

# Project Domain Prediction Techniques Using Machine Learning

Imana Azram<sup>1</sup>, L S Maurya<sup>2</sup>

<sup>1</sup>Research scholar, <sup>2</sup>Professor

<sup>1,2</sup>Department of Computer Science & Engineering

<sup>1,2</sup>Shri Ram Murti Smarak College of Engineering & Technology, 13 Km, Bareilly-Nainital Road, Ram Murti Puram, Bareilly - 243202, (U.P.) - INDIA

---

## ABSTRACT

Machine Learning has surpassed every other technology in the previous several years, including artificial intelligence. You may be able to expand your horizons of thinking and develop some very outstanding real-world projects if you use Machine Learning techniques. Students' chances of reaching professional success may be improved by giving them advice on which disciplines or tracks to follow based on information from other successful academics in their area as well as the students' personal histories and interests, according to the authors. A data mining and machine learning-based strategy was developed in response to the information provided. Several academic institutions and students from two various academic fields provided data for the objectives of training, evaluation, and certification. (hamed, B., 2017). Different Machine learning algorithms like gaussian naive bayes, support vector machine, logistic regression, stochastic gradient descent, decision tree, random forest were used to develop methods that had a high degree of accuracy in their predictions. In the future, the technique that has been established may be utilized to choose a new subject or stream (Mohamed, B., 2017).

**Keywords:** Artificial intelligence, Dataset, machine Learning, Prediction

---

## I. INTRODUCTION

Computing, data processing, systems control, complex algorithms, and artificial intelligence are just a few of the applications of computers that are covered in computer science classes. Aspects of computer science include programming, design, analysis, and theory. Projects in computer science engineering can require the development of several different types of software. It is possible to complete computer science project ideas by using Java, .NET, Oracle, and other technologies. Students in groups based on the team and the Moodle LMS were able to profit from the findings of the investigation. Learning management systems (LMSs) are used in educational

institutions as a form of electronic communication. By using the LMS platform, you will have quick and easy access to current content and information. The learning management system (LMS) includes a wide range of features, including data transmission, brand integration, collaborative learning, tailored learning, and an intuitive user interface, among others. Information on administration activities, official papers, daily reporting, educational course activities, and the sharing of teaching and learning program materials may all be documented and shared with the help of these technological advancements[1].

It is necessary to implement three elements for education to be successful: human capital theory; discourses; and the capability approach. Secondary schooling, which is the most essential of the three periods of education, has a significant impact on students' career goals and course choices, among other things. Every year, thousands of secondary school students throughout the globe are confronted with the dilemma of which career route to pursue the following graduation. The choice to pursue a career in a certain subject by a student may be a good indicator of their future. Choosing a stream (a set of topics) from a young age, on the other hand, may be problematic. Because of peer or family pressure or a lack of understanding, students who select the incorrect majors may wind up failing or dropping out of school entirely. As a consequence, if students wish to have a more meaningful future, they must exercise great caution while choosing the academic route that is best appropriate for them.

In the center of the stream, a diverse variety of factors influence the decision-making process. When students pick a stream based on the influence of celebrities, such as role models and politicians, as well as actors and actresses, they usually fail to discover their abilities and attitudes as a result of this. Because their recommendations are based on current trends and market demands, students can benefit from online career guidance provided by teachers and tutors, as well as coaching centers and tuition centers. However, students should avoid choosing the wrong stream based on their interests and early performance to avoid making this mistake. When it comes to systems that aid students in picking a stream or topic, data mining (DM) and machine learning (ML) techniques based on data of successful scholars from varied streams/subjects are not often employed. It was also observed that present methods of career counseling do not depend on hands-on experience or mathematical equations to assist students in making an educated choice about which area of study to follow after completing their undergraduate degree.

They are not the best alternatives for many students since they are expensive, time-consuming, and out of reach for many of them. Medical, non-medical, business, the arts, and social science are the most common options for students who desire to have a successful career in their chosen field. When a student achieves exceptional academic results, many people anticipate that they would pursue a career in the scientific field. The secondary school years are even more important in countries like India since it is during this period that students must make a choice on which career route to pursue after completing their secondary education. Students are not permitted to alter their majors after they have made a decision on which one they want to pursue. If you're a student in the business area, you can't take both the medical and non-medical tracks at the same time. A person's job path is most often determined by one of three methods: being influenced by family and friends, making one's own choices, and soliciting the assistance of specialists such as

career counselors. Among these are critical thinking and problem-solving abilities, the ability to find and analyze new information, interpersonal skills that enable people to work with others and engage effectively in cross-cultural situations, self-directed abilities that enable people to manage their work and complex projects, the ability to competently find resources and use tools, and the ability to communicate effectively in a variety of ways [2].

### **A. Objective**

1. To develop classification models for predicting suitable project domains of the engineering undergraduate students of CSE/IT.
2. To compare the accuracy of different developed classification models.
3. To use the above-developed classification models for predicting the project domain on new data.

## **II. LITERATURE REVIEW**

The creation of a powerful machine learning system is a top priority. A project is an excellent approach to improve your resume. Furthermore, machine learning is an area in which you may potentially spend your whole professional life. The opportunities for invention, construction, and progress are infinite. Last but not least, high-tech companies such as Google, Amazon, Microsoft, and others are having difficulty finding skilled Machine Learning Engineers. During this new era of machine learning and artificial intelligence, they have the chance to make a significant impact. Machine learning is the process through which a computer learns from the data it has previously gathered and uses that knowledge to generate predictions [3]. To achieve this aim, we will develop a project that uses machine learning to predict the price of a property using data from other properties in the area. For total beginners, this is the most effective method of learning about and comprehending machine learning.

Over the previous decade, there has been a total of \$32.320 trillion in fraud in the credit card processing industry. This is a difficulty that all financial institutions are seeking to solve via the use of Machine Learning. Machine learning (ML) will be used to identify fraudulent credit card transactions based on past banking data in this project [4]. As a result, machine learning is increasingly being used in data science applications for the analysis of complex relationships, which is becoming more common. It can learn even when it has not been expressly programmed. A model known as an Artificial Neural Network (ANN) has a long history in computers and data science, but it is gaining popularity and being used in a growing number of different applications. When neural networks are used to analyze complicated data sets that cannot be simplified using normal statistical approaches [5], the ability to analyze these data sets is increased. Furthermore, it has the capability of detecting non-linear correlations between dependent and independent variables using indirect methods of analysis. The use of artificial neural networks (ANN) in a variety of sectors, including healthcare, climate and weather, financial markets, pattern identification, classification, forecasting, and prediction, is gaining prominence [6].

Students' progress monitoring is used by universities all around the world to ensure that their students succeed in their studies. Every semester, schools and institutions amass massive amounts

of information. This information may be used by advisors and instructors to spot trends and patterns, which in turn can help students thrive in their academic endeavors. Due to the comprehensive data-keeping rules of universities, educational data may be available in a range of granularities. For example, In the majority of cases, data is dispersed throughout several databases and systems, each of which is linked to a distinct set of databases and systems. These platforms are often connected to the university's primary content management system. Typical examples of this kind of software are content management systems such as Moodle and Banner. Universities understand the importance of their data and the potential it has for assisting them in making better decisions regarding student recruitment and retention. For colleges, data mining provides channels and techniques that may be used to extract relevant information that can be used to benefit students, which is a top priority [7]. Among the concepts used in educational data mining are data mining, machine learning, statistics, pedagogical techniques, psychology, recommender systems, and visualizations, to name a few. Educational data mining is an interdisciplinary field that helps advisors and instructors better understand student performance and devise intervention strategies by using data mining, machine learning, and statistics [8]. Student retention and satisfaction are the primary goals of the final result, which is a complete approach. The use of educational data mining technologies may result in a better knowledge of the factors that impact students' performance. Computer-supported learning analytics, computer-supported behavioral analytics, computer-supported visualization analytics, evaluation of learning materials, computer-supported predictive analytics, self-learning behavior, and social network analysis are just a few of the techniques described in the paper, which include, but are not limited to, About classification rates, CSPA is quite efficient [9]. It may also be used to find trends and, as a result, to develop learning models in educational data, which is extremely useful. Keeping track of a student's progress is a difficult task to do. Each student, as a person, brings to the table a unique set of circumstances that influence his or her ability to succeed. Consequently, a solution that is one size fits all will fall short in at least one aspect. As a result, it is obvious that, depending on the goals of the research, there are several methods for assessing performance. A significant effect has been made by using educational data mining to monitor student success [10]. These types of projections have been made using methods such as classification and regression analysis. Advisors and instructors may utilize these tactics to offer better-informed suggestions to their clients and students. In addition, using predictive analysis, it is feasible to measure individual student performance, which is not achievable with aggregate analysis methodologies like those used in visualization analytics. Predictive analysis is performed. In recent years, academics have been interested in the prospect of forecasting student achievement based on their grade point average (GPA) [11]. Students' performance in previous semesters may be utilized to predict how well they would do this semester, based on their previous performance. Some studies include pre-university courses and SAT scores in their findings, while others depend on information gathered from discussion forums in which students often participate [12].

### **III. PROPOSED WORK**

It is feasible to categorize both structured and unstructured data, and the classification technique may be used for either or both types of data. The initial stage in the technique is to predict the class of data points that have been presented. The phrases target, label, and category are some of

the most popular terminologies used to refer to the many types of classes. Classification algorithms are used in machine learning to estimate the likelihood or probability that the next data will fall into one of the given categories based on the training data that has been provided.

### **A. Gaussian Bayes**

A classification technique based on the assumption that predictors in a dataset are independent is known as the Gaussian Bayes algorithm. This suggests that it is supposing that the characteristics are unconnected to one another.

### **B. SVM (Support vector machine)**

Classification and regression issues may be solved using Support Vector Machines (SVMs), one of the most common supervised learning methods. However, it is mostly utilized in Machine Learning to solve Classification difficulties. Finding the optimum line or decision boundary that can divide n-dimensional space into classes so that we may easily classify fresh data points in the future. A hyperplane denotes the border of the optimal choice. As a result, SVM selects the hyperplane's most extreme points/vectors for use [13].

### **C. Stochastic Gradient Descent**

SGD is a simple yet effective approach for fitting linear classifiers and regressors under convex loss functions, such as those used by (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around for a long time in the machine learning area, it has only just begun to get substantial attention in the context of large-scale learning and data mining. There are no specific machine learning models that SGD can be used to; rather, it is an optimization strategy that may be applied to any model. This is the only method a model can be taught [14]. In many cases, a sci-kit-learn comparable estimator will be available for SGD Classifier or SGD Regressor instances, which may use a different optimization technique.

### **D. Logistic Regression (For Multiclass)**

If you are working in the setting of logistic regression, it is feasible to handle multi-class classification problems by using multinomial logistic regression. When using logistic regression, the default configuration is to deal with two-class classification problems. When dealing with multi-class classification problems, numerous extensions, such as one-vs-rest, make it possible to use logistic regression.

Not a new approach, but rather an extension of the logistic regression model that uses a cross-entropy loss function and predicts a multinomial probability distribution to natively handle multi-class classification challenges, rather than a new technique.

### **E. Decision Tree**

While a tree has multiple real-world analogs, its significance in machine learning goes far beyond classification to encompass regression as well as other operations on data. Indecision analysis, a decision tree may be used to show alternatives and decision-making in a visually appealing and

understandable manner. The decision-making process is organized in a way that is evocative of a tree's structure. When it comes to machine learning and data mining, this technique is usually used to generate a plan for accomplishing a certain goal.

It's simple enough to get the essence of. It's feasible to envision trees in one's head.

- The amount of time spent on data preparation is small. Other operations often need the normalization of data, the generation of dummy variables, and the elimination of blank values, among other things. You should be aware, however, that this module does not handle the case of missing values.
- In general, the cost of using the tree (i.e., predicting data) is inversely proportional to the number of training data points that are used.
- Ability to cope with both numerical and categorical information Categorical variables, on the other hand, is not supported by the present version of sci-kit-learn. Other solutions are more specialized for datasets that only include a single kind of variable, such as categorical data. There is information about algorithms that can be found on this website.

## **F. Random Forest**

When it comes to classification and regression, it is usual practice to use supervised machine learning techniques such as a random forest. It employs a mixture of diverse samples to generate decision trees, which are subsequently averaged, to do classification and regression analysis.

One of the most important characteristics of the Random Forest Algorithm is that it can handle both continuous and categorical data sets, making it ideal for both regression and classification applications. Problems with classification are made simpler to address when this program is used.

During the row sampling procedure, this phase is referred to as the bootstrap phase. As a consequence, it is now able to train each model independently, which provides results. All of the models are integrated to produce the final result, which is selected by a majority vote of the participants. As the name indicates, aggregation entails combining all of the results and then providing an output based on a majority of the votes cast in a particular poll or vote round.

- In this way, the curse of dimensionality is avoided since each tree only takes into account one or a few attributes.
- To add another layer of complexity, each tree is constructed from a distinct collection of data and characteristics. As a consequence, we can make full advantage of the CPU's processing power while simultaneously constructing random forests.
- Because the decision tree only has access to 30% of the data, a random forest does not need a separate training and testing set of data for training and testing.
- Because the decision is determined by majority voting and average voting, there is a feeling of predictability [15].

## **IV. RESEARCH DESIGN AND METHODOLOGY**

### **A. Data Collection**

Data is collected by designing the google form. Our google form will consist of 14 questions where 3 questions related to username, contact no. and email ID respectively and 10 questions in the form will be related with the input feature for classifier building where students rated themselves on various skills needed to create a project on IT/CSE domain. 1 Question in the google form will be related to the output parameter that is the respondent's project domain. Total 400 observations were collected. Our questionnaire will contains question like this–

1. User Name
2. Contact No.
3. Email ID
4. Rate your knowledge self in C/C++
5. Rate your knowledge self in Java
6. Rate your knowledge self in Python
7. Rate your knowledge self in Web Technology
8. Rate your knowledge self in DBMS
9. Rate your knowledge self in Software Engineering
10. Rate your knowledge self in Data Structure
11. Rate your knowledge self in Algorithms
12. Rate your knowledge self in Mathematics
13. Rate your knowledge self in Statistics
14. Which domain do you prefer for a project?

### **B. Methods**

To accomplish our research work we will be adopting the following research methodology. The Algorithm of the research methodology to be used consists of following steps.

Step 1 - Start

Step 2 - Conceptualization of the idea

Step 3 - Problem statement/Topic Finalization

Step 4 – Identification and selection of input parameters

Step 5 – Identification and selection of output parameter/domain

Step 6 - Preparing questionnaire for the data collection

Step 7 – Designing google form on the basis of questionnaire

Step 8 – Data Collection

Step 9 – Data Cleaning

Step 10 – Identification and selection of suitable classification algorithm

Step 11 – Implementation of classification algorithm on training data

Step 12 – Testing the classification model to predict the accuracy score

Step 13 – Generate the classification report

Step 14 – Generate the heat map

Step 15 – Comparing the accuracy of different classification models

Step 16 - Stop

### **a. Feature Engineering**

Any machine learning algorithm needs well transformed data to predict the outcomes. So, it's a best practice to clean and transform the data before applying any machine learning algorithm to improve the accuracy. The process of cleaning and transforming the features to improve accuracy of machine learning is called feature engineering.

#### Encoding Categorical Features

Where values of rating features are categorized as Excellent, very good, good, average, poor. Since, obtained data are categorical in nature therefore we have to convert them into numerical values as machine learning algorithms work best on numerical data. We have converted each categorical value as - Excellent - 5, Very good - 4, Good - 3, Average - 2, Poor - 1 and for Dependent feature, we used label encoder to encode the dependent variable.

0 - AI, 1 - Cloud, 2 - IOT, 3 - IP, 4 - MA, 5 - Others, 6 - WA these are our encoded dependent variable values.

### **b. Model Preparation**

#### Training and Testing

Our dependent variable is Project domain and Independent variables are the rating of students. For training, we chose 80% of data and 20% data for testing.

**Six popular classification algorithms are used for training our data.**



These are:

- Gaussian Naive Bayes
- SVM (Support Vector Machine)
- SGD (Stochastic Gradient Descent)
- Logistic Regression (For Multiclass)
- Decision Tree
- Random Forest

## V. RESULTS

### A. Accuracy

In this section, we have described the obtained results from the s experiments that have been proposed for this study. We calculate the Accuracy score, R2 Value, Log loss, Mean squared error, Confusion matrix, Heatmap and ROC curve for each algorithm.

**Table 1.** Accuracy score, Hyperparameter values and random state of each classification algorithm.

S. No.	Algorithm	Hyper-parameter values	Random State	Percentage Accuracy score
1	Gaussian Naive Bayes	Default	24	43.75%
2	Support Vector Machine	kernel='rbf', gamma=0.5, c=0.1, probability =True	95	45.0%
3	Stochastic Gradient Descent	loss = 'hinge	42	42.5%
4	Logistic Regression	multi_class = 'multinomial', Solver = 'lbfgs'	42	47.5%

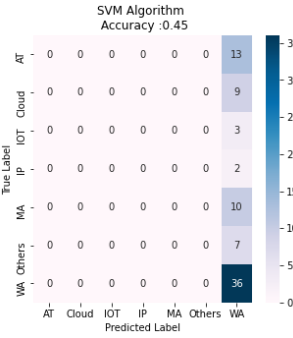
S. No.	Algorithm	Hyper-parameter values	Random State	Percentage Accuracy score
5	Decision Tree	max_depth = 5	42	38.75%
6	Random Forest	n_estimators = 200, random_state = 44	42	41.25%

### B. Confusion Matrix and Heatmap

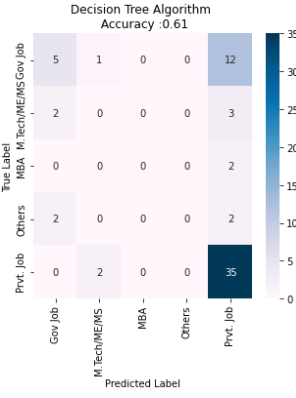
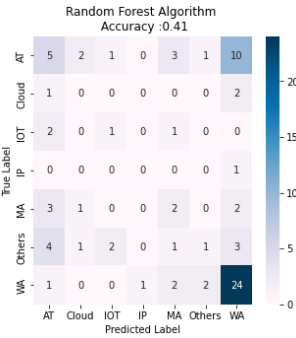
Since we know the actual value of our test data, now that we have our predicted value so with this we can develop a confusion matrix, which describes the performance of classification algorithms on a set of test data.

**Table 2.** Confusion matrix, heatmap of each algorithm

S. No.	Algorithm	Confusion Matrix
1	Gaussian Naive Bayes	<p>[[ 3, 0, 0, 1, 2, 2, 5],                      [ 1, 0, 0, 2, 0, 0, 3],                      [ 3, 0, 1, 0, 1, 0, 0],                      [ 0, 0, 0, 0, 0, 0, 3],                      [ 0, 1, 3, 1, 1, 0, 5],                      [ 0, 0, 6, 0, 0, 0, 2],                      [ 0, 0, 2, 0, 2, 0, 30]]</p>

S. No.	Algorithm	Confusion Matrix																																																																
2	Support Vector Machine	<p>[[ 0, 0, 0, 0, 0, 0, 13],                      [ 0, 0, 0, 0, 0, 0, 9],                      [ 0, 0, 0, 0, 0, 0, 3],                      [ 0, 0, 0, 0, 0, 0, 2],                      [ 0, 0, 0, 0, 0, 0, 10],                      [ 0, 0, 0, 0, 0, 0, 7],                      [ 0, 0, 0, 0, 0, 0, 36]]</p>  <p style="text-align: center;">SVM Algorithm Accuracy: 0.45</p> <table border="1" style="font-size: small; margin: auto;"> <tr> <td>True Label \ Predicted Label</td> <td>AT</td> <td>Cloud</td> <td>IOT</td> <td>IP</td> <td>MA</td> <td>Others</td> <td>WA</td> </tr> <tr> <td>AT</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>13</td> </tr> <tr> <td>Cloud</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>9</td> </tr> <tr> <td>IOT</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>3</td> </tr> <tr> <td>IP</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>2</td> </tr> <tr> <td>MA</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>10</td> </tr> <tr> <td>Others</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>7</td> </tr> <tr> <td>WA</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>36</td> </tr> </table>	True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA	AT	0	0	0	0	0	0	13	Cloud	0	0	0	0	0	0	9	IOT	0	0	0	0	0	0	3	IP	0	0	0	0	0	0	2	MA	0	0	0	0	0	0	10	Others	0	0	0	0	0	0	7	WA	0	0	0	0	0	0	36
True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA																																																											
AT	0	0	0	0	0	0	13																																																											
Cloud	0	0	0	0	0	0	9																																																											
IOT	0	0	0	0	0	0	3																																																											
IP	0	0	0	0	0	0	2																																																											
MA	0	0	0	0	0	0	10																																																											
Others	0	0	0	0	0	0	7																																																											
WA	0	0	0	0	0	0	36																																																											
3	Stochastic Gradient Descent	<p>[[ 6, 0, 0, 0, 0, 0, 16],                      [ 1, 0, 0, 0, 0, 0, 2],                      [ 0, 0, 0, 0, 1, 0, 3],                      [ 0, 0, 0, 0, 0, 0, 1],                      [ 0, 0, 0, 0, 0, 0, 8],                      [ 5, 0, 0, 0, 1, 0, 6],                      [ 2, 0, 0, 0, 0, 0, 28]]</p>																																																																

S. No.	Algorithm	Confusion Matrix																																																																
		<p>SGD Algorithm Accuracy :0.42</p> <table border="1"> <tr><th>True Label \ Predicted Label</th><th>AT</th><th>Cloud</th><th>IOT</th><th>IP</th><th>MA</th><th>Others</th><th>WA</th></tr> <tr><th>AT</th><td>6</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>16</td></tr> <tr><th>Cloud</th><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td></tr> <tr><th>IOT</th><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>3</td></tr> <tr><th>IP</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><th>MA</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td></tr> <tr><th>Others</th><td>5</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>6</td></tr> <tr><th>WA</th><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>28</td></tr> </table>	True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA	AT	6	0	0	0	0	0	16	Cloud	1	0	0	0	0	0	2	IOT	0	0	0	0	1	0	3	IP	0	0	0	0	0	0	1	MA	0	0	0	0	0	0	8	Others	5	0	0	0	1	0	6	WA	2	0	0	0	0	0	28
True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA																																																											
AT	6	0	0	0	0	0	16																																																											
Cloud	1	0	0	0	0	0	2																																																											
IOT	0	0	0	0	1	0	3																																																											
IP	0	0	0	0	0	0	1																																																											
MA	0	0	0	0	0	0	8																																																											
Others	5	0	0	0	1	0	6																																																											
WA	2	0	0	0	0	0	28																																																											
4	Logistic Regression	<p>[[ 8, 0, 0, 0, 1, 0, 13], [ 1, 0, 0, 0, 0, 0, 2], [ 0, 0, 0, 0, 1, 0, 3], [ 0, 0, 0, 0, 0, 0, 1], [ 0, 1, 0, 0, 3, 0, 4], [ 4, 0, 2, 0, 1, 0, 5], [ 3, 0, 0, 0, 0, 0, 27]]</p> <p>Logistic Regression Accuracy :0.47</p> <table border="1"> <tr><th>True Label \ Predicted Label</th><th>AT</th><th>Cloud</th><th>IOT</th><th>IP</th><th>MA</th><th>Others</th><th>WA</th></tr> <tr><th>AT</th><td>8</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>13</td></tr> <tr><th>Cloud</th><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td></tr> <tr><th>IOT</th><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>3</td></tr> <tr><th>IP</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><th>MA</th><td>0</td><td>1</td><td>0</td><td>0</td><td>3</td><td>0</td><td>4</td></tr> <tr><th>Others</th><td>4</td><td>0</td><td>2</td><td>0</td><td>1</td><td>0</td><td>5</td></tr> <tr><th>WA</th><td>3</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>27</td></tr> </table>	True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA	AT	8	0	0	0	1	0	13	Cloud	1	0	0	0	0	0	2	IOT	0	0	0	0	1	0	3	IP	0	0	0	0	0	0	1	MA	0	1	0	0	3	0	4	Others	4	0	2	0	1	0	5	WA	3	0	0	0	0	0	27
True Label \ Predicted Label	AT	Cloud	IOT	IP	MA	Others	WA																																																											
AT	8	0	0	0	1	0	13																																																											
Cloud	1	0	0	0	0	0	2																																																											
IOT	0	0	0	0	1	0	3																																																											
IP	0	0	0	0	0	0	1																																																											
MA	0	1	0	0	3	0	4																																																											
Others	4	0	2	0	1	0	5																																																											
WA	3	0	0	0	0	0	27																																																											
5	Decision Tree	<p>[[ 3, 0, 4, 0, 1, 1, 13], [ 1, 0, 0, 0, 0, 0, 2], [ 2, 0, 0, 0, 0, 1, 1], [ 0, 0, 0, 0, 0, 0, 1], [ 0, 1, 0, 0, 3, 1, 3], [ 7, 0, 0, 0, 1, 1, 3],</p>																																																																

S. No.	Algorithm	Confusion Matrix																																																								
		<p>[ 1, 1, 1, 0, 2, 1, 24]]</p>  <p>Decision Tree Algorithm Accuracy :0.61</p> <table border="1"> <tr> <td>Gov Job</td> <td>5</td> <td>1</td> <td>0</td> <td>0</td> <td>12</td> </tr> <tr> <td>M.Tech/ME/MS</td> <td>2</td> <td>0</td> <td>0</td> <td>0</td> <td>3</td> </tr> <tr> <td>MBA</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>2</td> </tr> <tr> <td>Others</td> <td>2</td> <td>0</td> <td>0</td> <td>0</td> <td>2</td> </tr> <tr> <td>Pvt. Job</td> <td>0</td> <td>2</td> <td>0</td> <td>0</td> <td>35</td> </tr> </table>	Gov Job	5	1	0	0	12	M.Tech/ME/MS	2	0	0	0	3	MBA	0	0	0	0	2	Others	2	0	0	0	2	Pvt. Job	0	2	0	0	35																										
Gov Job	5	1	0	0	12																																																					
M.Tech/ME/MS	2	0	0	0	3																																																					
MBA	0	0	0	0	2																																																					
Others	2	0	0	0	2																																																					
Pvt. Job	0	2	0	0	35																																																					
6	Random Forest	<p>[[ 5, 2, 1, 0, 3, 1, 10],                      [ 1, 0, 0, 0, 0, 0, 2],                      [ 2, 0, 1, 0, 1, 0, 0],                      [ 0, 0, 0, 0, 0, 0, 1],                      [ 3, 1, 0, 0, 2, 0, 2],                      [ 4, 1, 2, 0, 1, 1, 3],                      [ 1, 0, 0, 1, 2, 2, 24]]</p>  <p>Random Forest Algorithm Accuracy :0.41</p> <table border="1"> <tr> <td>AT</td> <td>5</td> <td>2</td> <td>1</td> <td>0</td> <td>3</td> <td>1</td> <td>10</td> </tr> <tr> <td>Cloud</td> <td>1</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>2</td> </tr> <tr> <td>IOT</td> <td>2</td> <td>0</td> <td>1</td> <td>0</td> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td>MA</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>Others</td> <td>3</td> <td>1</td> <td>0</td> <td>0</td> <td>2</td> <td>0</td> <td>2</td> </tr> <tr> <td>WA</td> <td>4</td> <td>1</td> <td>2</td> <td>0</td> <td>1</td> <td>1</td> <td>3</td> </tr> <tr> <td>WA</td> <td>1</td> <td>0</td> <td>0</td> <td>1</td> <td>2</td> <td>2</td> <td>24</td> </tr> </table>	AT	5	2	1	0	3	1	10	Cloud	1	0	0	0	0	0	2	IOT	2	0	1	0	1	0	0	MA	0	0	0	0	0	0	1	Others	3	1	0	0	2	0	2	WA	4	1	2	0	1	1	3	WA	1	0	0	1	2	2	24
AT	5	2	1	0	3	1	10																																																			
Cloud	1	0	0	0	0	0	2																																																			
IOT	2	0	1	0	1	0	0																																																			
MA	0	0	0	0	0	0	1																																																			
Others	3	1	0	0	2	0	2																																																			
WA	4	1	2	0	1	1	3																																																			
WA	1	0	0	1	2	2	24																																																			

### C. Classification Report

After getting the values confusion matrix, classification report is developed with parameters precision, recall, f1-score and support.

**Table 3.** Classification report of each algorithm

S. No.	Algorithm	Classification report
1	Gaussian Naive Bayes	precision recall f1-score support AT 0.43 0.23 0.30 13 Cloud 0.00 0.00 0.00 6 IOT 0.08 0.20 0.12 5 IP 0.00 0.00 0.00 3 MA 0.17 0.09 0.12 11 Others 0.00 0.00 0.00 8 WA 0.62 0.88 0.73 34 accuracy 0.44 80 macro avg 0.19 0.20 0.18 80 weighted avg 0.36 0.44 0.38 80
2	Support Vector Machine	precision recall f1-score support AT 0.00 0.00 0.00 13 Cloud 0.00 0.00 0.00 9 IOT 0.00 0.00 0.00 3 IP 0.00 0.00 0.00 2 MA 0.00 0.00 0.00 10 Others 0.00 0.00 0.00 7 WA 0.45 1.00 0.62 36 accuracy 0.45 80 macro avg 0.06 0.14 0.09 80

S. No.	Algorithm	Classification report
		weighted avg 0.20 0.45 0.28 80
3	Stochastic Gradient Descent	precision recall f1-score support AT 0.43 0.27 0.33 22 Cloud 0.00 0.00 0.00 3 IOT 0.00 0.00 0.00 4 IP 0.00 0.00 0.00 1 MA 0.00 0.00 0.00 8 Others 0.00 0.00 0.00 12 WA 0.44 0.93 0.60 30 accuracy 0.42 80 macro avg 0.12 0.17 0.13 80 weighted avg 0.28 0.42 0.32 80
4	Logistic Regression	precision recall f1-score support AT 0.50 0.36 0.42 22 Cloud 0.00 0.00 0.00 3 IOT 0.00 0.00 0.00 4 IP 0.00 0.00 0.00 1 MA 0.50 0.38 0.43 8 Others 0.00 0.00 0.00 12 WA 0.49 0.90 0.64 30

S. No.	Algorithm	Classification report
		accuracy 0.48 80 macro avg 0.21 0.23 0.21 80 weighted avg 0.37 0.47 0.40 80
5	Decision Tree	precision recall f1-score support AT 0.21 0.14 0.17 22 Cloud 0.00 0.00 0.00 3 IOT 0.00 0.00 0.00 4 IP 0.00 0.00 0.00 1 MA 0.43 0.38 0.40 8 Others 0.20 0.08 0.12 12 WA 0.51 0.80 0.62 30 accuracy 0.39 80 macro avg 0.19 0.20 0.19 80 weighted avg 0.32 0.39 0.34 80
	Random Forest	precision recall f1-score support AT 0.31 0.23 0.26 22 Cloud 0.00 0.00 0.00 3 IOT 0.25 0.25 0.25 4 IP 0.00 0.00 0.00 1 MA 0.22 0.25 0.24 8



S. No.	Algorithm	Classification report
		Others 0.25 0.08 0.12 12
		WA 0.57 0.80 0.67 30 accuracy 0.41 80 macro avg 0.23 0.23 0.22 80 weighted avg 0.37 0.41 0.38 80

#### D. Tree Plot of Decision Tree

Fig 1 represents the tree plot of our decision tree algorithm with max\_depth = 5 and gini impurity is chosen with entropy as the information gain.

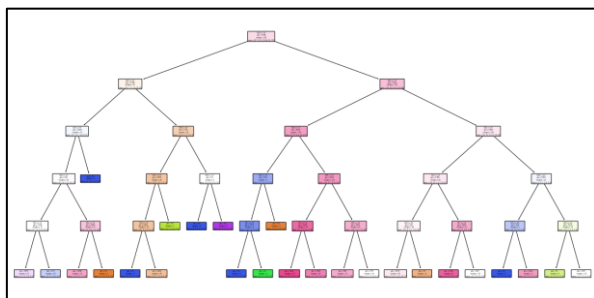


Fig 1. Decision tree with depth = 5

#### E. Mean Squared Error, R2 Value, Log Loss

MSE, R2 value and log loss are calculated for each algorithm. MSE represents how well our model is fitting with the actual data. Smaller the value of MSE, the better will be the model. R2 score is the coefficient of determination which tells how well our model is performing whereas log loss is the representation of the probabilities of each prediction with respect to actual data.

Table 4. Mean squared error R2 value and log loss.

S. No.	Algorithm	MSE	R2 Value	Log Loss
1	Gaussian Naive bayes	6.7	-0.25	1.76
2	SVM	10.07	-0.78	1.62

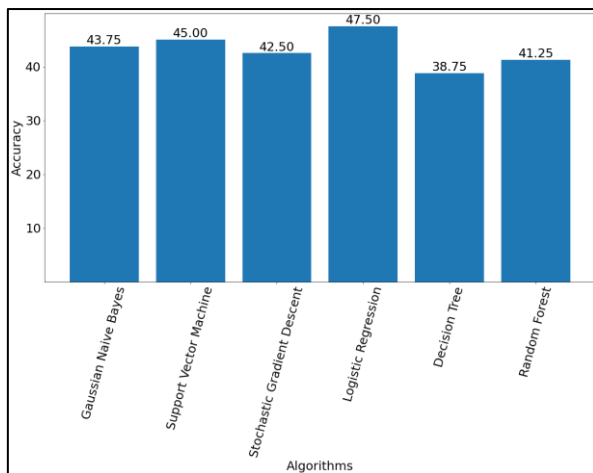
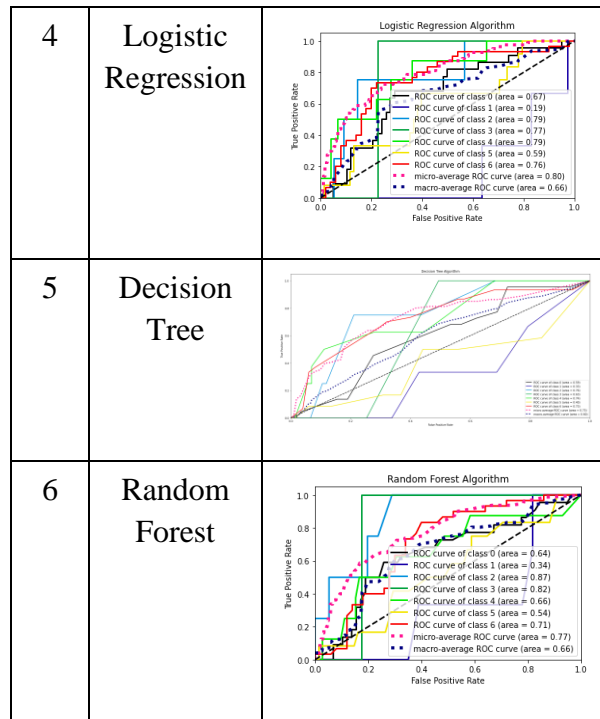
3	SGD	11.55	-0.80	1.53
4	Logistic Regression	10.66	-0.66	1.51
5	Decision Tree	11.31	-0.76	5.18
6	Random Forest	9.61	-0.50	2.0

### F. ROC CURVE (Receiving Operating Characteristic) Curve.

ROC curve shows the performance of classifiers where if any classifier comes closer to the top right corner then that algorithm is performing well whereas if classifiers come closer to the diagonal of the ROC space then the accuracy of those classifiers will be less.

**Table 5.** ROC curve representation

S. No.	Algorithm	ROC Curve
1	Gaussian Naive Bayes	<p>Gaussian Naive Bayes Algorithm</p> <ul style="list-style-type: none"> <li>ROC curve of class 0 (area = 0.58)</li> <li>ROC curve of class 1 (area = 0.73)</li> <li>ROC curve of class 2 (area = 0.71)</li> <li>ROC curve of class 3 (area = 0.71)</li> <li>ROC curve of class 4 (area = 0.57)</li> <li>ROC curve of class 5 (area = 0.27)</li> <li>ROC curve of class 6 (area = 0.77)</li> <li>micro-average ROC curve (area = 0.74)</li> <li>macro-average ROC curve (area = 0.63)</li> </ul>
2	Support Vector Machine	<p>SVM Algorithm</p> <ul style="list-style-type: none"> <li>ROC curve of class 0 (area = 0.39)</li> <li>ROC curve of class 1 (area = 0.47)</li> <li>ROC curve of class 2 (area = 0.90)</li> <li>ROC curve of class 3 (area = 0.57)</li> <li>ROC curve of class 4 (area = 0.51)</li> <li>ROC curve of class 5 (area = 0.47)</li> <li>ROC curve of class 6 (area = 0.69)</li> <li>micro-average ROC curve (area = 0.76)</li> <li>macro-average ROC curve (area = 0.61)</li> </ul>
3	Stochastic Gradient Descent	<p>SGD Algorithm</p> <ul style="list-style-type: none"> <li>ROC curve of class 0 (area = 0.66)</li> <li>ROC curve of class 1 (area = 0.23)</li> <li>ROC curve of class 2 (area = 0.78)</li> <li>ROC curve of class 3 (area = 0.75)</li> <li>ROC curve of class 4 (area = 0.79)</li> <li>ROC curve of class 5 (area = 0.41)</li> <li>ROC curve of class 6 (area = 0.70)</li> <li>micro-average ROC curve (area = 0.79)</li> <li>macro-average ROC curve (area = 0.62)</li> </ul>



**Fig 2.** Accuracy score for each algorithm

## V. CONCLUSIONS AND FUTURE WORK

In this study we proposed a machine learning based algorithm for prediction of suitable project domains of the engineering undergraduate students of CSE/IT. Then we calculated the accuracy of each classifier. We found that logistic regression performance is very good as compared to other algorithms. Performance sequences of each algorithm are represented as - **Logistic Regression > Support Vector Machine > Gaussian Naive Bayes > Stochastic Gradient Descent > Random Forest > Decision Tree**. Various other techniques have been implemented in order to get insight into the performance of each algorithm. We calculated R2 Score (Coefficient of determination), log loss, Mean squared error and classification report with parameters precision, recall, f1-score, support and compared the values of each algorithm. Then we calculated the confusion matrix and plotted the heat map of each algorithm with their ROC curve. In future, we can use ensemble learning techniques like Voting ensemble, bagging ensemble, gradient

boosting, etc which are very powerful machine learning algorithms which might give more accuracy while testing the data.

## REFERENCES

- [1] Ahmadzadeh, H., & Masehian, E. (2015). Modular robotic systems: Methods and algorithms for abstraction, planning, control, and synchronization. *Artificial Intelligence*, 223, 27-64.
- [2] Ramanan, B., Drabeck, L., Woo, T., Cauble, T., & Rana, A. (2020, December). ~ PB&J~ Easy Automation of Data Science/Machine Learning Workflows. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 361-371). IEEE.
- [3] Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.
- [4] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019, January). Real-time credit card fraud detection using machine learning. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 488-493). IEEE.
- [5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- [6] Ahmed, E., Usman, M., Anwar, S., Ahmad, H. M., Nasir, M. W., & Malik, M. A. I. (2021). Application of ANN to predict performance and emissions of SI engine using gasoline-methanol blends. *Science Progress*, 104(1), 00368504211002345.
- [7] Romero, C., & Ventura, S. (2013). *Data mining in education*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.
- [8] Azimi, S., Popa, C. G., & Cucić, T. (2020). Improving Students Performance in Small-Scale Online Courses—A Machine Learning-Based Intervention. arXiv preprint arXiv:2012.01187.
- [9] Saqr, M., & López-Pernas, S. (2021). Modelling diffusion in computer-supported collaborative learning: a large scale learning analytics study. *International Journal of Computer-Supported Collaborative Learning*. <https://doi.org/10.1007/s11412-021-09356-4>
- [10] Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. *Egitim ve Bilim*, 45(202).
- [11] Feldman, D. B., & Kubota, M. (2015). Hope, self-efficacy, optimism, and academic achievement: Distinguishing constructs and levels of specificity in predicting college grade-point average. *Learning and Individual Differences*, 37, 210-216.

- [12] González-Castaño, F. J., García-Palomares, U. M., & Meyer, R. R. (2004). Projection Support Vector Machine Generators. *Machine Learning*, 54(1), 33–44. <https://doi.org/10.1023/b:mach.0000008083.47006.86>
- [13] Amitha, A., & Meleet, M. (2019). A System for Recommendation of Medication Using Gaussian Naïve Bayes Classifier. *International Journal of Innovative Research in Computer Science & Technology*, 7(3), 100–103. <https://doi.org/10.21276/ijrcst.2019.7.3.13>
- [14] Mantas, C. J., Castellano, J. G., Moral-García, S., & Abellán, J. (2018). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing*. <https://doi.org/10.1007/s00500-018-3628-5>
- [15] Bhukya, D. P., & Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. *International Journal of Computer and Electrical Engineering*, 660–665. <https://doi.org/10.7763/ijcee.2010.v2.208>