: $35,000-$50,000, $50,000-$75,000, etc.
* **Regression:** Regression is used to decide which variables will actually apply to your analysis. Regression analysis is used to smooth large amounts of data. This will help you get a handle on your data, so you’re not overburdened with unnecessary data.
* **Clustering:** Clustering algorithms are used to properly group data, so that it can be analysed with like data. They’re generally used in unsupervised learning, when not a lot is known about the relationships within your data.
* **Remove duplicate data**

1. **Data transformation**

With data cleaning, we’ve already begun to modify our data, but data transformation will begin the process of turning the data into the proper format(s) you’ll need for analysis and other downstream processes.

This generally happens in one or more of the below:

* Aggregation
* Normalization
* Feature selection
* Discreditation
* Concept hierarchy generation

1. **Data reduction**

Data reduction not only makes the analysis easier and more accurate, but cuts down on data storage.

It will also help identify the most important features to the process at hand.

* Attribute selection
* Numerosity reduction
* Dimensionality reduction

**4.2.3 Data pre-processing methods used**

* **Feature scaling**

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization.

**Standard scalar** standardizes features of the data set by scaling to unit variance. Standardization improves the convergence rate during the optimization process.

* **Splitting of dataset- Train and Test split**

The train-test-split function is for splitting a single dataset for two different purposes: training and testing. The training subset is for building your model.   
The testing subset is for using the model on unknown data to evaluate the performance of the model.

*Chapter 5*

**MODEL BUILDING AND IMPLEMENTATION**

After preparing the data for training it is an important step to select effective algorithms to train the model. In this project, we have used two classification algorithms i.e., **Logistic Regression and K – nearest neighbour.**

**5.1 Logistic Regression**

**Regression** is a method to determine a statistical relationship between dependent variable and one or more independent variables. **Logistic regression** is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

**5.1.1 Types of logistic regression**

There are two types of logistic regression:

* **Binary** (e.g., Tumour Malignant or Benign)
* **Multivariate** (e.g., Cats, dogs or Sheep's, etc.)

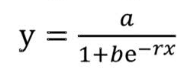
**5.1.2 Why logistic regression?**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems; it is a predictive analysis algorithm and based on the concept of probability. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function.

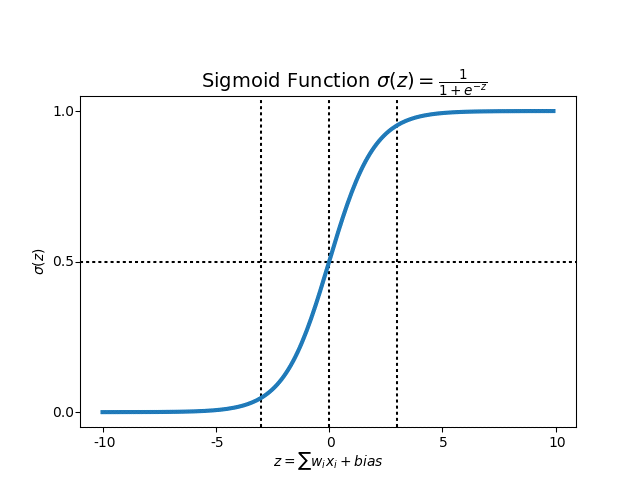
The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

Basic formula of Logistic regression model:



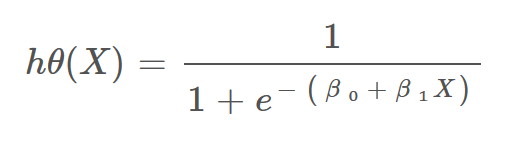
**5.1.3 Sigmoid function and Hypothesis representation**

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities, as shown in fig. 5.1:



**Fig. 5.1** Sigmoid function

Hypothesis representation:



**5.2 K – NEAREST**

K-Nearest Neighbors is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data.

**Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So, for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs’ images and based on the most similar features it will put it in either cat or dog category.

**5.2.1 Why do we need a K-NN algorithm?**

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram or fig. 5.2



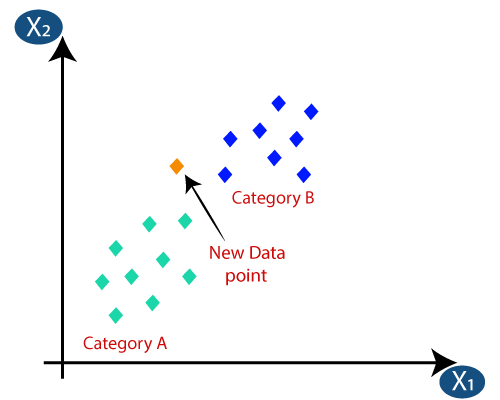
**Fig. 5.2** K-NN example 1

**5.2.2 How does K-NN work?**

The K-NN working can be explained on the basis of the below algorithm:

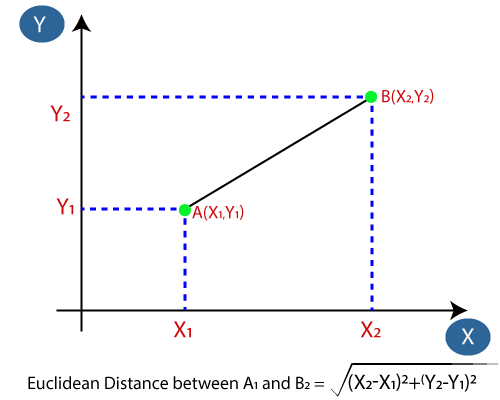
* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below figure 5.3: Select the best attribute using Attribute Selection Measures (ASM) to split the records.



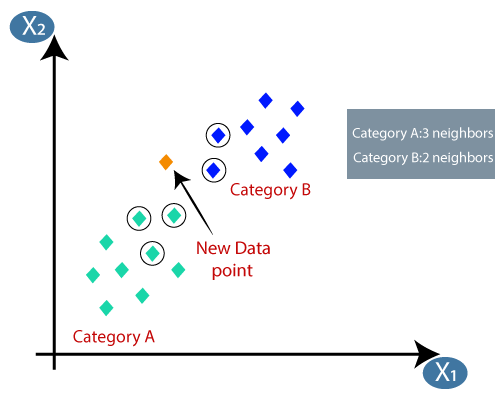
**Fig. 5.3** new point classification (KNN)

* Firstly, we will choose the number of neighbors, so we will choose the k=5.
* Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as shown in Fig.5.4:



**Fig. 5.4** Euclidean distance

* By calculating the Euclidean distance, we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider Fig. 5.5:



**Fig. 5.5** New point classified (KNN)

* As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

**5.2.3 How to select the value of K in K-NN algorithm?**

Below are some points to remember while selecting the value of K in the K-NN algorithm:

* There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
* A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
* Large values for K are good, but it may find some difficulties.

**5.3 Advantages and Disadvantages**

The table 5.1 describes the advantages and disadvantages of logistic regression and K –Nearest Neighbour

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | **K – Nearest Neighbour** |
| **Advantages** | 1. Logistic Regression performs well when the dataset is linearly separable. 2. Logistic regression is less prone to over-fitting but it can over fit in high dimensional datasets. 3. Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative). 4. Logistic regression is easier to implement, interpret and very efficient to train | 1. It is simple to implement. 2. It is very useful for nonlinear data because there is no assumption about data in this algorithm. 3. It is a versatile algorithm as we can use it for classification as well as regression. 4. It has relatively high accuracy but there are much better supervised learning models than KNN. 5. It is robust to the noisy training data. 6. It can be more effective if the training data is large. |
| **Disadvantages** | 1. Main limitation of Logistic Regression is the **assumption of linearity** between the dependent variable and the independent variables 2. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to over fit. 3. Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. | 1. Always needs to determine the value of K which may be complex some time. 2. The computation cost is high because of calculating the distance between the data points for all the training samples. |

**Table no. 5.1** Advantages and disadvantages

*Chapter 6*

**FRONTEND**

Hypertext Markup Language (HTML) is one of the three main components of modern web pages, along with Cascading Style Sheets (CSS) and JavaScript. HTML indicates to the browser what elements should be included in the webpage (and in what order). CSS indicates how each element should be styled. JavaScript provides a means for webpage authors to manipulate these elements programmatically and in response to actions by the end user.

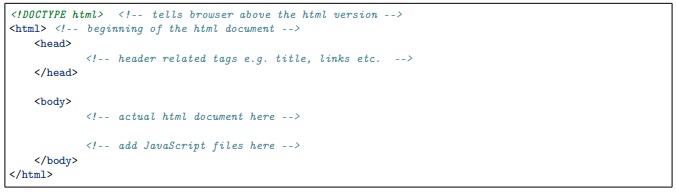
**6.1 HTML**

**HTML** stands for Hypertext Markup Language. It is used to design web pages using a markup language. HTML is the combination of Hypertext and Markup language. Hypertext defines the link between the web pages. A markup language is used to define the text document within tag which defines the structure of web pages. This language is used to annotate (make notes for the computer) text so that a machine can understand it and manipulate text accordingly. Most markup languages (e.g., HTML) are human-readable. The language uses tags to define what manipulation has to be done on the text.

**6.1.1. Introduction**

The basic structure for an HTML page is shown below.

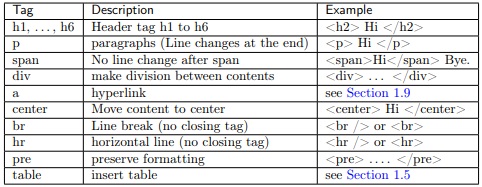
* Entries inside the /< . . . /> are known as tags. Most of the tags has an opening and closing e.g., <head> (opening head) and </head> (closing head). Some of the tags do not have closing tags e.g., <! DOCTYPE> and <br/>. We need to write the HTML codes inside the tags.
* The comments are written between ‘!-’ and ‘->’.
* Here Line 1 gives the details of the ‘HTML version’ to the web-browser. The ‘html’ tells it is version 5.
* The ‘head’ tag (Lines 3-5) contains the header related tags e.g., ‘title for the page’ and ‘links for the CSS files’ etc.
* The ‘body’ tag (7-11) contains the actual HTML code which is displayed on the web-browser. Also, we add all the JavaScript related codes just before the closing body tag (</body>).
* The figure 6.1 describes the basic structure of HTML:



**Fig. 6.1** HTML basic structure

**6.1.2 Basic tags**

The Table 6.1 shows the list of tags which are required for writing the basic ‘HTML’ codes i.e., without any style e.g., bold, italics and numbering etc.



**Table no 6.1** HTML basic tags

**6.1.3 Attributes**

Various attributes in HTML are: -

1. **Core attributes**

Table 6.2 shows core attributes used in HTML.

|  |  |  |
| --- | --- | --- |
| **Name** | **Values** | **Description** |
| Id | user defined names | <p id=’p\_1’> Hi </p> |
| class | user defined names | <p class=p\_class’> Hi</p> |
| style | CSS styles | <p style=’color: red; font-weight: bold;”> Hi</p> |
| align | left, right, centre | horizontal alignment |
| Width | numeric value or % value | width of images and tables etc. |
| Height | numeric value | height of images and tables etc |

**Table no 6.2** core attributes

1. **Tables**

Table 6.3 shows table tag and its attributes.

|  |  |
| --- | --- |
| **Tag** | **Description** |
| Table | beginning and end of table |
| Tr | row of table |
| th | header cell |
| td | data cell |
| **Attributes** | **Description** |
| rowspan | number of rows to merge |
| Colspan | number of columns to merge |
| Border | width of border |
| Cell padding | width of whitespace between two border |
| Cellspacing | width of whitespace within a border |
| Bgcolor | background color |
| Bordercolor | color of border |
| Width | width of table (numeric or %) |
| Height | height of table (numeric) |
| Caption | caption for table |

**Table no 6.3** Table tag attributes

1. **Text formatting**

Table 6.4 shows text formatting tags used in HTML.

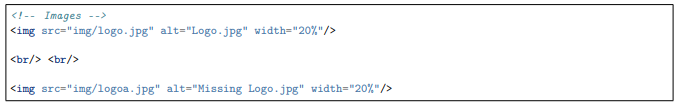
|  |  |
| --- | --- |
| **Tag** | **Description** |
| B | Bold |
| I | Italic |
| U, ins | Underline |
| Strike, del | Strike |
| Sup | Superscript |
| Sub | Subscript |
| Big | Big size text |
| Small | Small size text |

**Table no 6.4** Text formatting

1. **Images**

Image tag has two important attributes i.e., ‘src’ and ‘alt’ as described below,

* src: tells the location of ‘image’ file the image ‘logo.jpg’ will be searched inside the folder ‘img’.
* alt: is the ‘alternate text’ which is displayed if image is not found. For example, in Line 6, the name of the image is incorrectly written i.e. ‘logo a’ (instead of ‘logo’), therefore the value of ‘alt’ i.e. ‘Missing Logo.jpg’ will be displayed as shown in Fig. 6.2.



**Fig. 6.2** img tag

1. **Forms**

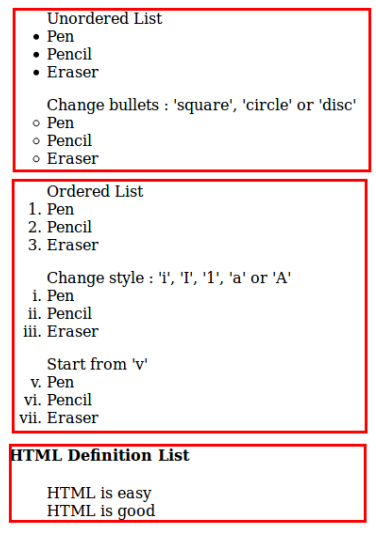
The below table 6.5 describes the form tag and its attributes:



**Table no 6.5** form tag attributes

1. **Lists**

Fig. 6.3 shows two types of lists used in HTML. i.e., unordered list and ordered list.

****

**Fig. 6.3** Lists in html

**6.2 CSS**

**C**ascading **S**tyle **S**heets, fondly referred to as CSS, is a simple design language intended to simplify the process of making web pages presentable.

CSS handles the look and feel part of a web page. Using CSS, you can control the colour of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colours are used, layout designs, and variations in display for different devices and screen sizes as well as a variety of other effects.

Types of CSS:

* 1. **Inline CSS**:  Inline CSS contains the CSS property in the body section attached with element is known as inline CSS. This kind of style is specified within an HTML tag using the style attribute.
  2. **Embedded CSS**: This can be used when a single HTML document must be styled uniquely. The CSS rule set should be within the HTML file in the head section i.e. the CSS is embedded within the HTML file.
  3. **External CSS**: External CSS contains separate CSS file which contains only style property with the help of tag attributes (For example class, id, heading, … etc). CSS property written in a separate file with .css extension and should be linked to the HTML document using **link** tag. This means that for each element, style can be set only once and that will be applied across web pages.

CSS is easy to learn and understand but it provides powerful control over the presentation of an HTML document. Most commonly, CSS is combined with the markup languages HTML or XHTML.

**Properties of CSS:**

Inline CSS has the highest priority, then comes Internal/Embedded followed by External CSS which has the least priority. Multiple style sheets can be defined on one page. If for an HTML tag, styles are defined in multiple style sheets then the below order will be followed.

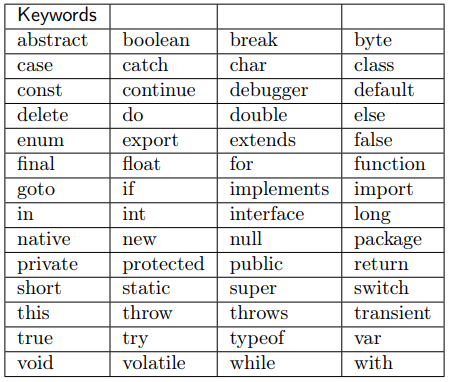
* As Inline has the highest priority, any styles that are defined in the internal and external style sheets are overridden by Inline styles.
* Internal or Embedded stands second in the priority list and overrides the styles in the external style sheet.
* External style sheets have the least priority. If there are no styles defined either in inline or internal style sheet then external style sheet rules are applied for the HTML tags.

**6.3 JavaScript**

JavaScript is a cross-platform, object-oriented scripting language used to make Webpages interactive (e.g., having complex animations, clickable buttons, popup menus, etc.). There are also more advanced server-side versions of JavaScript such as Node.js, which allow you to add more functionality to a website than downloading files (such as real-time collaboration between multiple computers). Inside a host environment (for example, a web browser), JavaScript can be connected to the objects of its environment to provide programmatic control over them. JavaScript contains a standard library of objects, such as Array, Date, and Math, and a core set of language elements such as operators, control structures, and statements. Core JavaScript can be extended for a variety of purposes by supplementing it with additional objects

**6.3.1 Keywords**

Below table 6.6 includes the reserved keywords in the JavaScript which cannot be used as ‘variable’ and ‘function’ names etc.



**Table no 6.6** list of keywords in JS

**6.3.2 Datatypes**

JavaScript has three types of data,

* Numbers: 123, 32.32
* Strings: “Pranav”, “Krishna Patel”, “123”
* Boolean: true, false
  + 1. **Operators**

**1. Assignment operators:** Different assignment operators used in JavaScript are ‘=’, ‘-=’, ‘\*=’, ‘/=’ and ‘%=’.

**2. Comparison operators:** Comparison operators used are ‘==’, ‘===’, ‘! =’, ‘>’, ‘<’, ‘>=’, and ‘<=’.

**3. Logical operators:** Logical operators used are ‘&&’ (logical and), ‘||’ (logical or) and ‘!’ (logical not).

**2. Bitwise operator:** Bitwise operators used are ‘&’ ( and), ‘|’ (or), ‘^’ (xor) and ‘~’ (not).

**6.4 PHP**

**6.4.1 What is PHP?**

PHP started out as a small open-source project that evolved as more and more people found out how useful it was. Rasmus Lerdorf unleashed the first version of PHP way back in 1994. PHP is a recursive acronym for "PHP: Hypertext Pre-processor".  It is a widely-used, open-source scripting language. Its scripts are executed on the server. It is free to download and use

**6.4.2 What is a PHP file?**

PHP files can contain text, HTML, CSS, JavaScript, and PHP code. Its code is executed on the server, and the result is returned to the browser as plain HTML. Files of PHP have extension ".php".

**6.4.3 What can PHP do?**

* PHP can generate dynamic page content
* PHP can create, open, read, write, delete, and close files on the server
* PHP can collect form data
* PHP can send and receive cookies
* PHP can add, delete, modify data in your database
* PHP can be used to control user-access
* PHP can encrypt data.

With PHP you are not limited to output HTML. You can output images, PDF files, and even flash movies. You can also output any text, such as XHTML and XML.

**6.4.4 Why PHP?**

* PHP runs on various platforms (Windows, Linux, UNIX, Mac OS X, etc.)
* PHP is compatible with almost all servers used today (Apache, IIS, etc.)
* PHP supports a wide range of databases
* PHP is free. Download it from the official PHP resource: [www.php.net](http://www.php.net/)
* PHP is easy to learn and runs efficiently on the server side.

*Chapter 7*

**SYSTEM DESIGN AND PROJECT CODE**

**7.1 System Design**

METHODOLOGY

**Registration and Login**

The students would be registered through a very simple method either by email id or Mobile number. The login credentials would be created and would be validated through every login attempt. Students Can See Various Fields.

**Discover yourself**

This section would enable the students to take a few tests to discover themselves in terms of their ability, interests, inclination, future plans etc. This would create a Student profile which would be used as a baseline for suggesting the possible career options. We will leverage the AI ML techniques to predict the way forward.

**Data Pre-processing**

We pre-process the data into required format. For Example, the data in data set will be stored in the form of words, nothing but alphabetic. We convert those into numerical format.

**Predicting the Skills**

From the Data By applying various machine algorithms on the data set ,we found more accuracy. At any one algorithm, thus it suits for the recommendation system to be accurate.

**Then Recommend the respected skill**

Individual students differ from the other students in their skills. Recommendation system helps to predict the inherent skill of a student and recommend the respected skill courses

**Knowledge Networking**

As the name indicates, this module would assist to harness the knowledge through various sources. This would also have a section to provide the information by students, which would be made available only post scrutiny by the admin team.

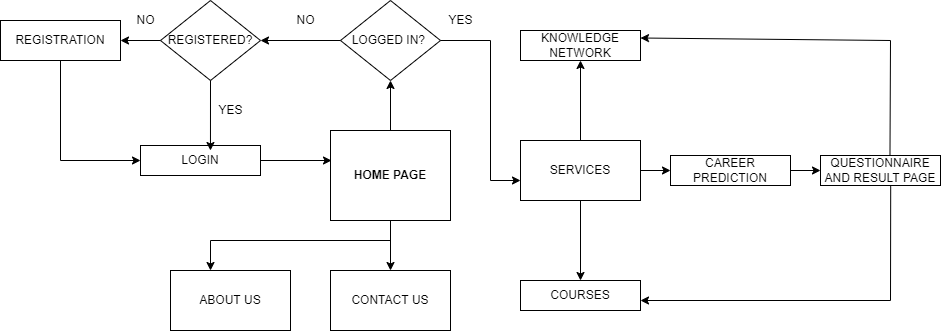
**Online courses**

This section as the names suggests consist of various courses for you to learn. This page has different section for programming languages, server-side languages, frontend and back end. And all these are links to various tutorials, courses (free and paid) present on the internet.

**Blogs**

This section has various job roles listed in it.

The following figure 7.1 shows the system design flow of our project:



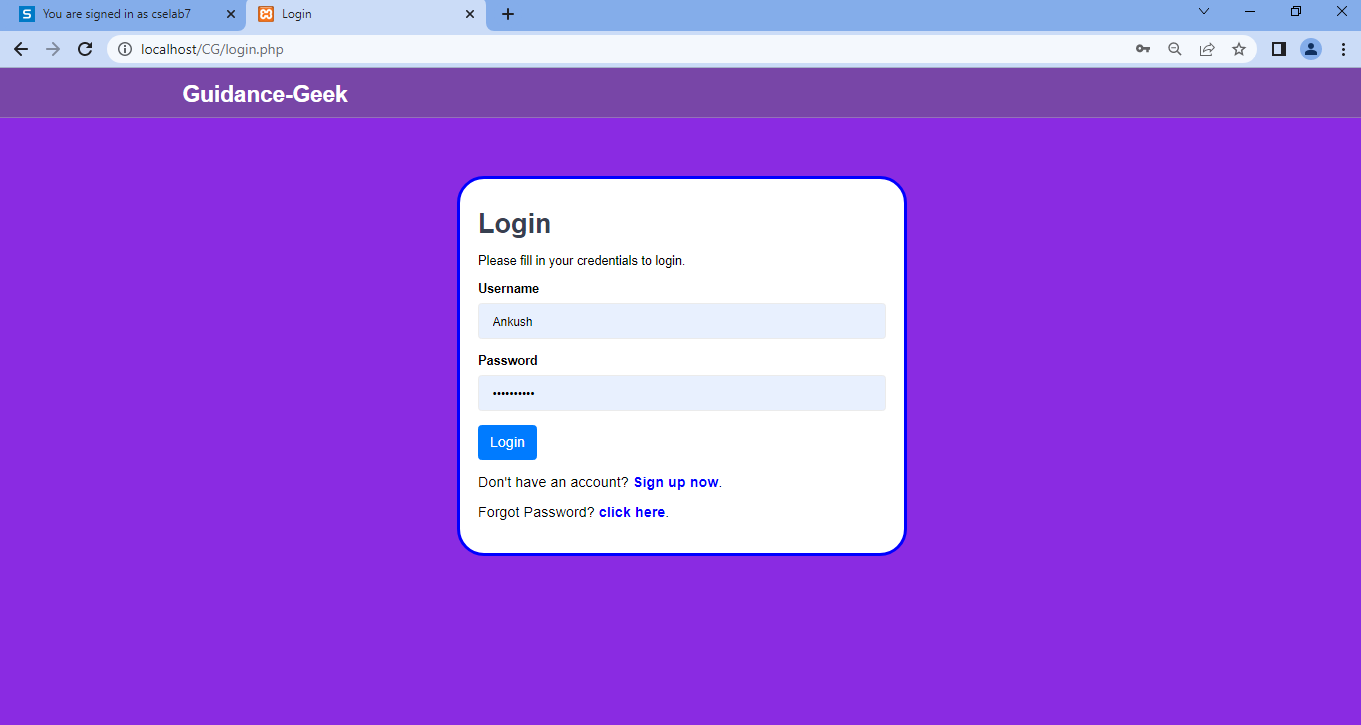
**Fig. 7.1** System design flowchart

**7.2 Website and project code**

* **Login page**

The login page consists of two fields i.e., username and password as shown in figure 7.2.

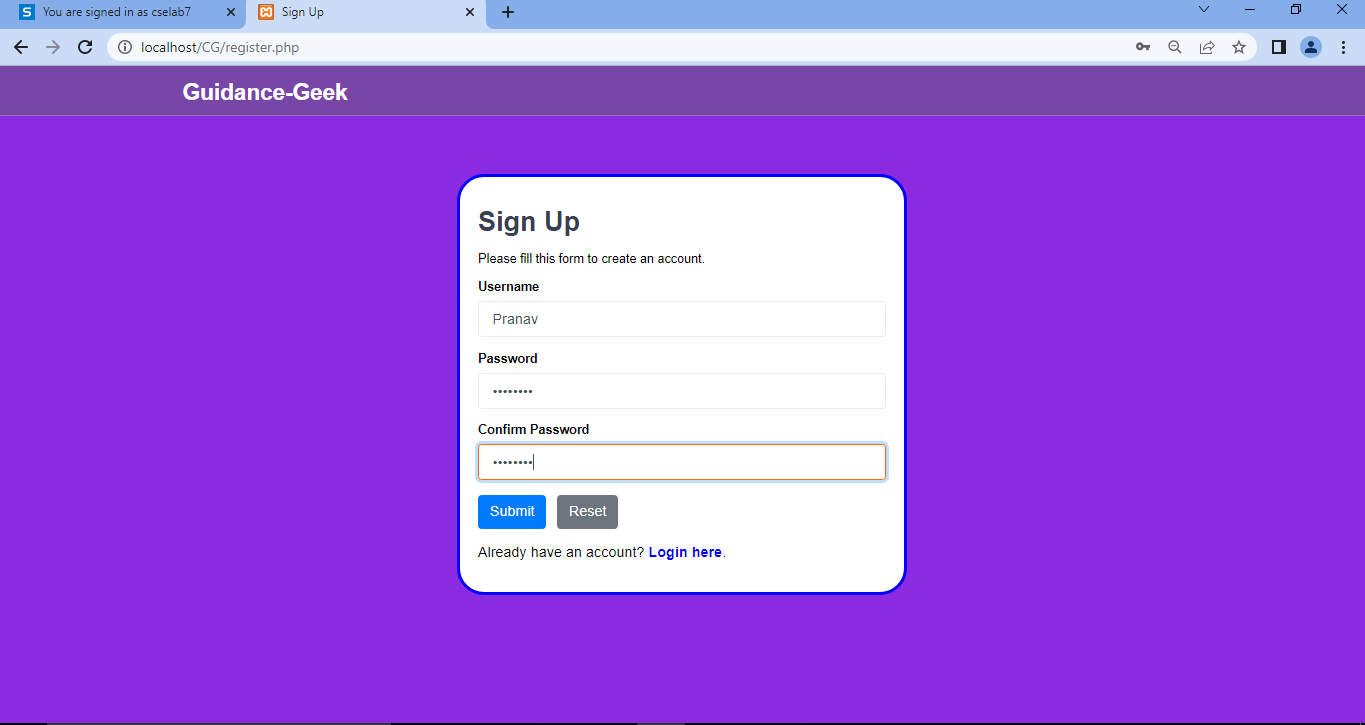
If the user is already registered on the website, he/she can directly login.

****

**Fig. 7.2** Login Page

* **Registration page**

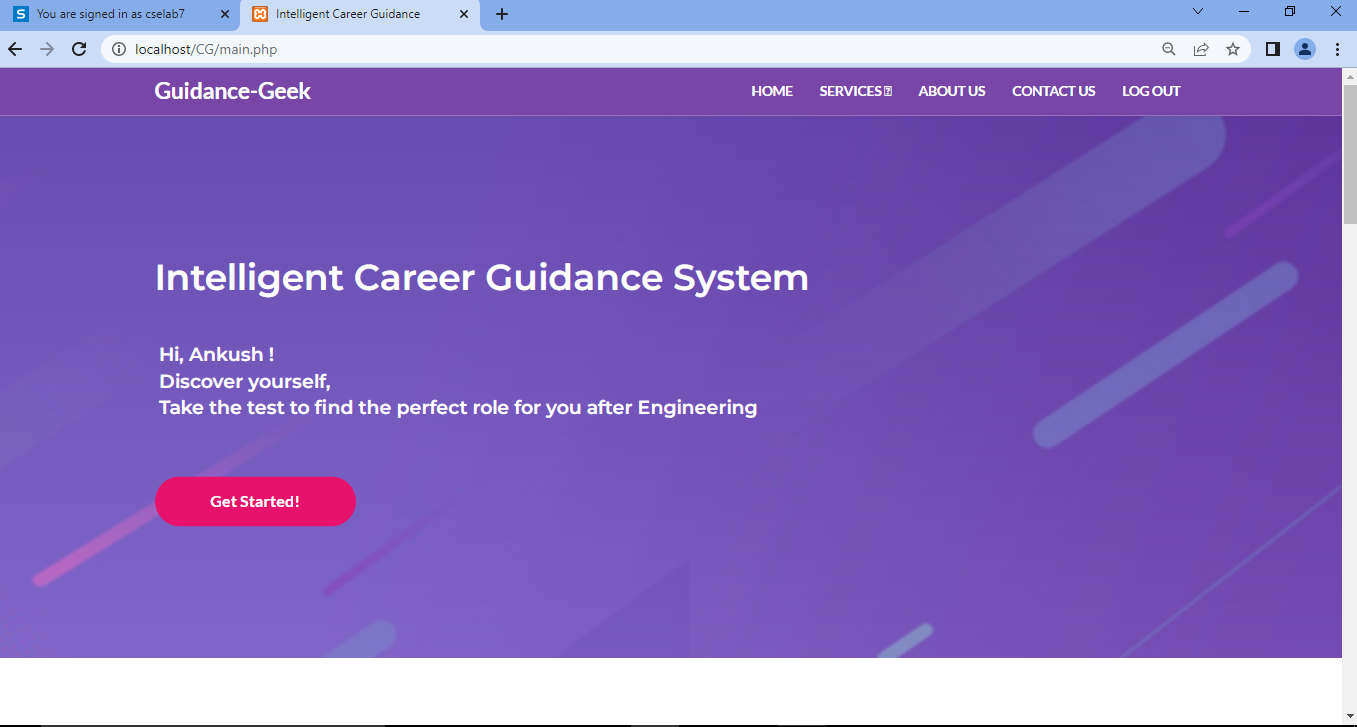
When a new user wants to register on the website, the user has to provide a unique username along with the password as shown in figure 7.3:

****

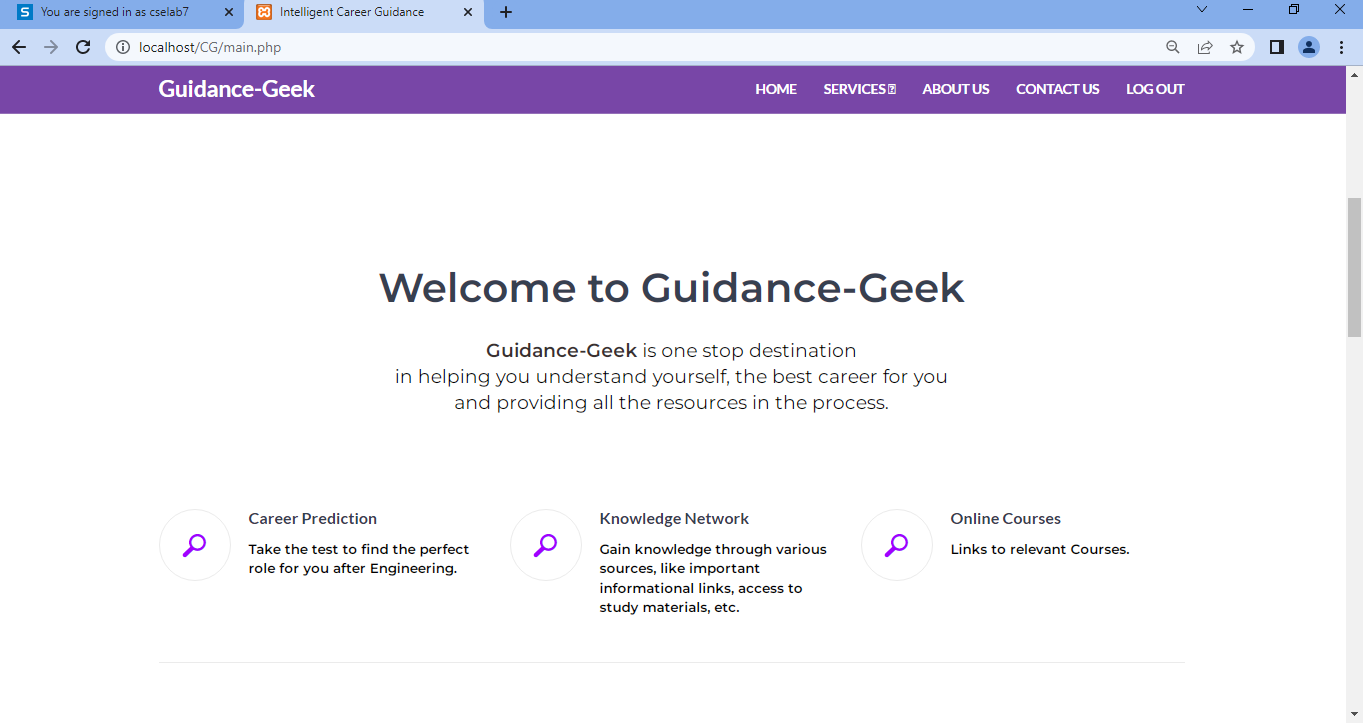
**Fig. 7.3** Registration Page

* **Home page**

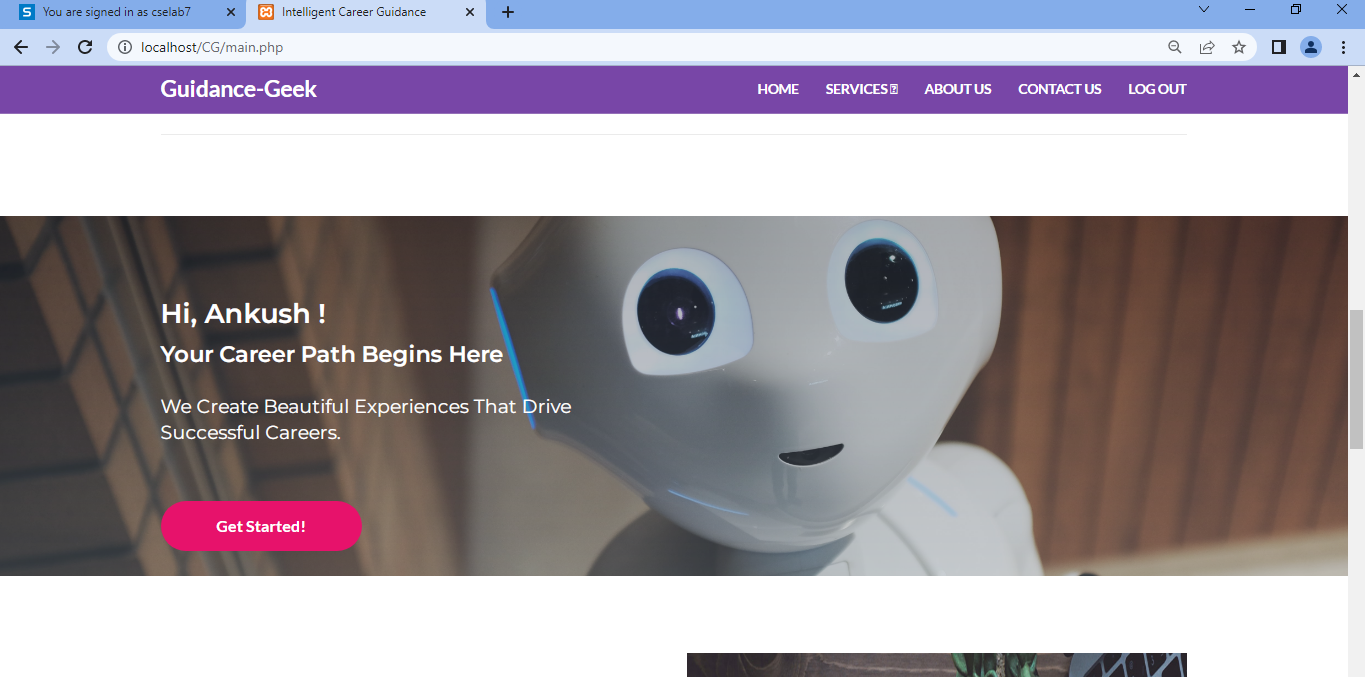
Home page is the interface which a user will see when the user opens the website. From the homepage user can access various services like career prediction, knowledge network blog, courses etc. Along with about us and contact us page. Figures 7.4, 7.5, 7.6 and 7.7 shows the interface of home page:

****

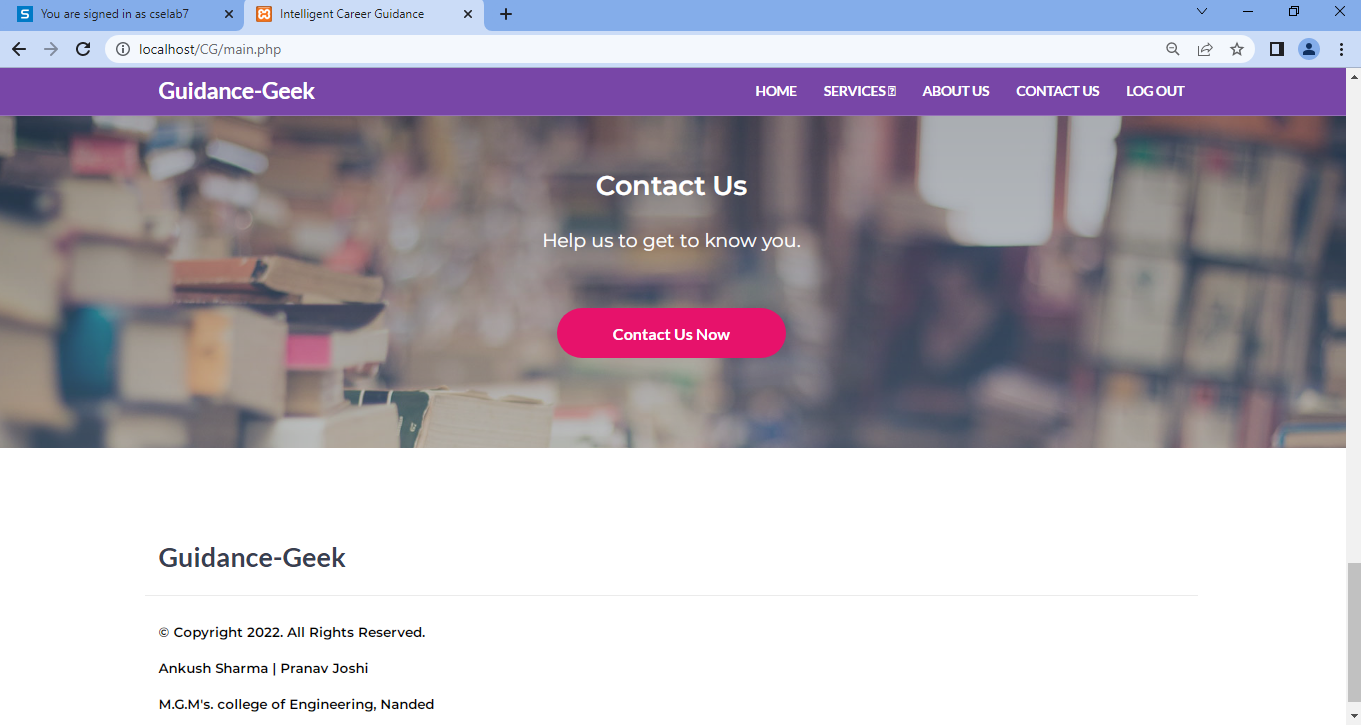
**Fig. 7.4** Home page (1)

****

**Fig. 7.5** Home page (2)

****

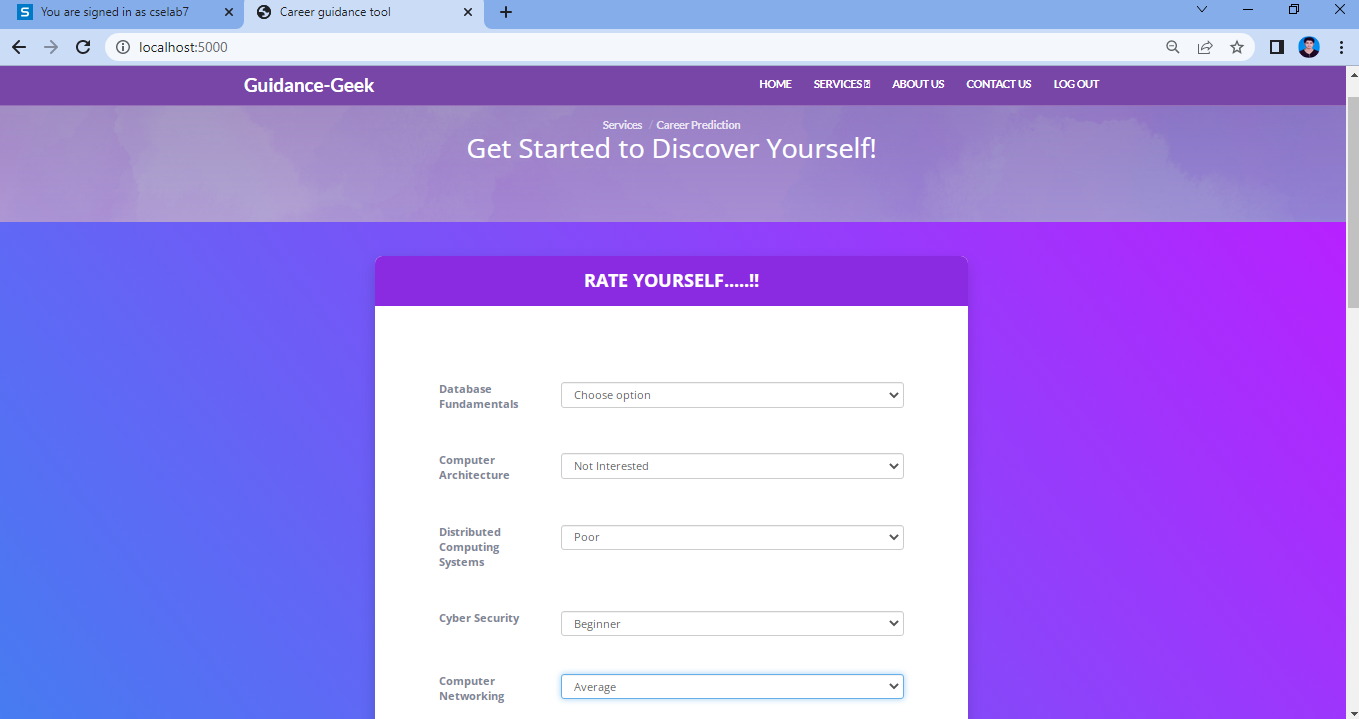
**Fig. 7.6** Home page (3)

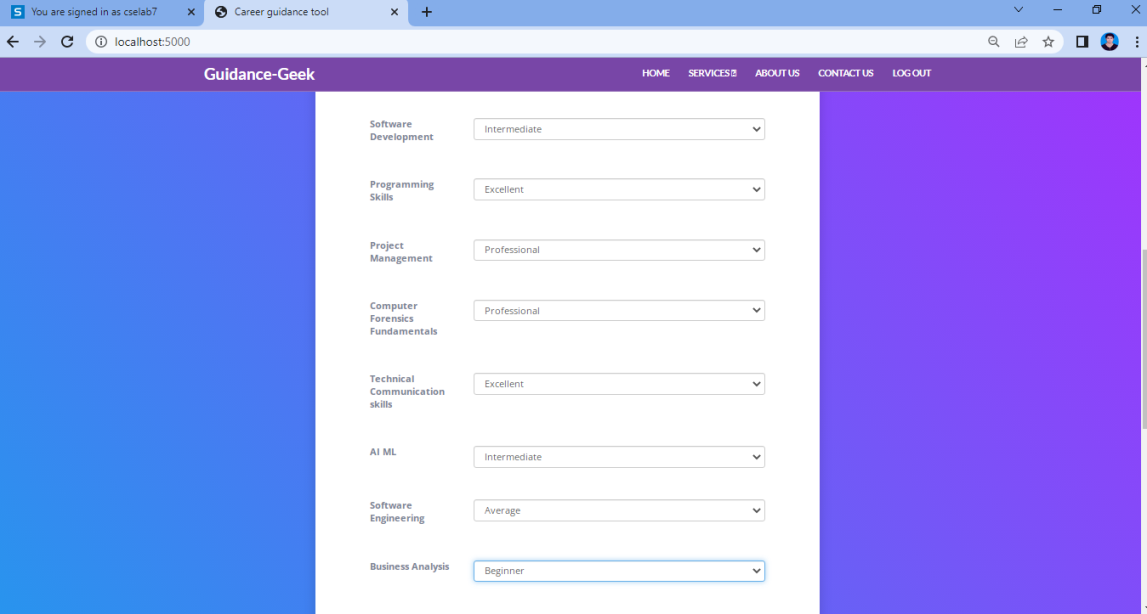
****

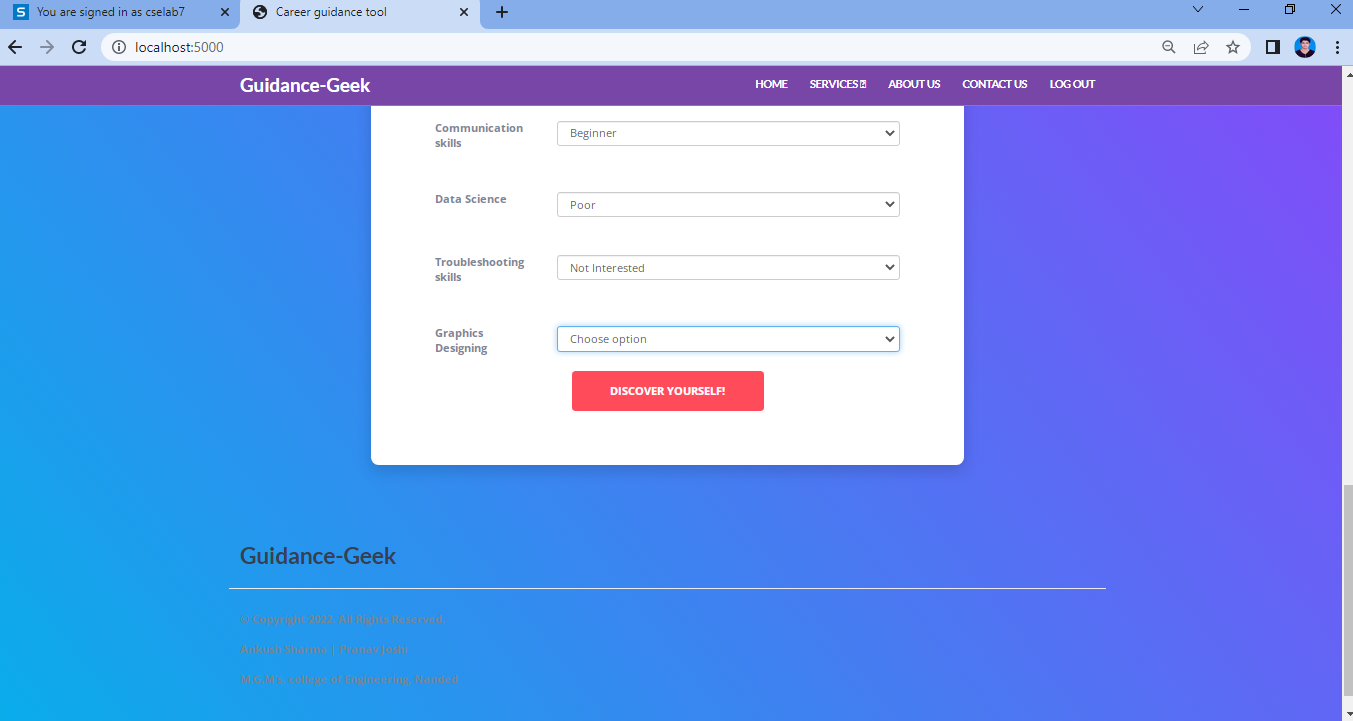
**Fig. 7.7** Home page (4)

* **Form (of different efficiency: not interested - professional)**

The form page is used to predict the job role based upon the proficiency of a candidate in various subjects/domains. Refer the fig. 7.8 for the form page:

****

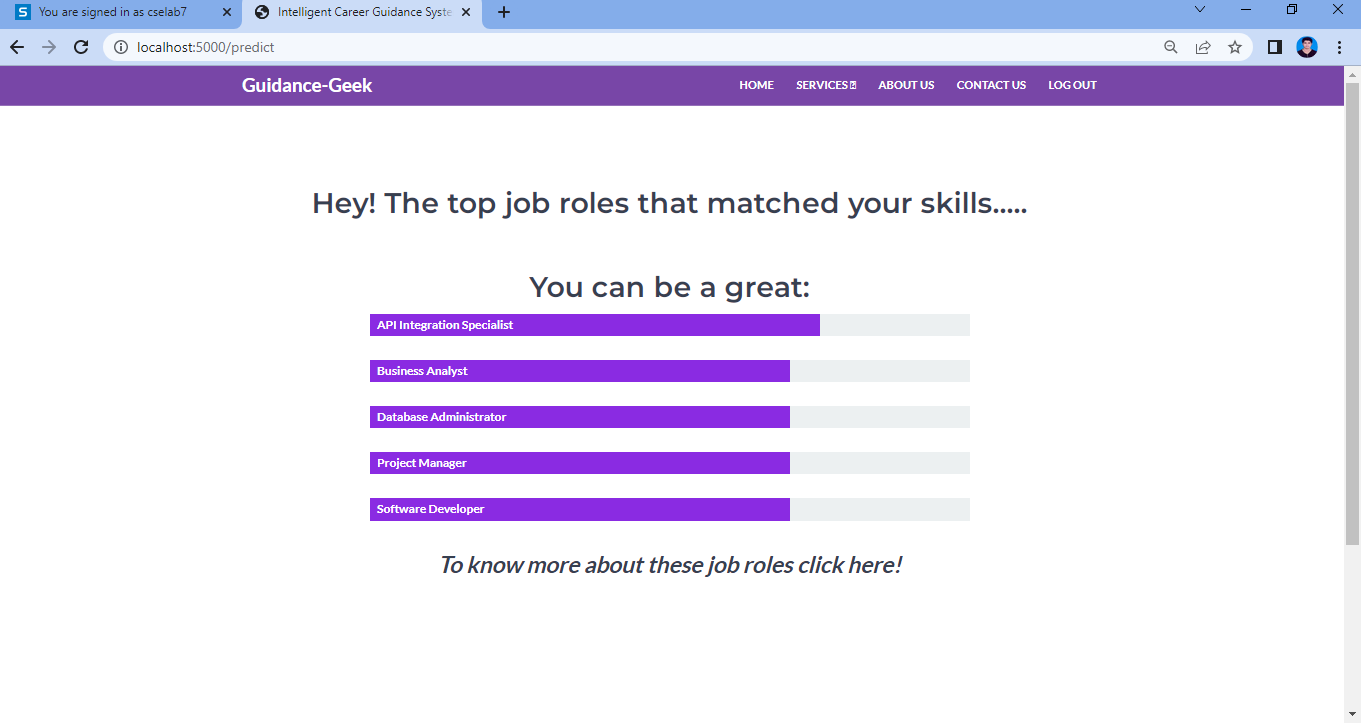
****

****

**Fig. 7.8** Form page

* **Result page**

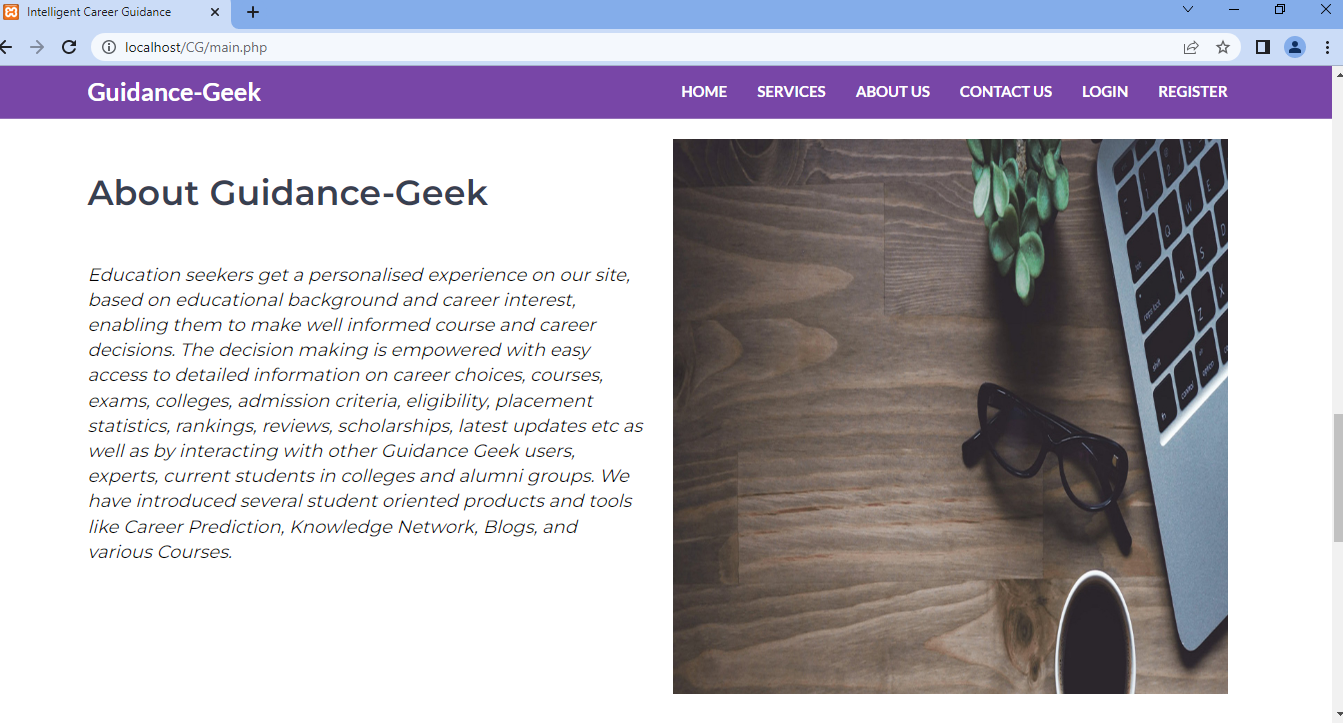
When the user submits the career prediction form, he/she will get the suggested job roles which are predicted using the machine learning model. Refer the below fig. 7.9 for the result page:

****

**Fig. 7.9** Result page

* **About us**

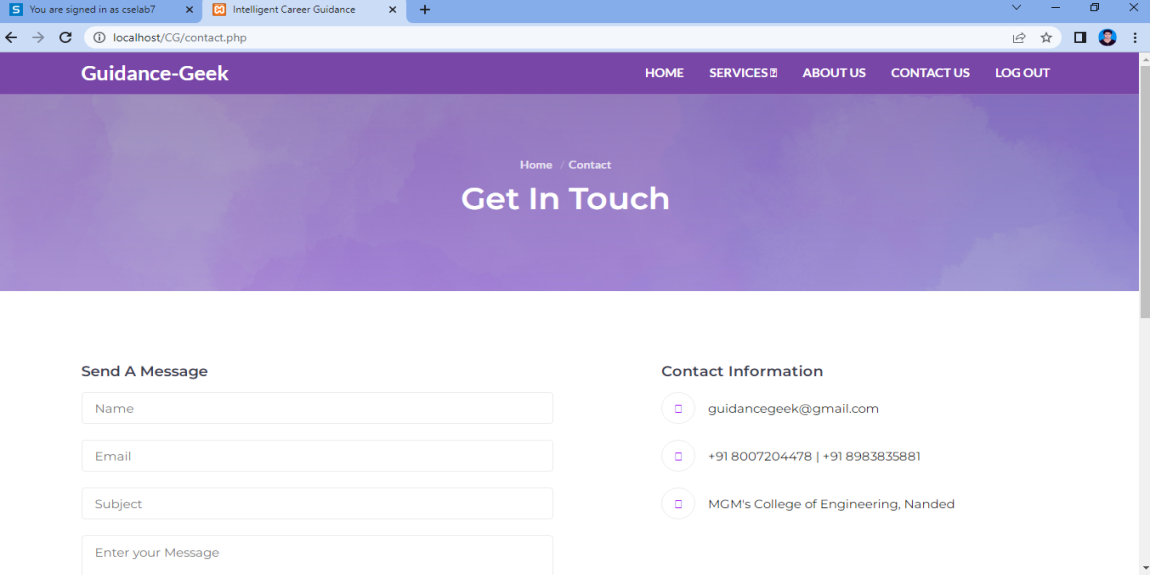
The about us page describes main functionalities of the website as shown in the figure 7.10:

****

**Fig. 7.10** About us page

* **Contact us**

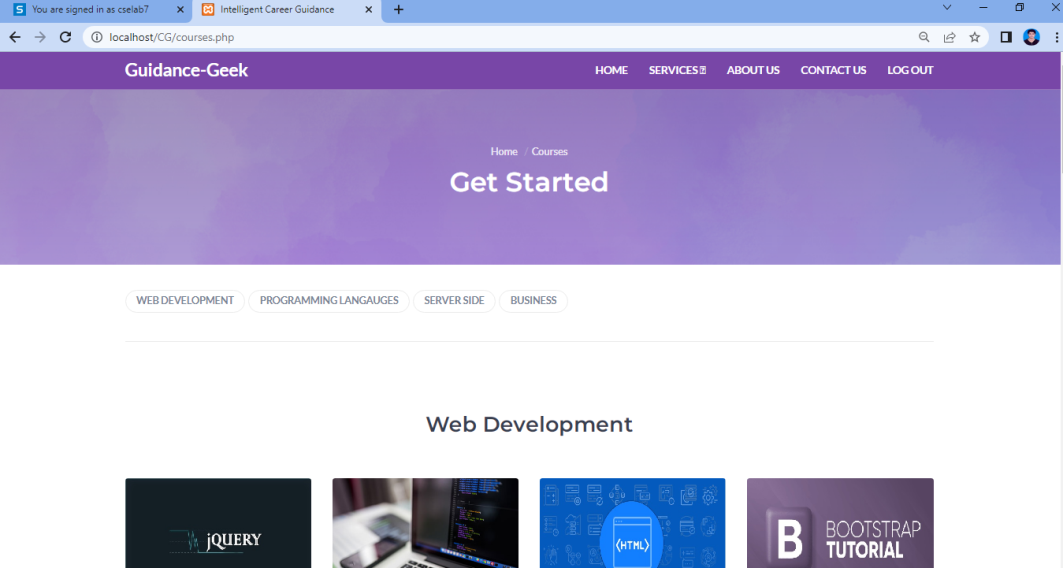
With the help of contact us page, user can provide us some feedbacks or any difficulties faced by the user. The below figure 7.11 shows the contact us page:

****

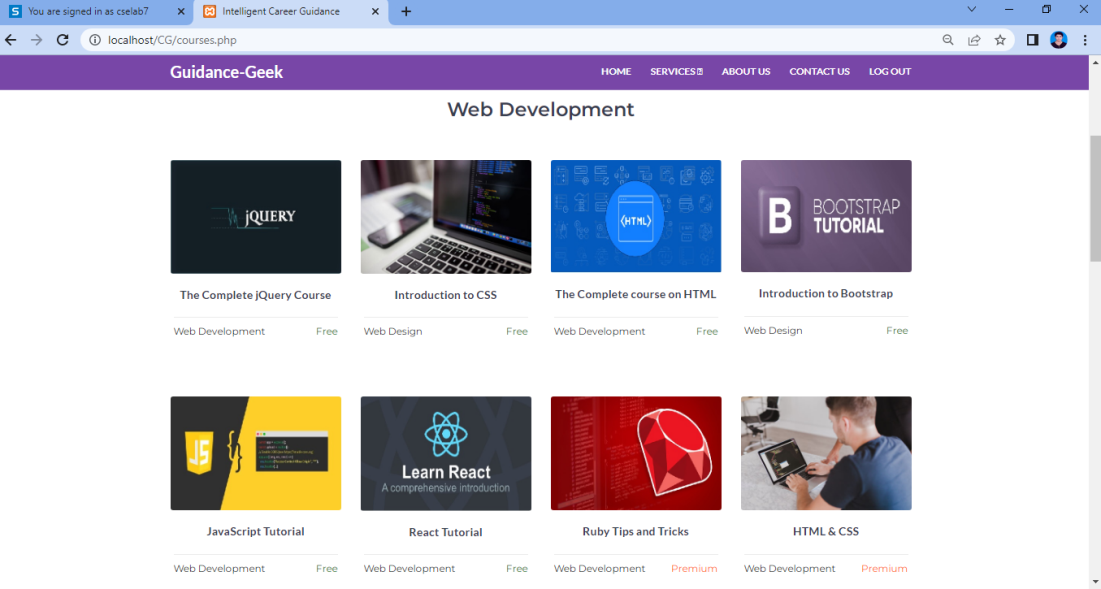
**Fig. 7.11** Contact us page

* **Courses page**

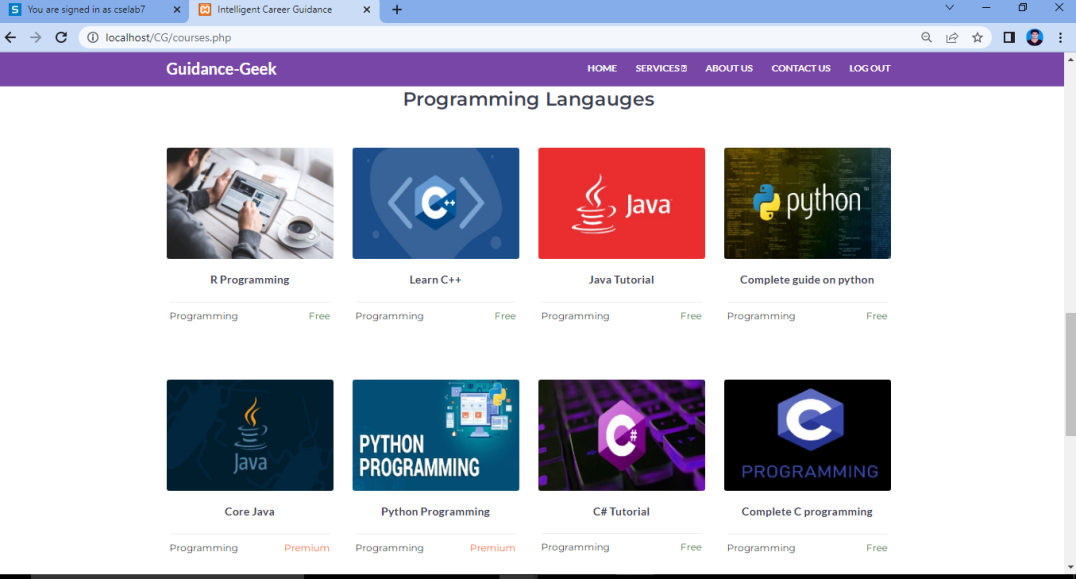
The courses page provides user some of the best resources which will be helpful for the user to pursue the career in various domains. The courses page contains free as well as the paid courses. Figures 7.12, 7.13, 7.14 and 7.15 represent the courses page.

****

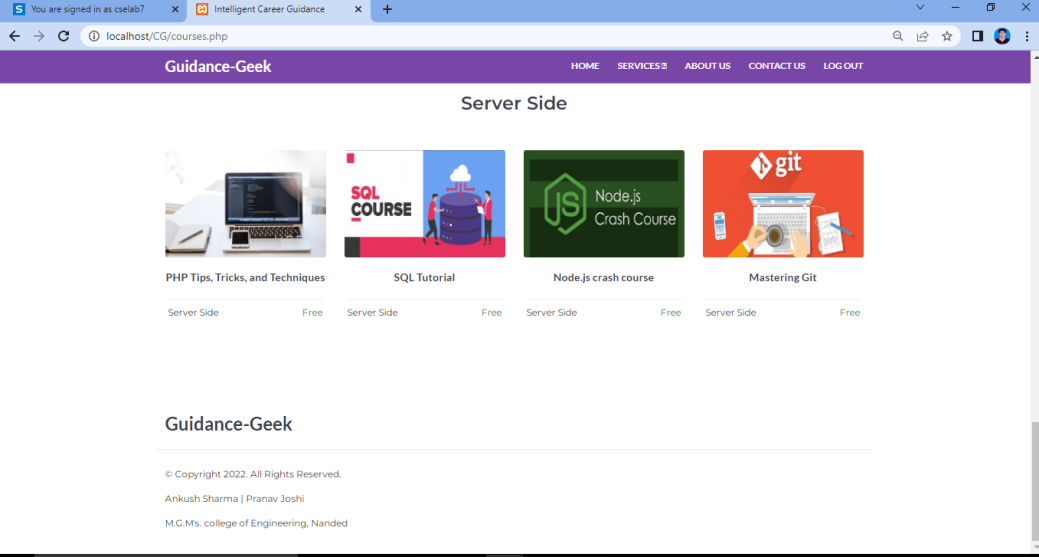
**Fig. 7.12** Courses page (1)

****

**Fig. 7.13** Courses page (2)

****

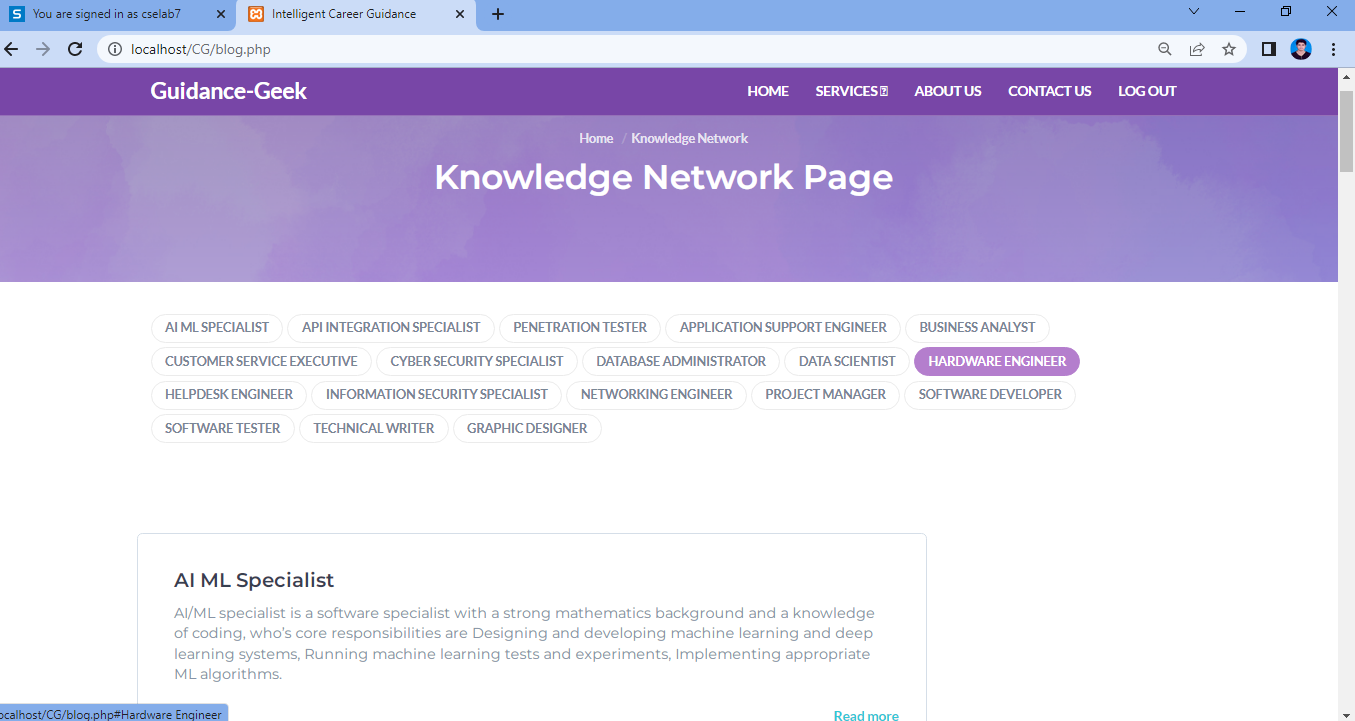
**Fig. 7.14** Courses page (3)

****

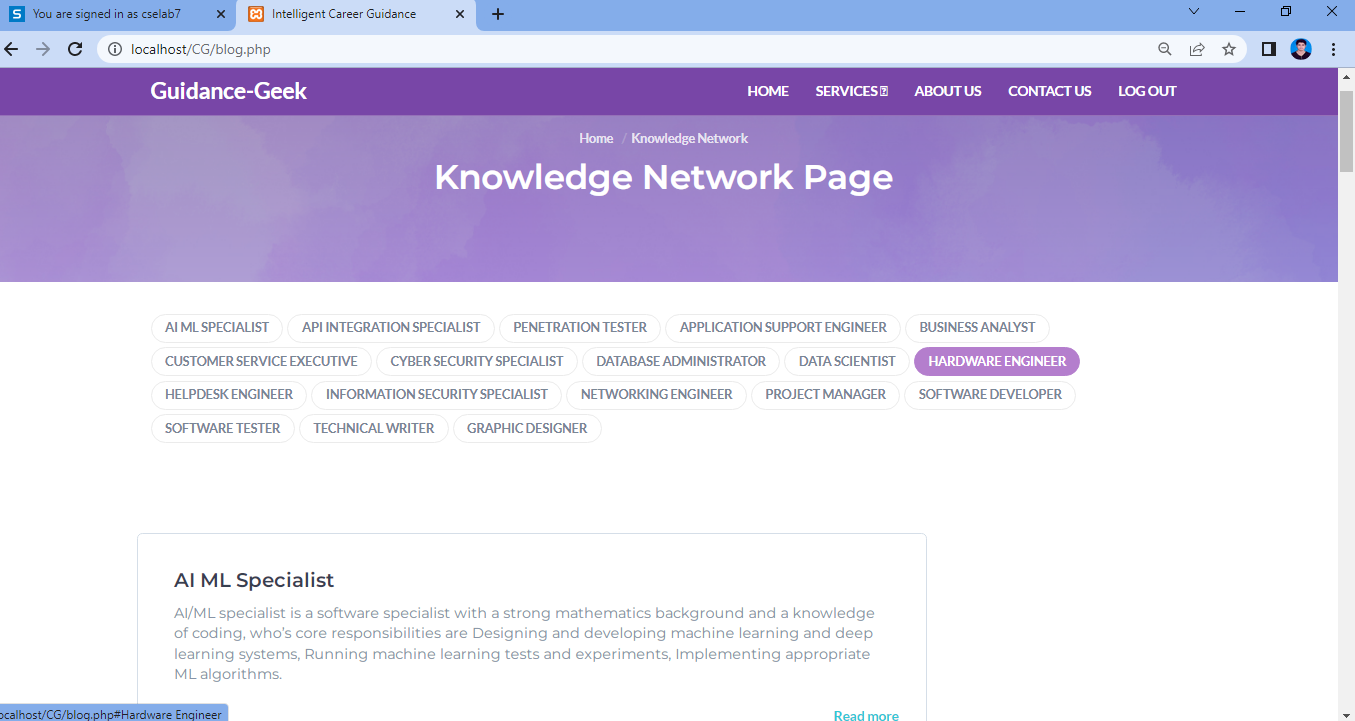
**Fig. 7.15** Courses page (4)

* **Knowledge network page**

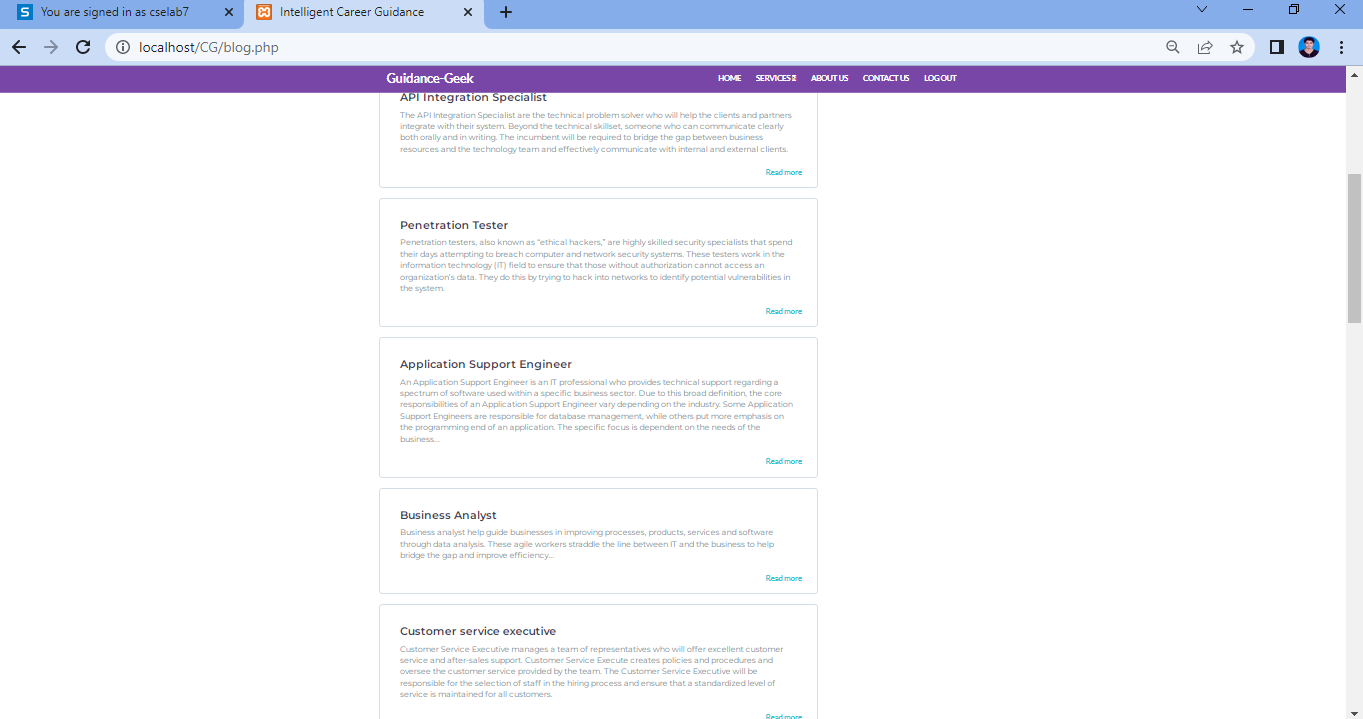
This page provides the information about the various job roles. Below figures 7.16, 7.17, 7.18, 7.19, 7.20, and 7.21 represent knowledge pages:

****

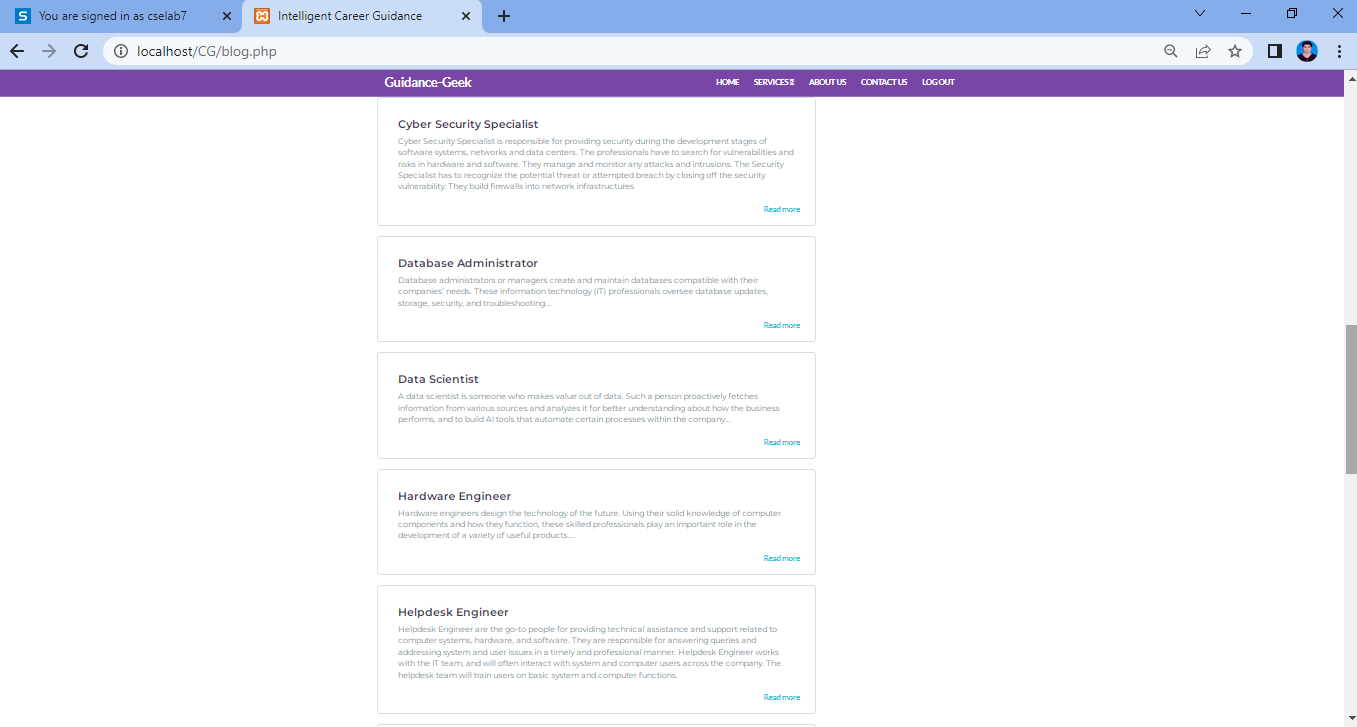
**Fig. 7.16** Knowledge Network page (1)

****

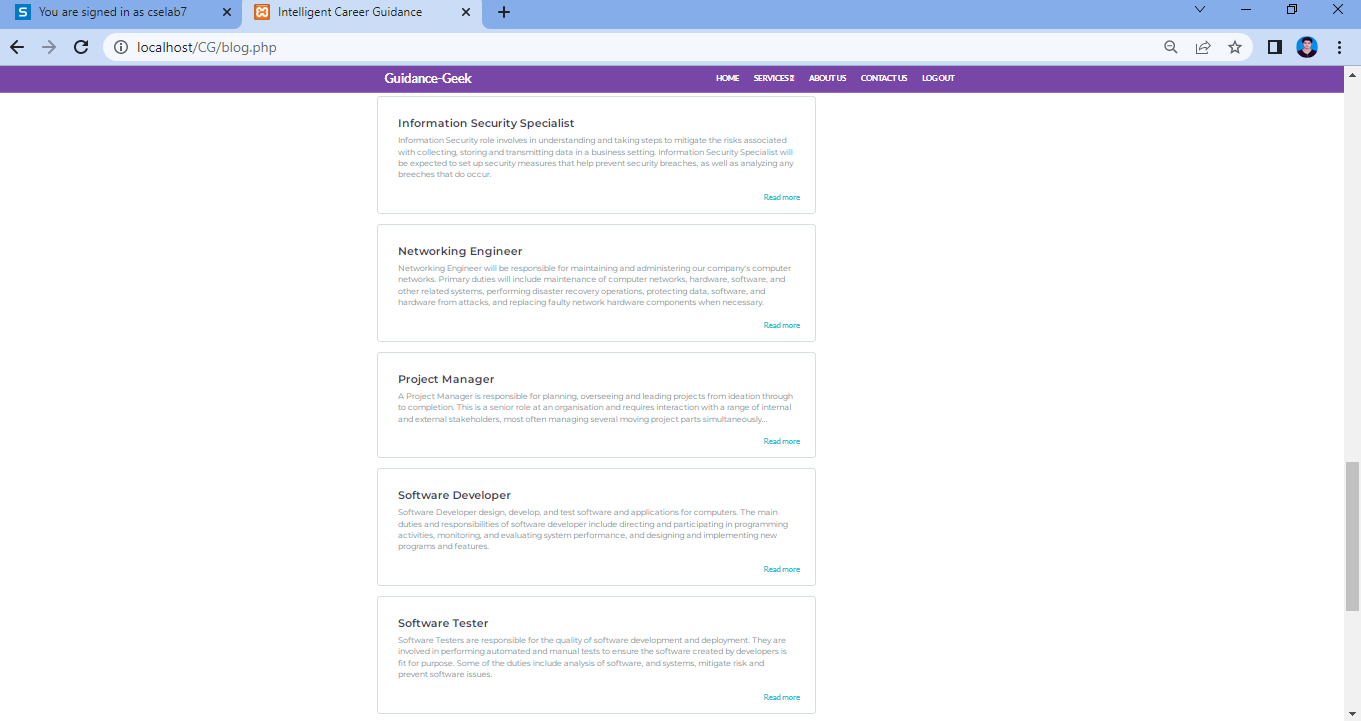
**Fig. 7.17** Knowledge Network page (2)

****

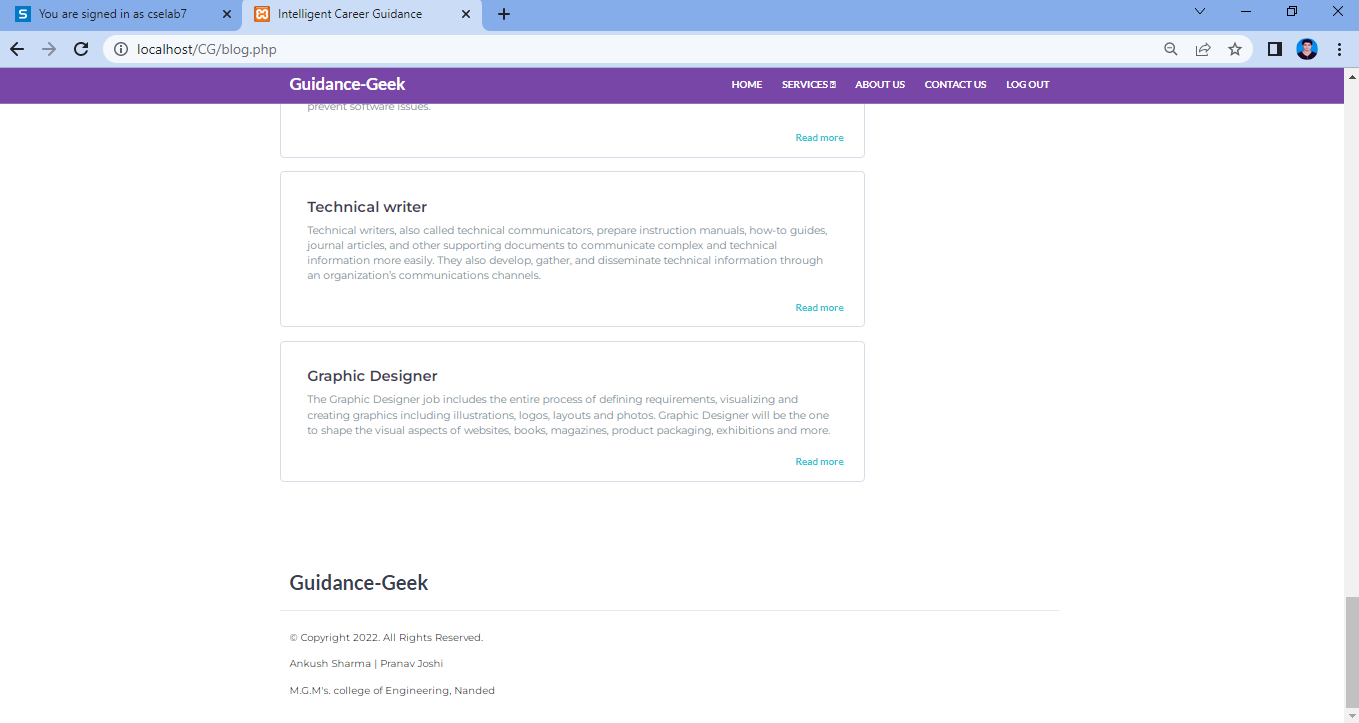
**Fig. 7.18** Knowledge Network page (3)

****

**Fig. 7.19** Knowledge Network page (4)

****

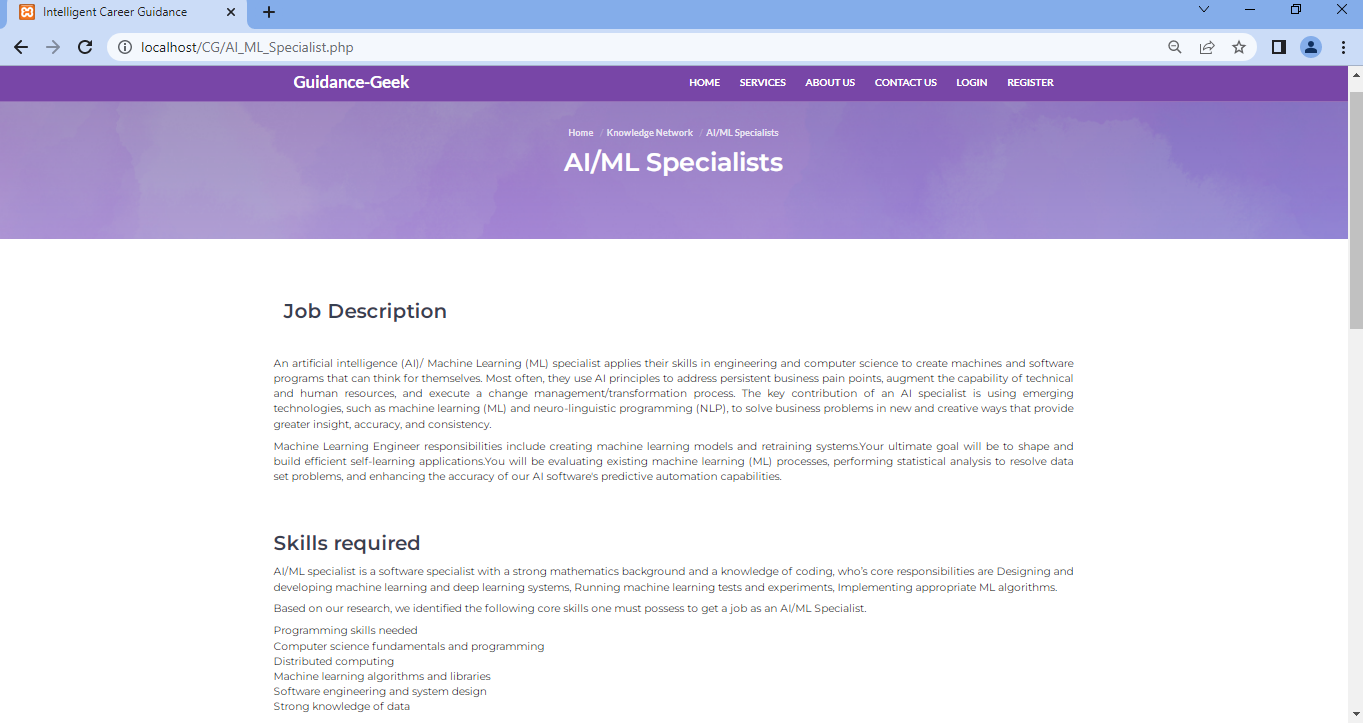
**Fig. 7.20** Knowledge Network page (5)

****

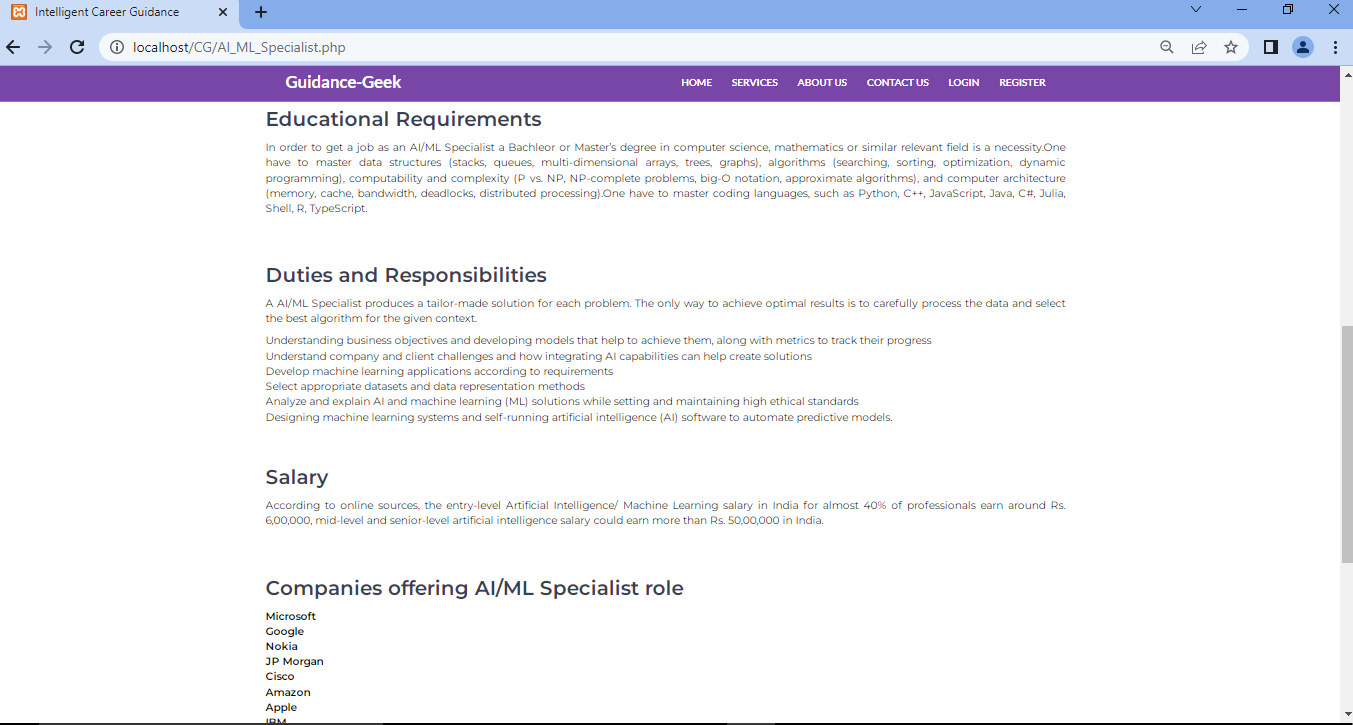
**Fig. 7.21** Knowledge Network page (6)

* **Job role example**

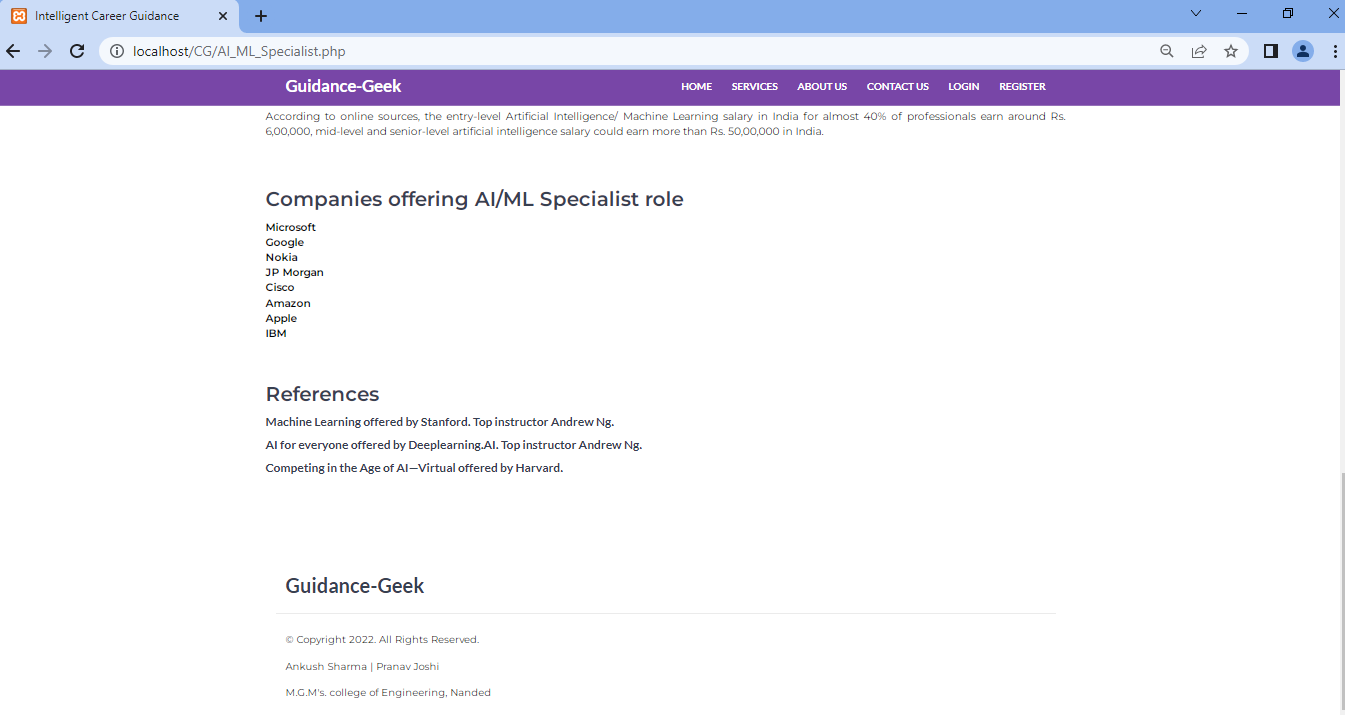
When the user clicks on any of the job role provided in the knowledge network page, he/she will get redirected to the job role page where the user can get in detail information about the corresponding job role along with the job description, skills and education requirement, roles and responsibilities, salary, companies which are offering this job role and some resources for the same. Below figures 7.22, 7.23 and 7.24 shows the job role page:

****

**Fig. 7.22** Job role page (1)

****

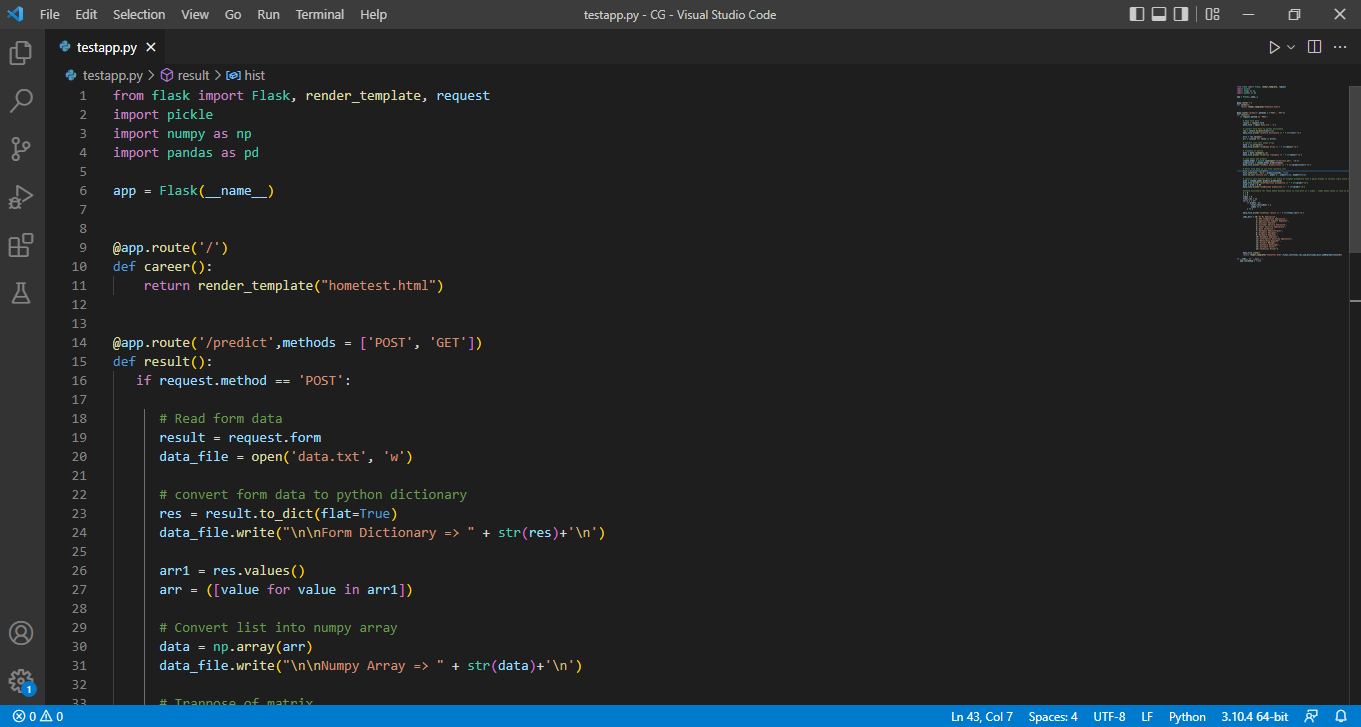
**Fig. 7.23** Job role page (2)

****

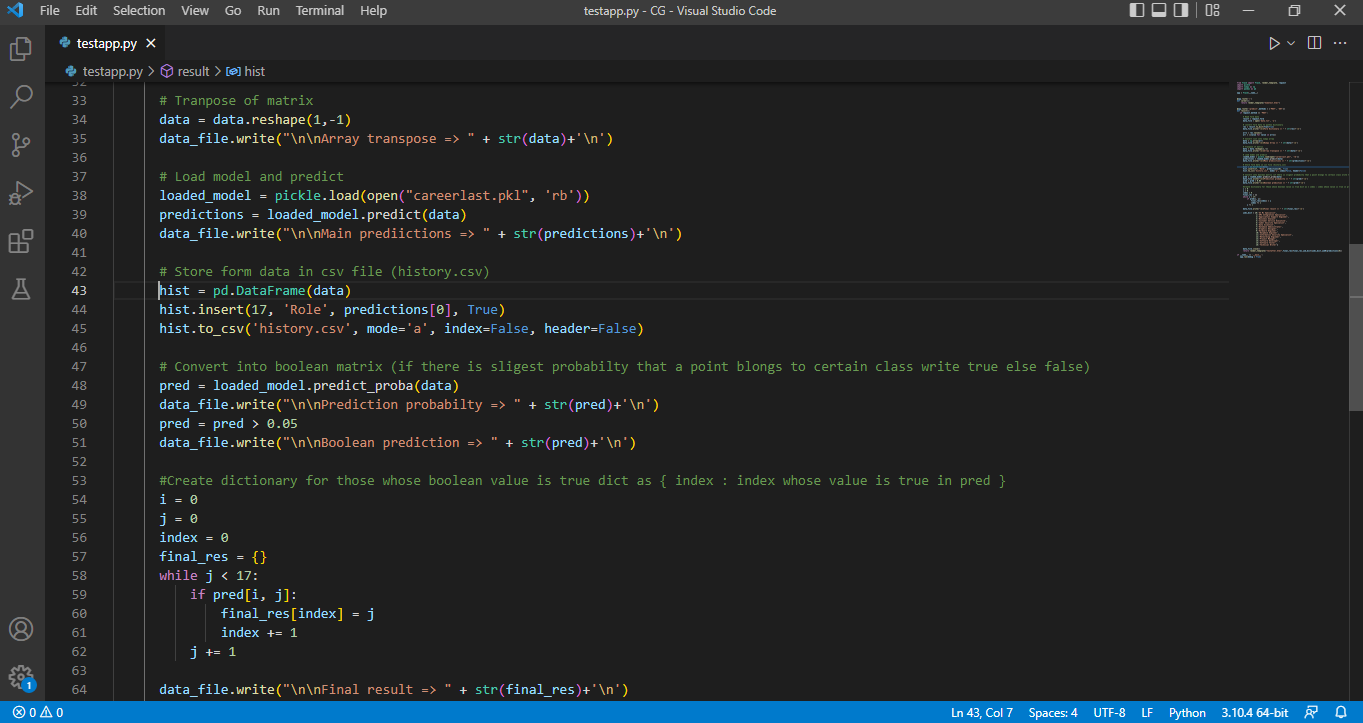
**Fig. 7.24** Job role page (3)

* **Flask code**

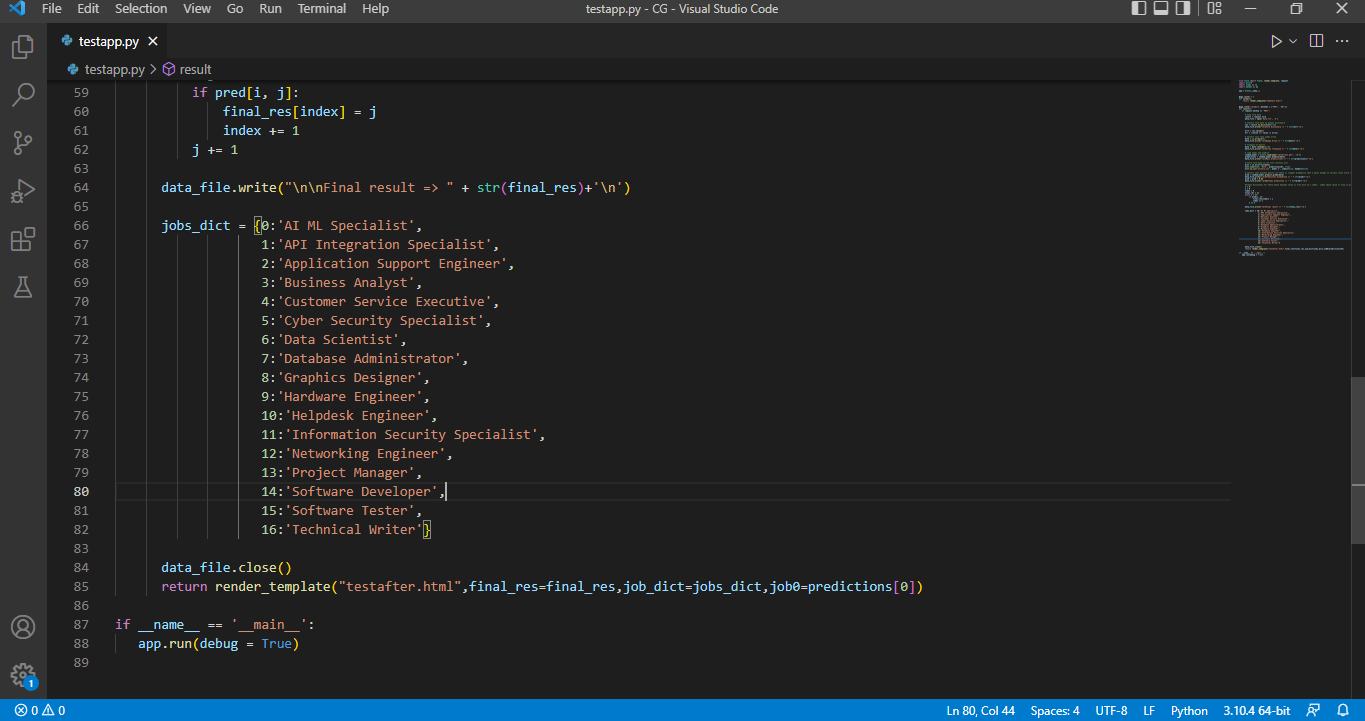
Figure 7.25, 7.26 and 7.27 shows flask code used run background processes on after receiving inputs from the form. Its uses pre-build machine learning model and predicts the outcome.

****

**Fig. 7.25** Flask Code (1)

****

**Fig. 7.26** Flask Code (2)

****

**Fig. 7.27** Flask Code (3)

* **Database**

Database used in this project is MYSQL. Below figure 7.28 shows entries of the data base name ‘my\_db’ and table name ‘users’.

****

**Fig. 7.28** Database

**CONCLUSION**

Machine learning is quickly growing in the field of computer science. It has applications in nearly every other field of study and is already being implemented commercially because it can solve too difficult or time-consuming problems for humans to solve.

In this ML project, with the help of proficiency of students in various subjects/domains we trained, tested the machine learning model and integrated it with the website so that a student can get the suggested job roles along with the descriptions of different jobs and the corresponding resources required for them.

The two algorithms used were logistic regression and K Nearest Neighbours. These are the two most popular and basic classification algorithm being used. Since none of the algorithm is better than the other and one’s performance is often credited to the nature of dataset being work upon. Both the model gave nearly same accuracy 92.78 and 99.01 respectively.

Thus, with the help of this model, students as well as different organisations can evaluate the expertise of a particular student with respect to one or more domains. So that it will help students to pursue more suitable job roles and organisations can put the students in particular domains, so that the organisation can expect better performance from their employees which will help them in various aspects.

**REFERENCE**

1. [https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning%20)
2. <https://monkeylearn.com/blog/data-preprocessing/>
3. <https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>
4. <http://theprofessionalspoint.blogspot.com/2019/03/advantages-and-disadvantages-of.html>
5. <https://www.w3schools.com/w3css/defaulT.asp>
6. <https://www.tutorialrepublic.com/php-tutorial/>
7. <https://flask.palletsprojects.com/en/2.1.x/tutorial/>
8. <https://javascript.info/>
9. <https://www.analyticsvidhya.com/blog/2020/09/integrating-machine-learning-into-web-applications-with-flask/>
10. <https://www.mysqltutorial.org/>