Prediction of Admission and Jobs in Engineering and Technology with Respect to Demographic Location

Submitted in partial fulfilment of the requirements

of the degree of

BACHELOR OF ENGINEERING (Third Year)
in
COMPUTER ENGINEERING
(A.Y. 2022-2023)

by
Group No.: C1
Ashutosh Sharma (Roll No: 13)
Karthik Shetty (Roll No: 18)
Rajan Shukla (Roll No: 23)
Hariom Soni (Roll No: 36)

Under the Guidance of
Dr. Sheetal Rathi
Professor and Dean Academic, TCET

Computer Engineering

Thakur College of Engineering & Technology

Thakur Village, Kandivali (East), Mumbai-400101

(Academic Year 2022-23)

CERTIFICATE

This is to certify that the project entitled **"Prediction of admission and jobs in engineering and technology with respect to demographic location"** is a bonafide work of Mr. Ashutosh Sharma (COMP_C - 13), Mr. Karthik Shetty (COMP_C - 18), Mr. Rajan Shukla (COMP_C – 23), Mr. Hariom Soni (COMP_C – 36) submitted to the Thakur College of Engineering and Technology, Mumbai (An Autonomous College affiliated to University of Mumbai) in partial fulfillment of the requirement for the award of the degree of "Bachelor of Engineering" in "Computer Engineering"

Signature: Signature:
Name of HOD: Dr. Harshali Patil   Name of Guide: Dr. Sheetal Rathi
Department of Computer Engineering   Professor and Dean Academic

ABSTRACT

The demand for skilled professionals in the engineering and technology industries in India has been on the rise, leading to an increase in competition for admission to technical courses and job opportunities. For the same purpose universities and colleges need to identify the employability of the students and the employment potential of different engineering and technical courses to increase the chances of students getting better jobs. This paper aims to investigate different machine learning algorithms and data mining techniques which are used to analyze the educational data available and predict the employment potential of different technical courses. By utilizing these techniques, universities and colleges can better understand the factors that impact employability, and make data-driven decisions about which courses to offer and which skills to prioritize in their curriculum. The findings of this paper will provide valuable insights into the usage of machine learning to predict the employment potential of technical courses, and help universities and colleges make more informed decisions about their curriculum and student support services

ACKNOWLEDGEMENT

Ashutosh Sharma – COMPC 13

Karthik Shetty – COMPC 18

Rajan Shukla – COMPC 23

Hariom Soni - COMPC 36

TABLE OF CONTENTS

TCET

DEPARTMENT OF COMPUTER ENGINEERING (COMP)
(Accredited by NBA for 3 years, 4ᵗʰ Cycle Accreditation w.e.f. 1ˢᵗ July 2022)
Choice Based Credit Grading Scheme (CBCGS)
Under TCET Autonomy

Chapter 1: LinkedIn Profile and Blog Writing

1.1 LinkedIn Profile Screenshots

Rajan Shukla (He/Him) · 1st

Thakur College of Engineering and Technology. Computer Engineering

Mumbai, Maharashtra, India · Contact info

500+ connections

Aditya Soni, Aaryan Saygaonkar, and 134 other mutual connections

🖅 Message   More

Thakur College Of Engineering and Technology

Open to work
Software Engineer, Web Developer, Website Developer and Java Developer roles
Show details
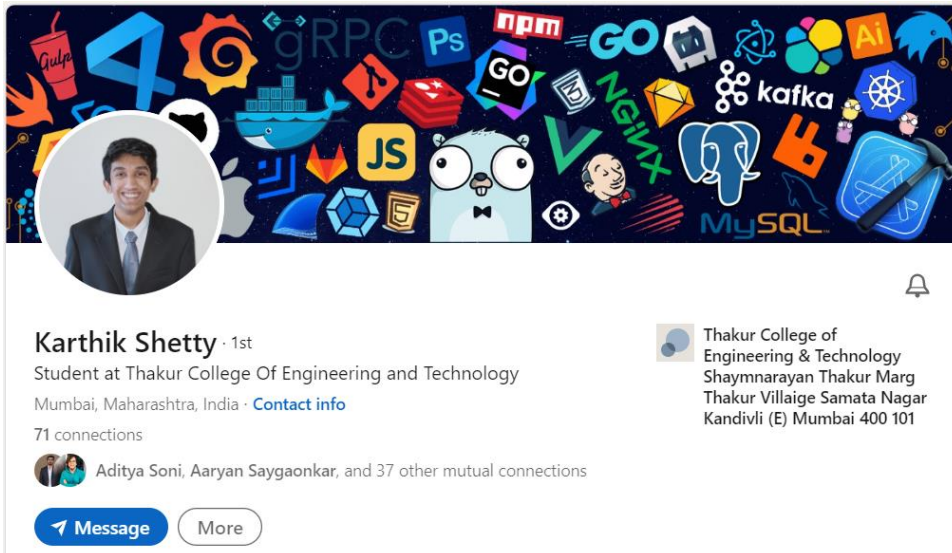
Analytics
👁 Private to you

71 profile views
Discover who's viewed your profile.

626 post impressions
Check out who's engaging with your posts.

133 search appearances
See how often you appear in search results.

Hariom Soni (He/Him) · 1st

Automation Testing | Website Development | Selenium Webdriver using Python | SQL, HTML, Python, C# | 2024 Graduation in Computer Engineering

Mumbai, Maharashtra, India · Contact info

32 connections

Aditya Soni, Aaryan Saygaonkar, and 24 other mutual connections

🖅 Message   More

Thakur College of Engineering & Technology Shaymnarayan Thakur Marg Thakur Villaige Samata Nagar Kandivli (E) Mumbai 400 101

1.2 Blog Screenshots

Understanding the Impact of Demographic Location on Admissions in Technical Courses

Prediction of Admissions in Technical Courses based on Demographic Location by Unsplash

In India, the competition for admission to technical courses and jobs in the engineering and technology industries is intense. With a growing demand for skilled professionals in these fields, it is important to understand the factors that influence an individual's chances of success. Demographic location, including factors such as geography, age, gender, and ethnicity, can play a significant role in determining an individual's admission and job opportunities in technical courses and the engineering and technology industries.

In 2020, Mishra Shakti and Mishra Anjali conducted a study on the impact of demographic location on admissions and jobs in technical courses in India. The study aimed to develop a predictive model that can accurately predict admission and job opportunities for individuals based on their demographic information and qualifications. The study used a large datasets of admission and job applicants from the engineering and technology industries in India and analyzed the data using multi-linear regression.

The results of the study indicated that demographic location has a significant impact on admission and job opportunities in technical courses and the engineering and technology industries in India. The predictive model developed in the study was found to be highly accurate, with an R-squared value of 0.90. The model accurately predicted admission and job opportunities for individuals based on their demographic information and qualifications.

3

TCET

DEPARTMENT OF COMPUTER ENGINEERING (COMP)
(Accredited by NBA for 3 years, 4th Cycle Accreditation w.e.f. 1st July 2022)
Choice Based Credit Grading Scheme (CBCGS)
Under TCET Autonomy

Which Sectors Have Hired The Most? Top Sectors
2016 - 2022

Taken from International Journal for Research in Applied Science & Engineering Technology

The study by Mishra Shakti and Mishra Anjali provides valuable insights into the relationship between demographic location and admission and job opportunities in technical courses and the engineering and technology industries in India. The results suggest that demographic location can have a significant impact on an individual's chances of success in these fields. The predictive model developed in the study can be used to help individuals make informed decisions about their education and career paths and to provide valuable information for universities and employers in India.

Future research in this area could explore the impact of other factors, such as socioeconomic status and prior experience, on admission and job opportunities in the engineering and technology industries in India. Additionally, further study could be done to evaluate the long-term impact of demographic location on career success in these industries in India.

In conclusion, understanding the impact of demographic location on admission and job opportunities in technical courses and the engineering and technology industries in India is critical for individuals seeking to enter or advance in these fields. The study by Mishra Shakti and Mishra Anjali provides valuable insights into this relationship and highlights the importance of demographic information in predicting admission and job opportunities in India.

Reference:

- Mishra Shakti, Mishra Anjali. (2020). Prediction of Admissions and Jobs in Technical Courses concerning Demographic Location Using a Multi-Linear Regression Model. [ResearchGate]. Available at: Mishra, Anjali & Kumar, Aishwary & Mishra, Shakti & Sanjay, H.. (2020). Prediction of Admissions and Jobs in Technical Courses with Respect to Demographic Location Using Multi-linear Regression Model. 10.1007/978-981-15-5788-0_65.
- https://www.ijraset.com/fileserve.php?FID=15097

1.3 URL (LinkedIn Profile/ Blog)

- Ashutosh Sharma

Profile - https://www.linkedin.com/in/ashutosh-sharma-a759ab207/

- Karthik Shetty

Profile - https://www.linkedin.com/in/karthikshetty27/

- Rajan Shukla

Profile - https://www.linkedin.com/in/rajanshukla04/

- Hariom Soni

Profile - https://www.linkedin.com/in/hariom-soni-4a9074200/

Blog Link: https://karthiks18.wixsite.com/rblc1/post/understanding-the-impact-of-demographic-location-on-admissions-and-jobs-in-technical-courses-in-indi

1.4 Count of Likes, Shares and Comments

Jan 30 · 2 min

Understanding the Impact of Demographic Location on Admissions in Technical Courses

In India, the competition for admission to technical courses and jobs in the engineering and technology industries is intense. With a...

👁 37   26 ♡

Views: 37

Likes: 26

1.5 Screenshots of Quiz

RBL-2 Session 1 Quiz (Blog Writing and LinkedIn Profile Building)

1032201072@tcetmumbai.in Switch account   Draft saved

*Indicates required question

Email *
☑ Record 1032201072@tcetmumbai.in as the email to be included with my response

Name *   0 points
Ashutosh Sharma

*
13

Branch *   1 point
Comp ▾

DIV *   1 point
C ▾

1. **What is blogging?** *   1 point
○ Creating photographic content
○ Publishing a video/audio content on web
◉ Publish written content on a website
○ All of above

---

2. **Creating and publishing video content on a blog is called?** *   1 point
◉ Vlog
○ News
○ Media blog
○ All of above

3. **The first page of a blog post is ___.html.** *   1 point
◉ index
○ blogpost
○ main
○ Header

---

4. **What is the maximum number of blog posts that can be indeed to your blog?** *   1 point
○ 5 per day
○ 0
◉ infinite
○ All the above

5. **What is crawling on a blog?** *   1 point
◉ Commenting on each post of the blog
○ Fetching page of the blog for indexing
○ Adding ads to each page separately
○ None of these

---

6. **On LinkedIn, your profile most closely resembles a ___.** *   1 point
○ dating profile
○ digital scrapbook
◉ resume
○ photo gallery

7. **Which of the following is a Social Media Classification tool?** *   1 point
○ Social Publishing
○ Social Networking
○ Photo-based Social Networking
◉ All the above

---

8. **___ social network is considered the most popular for business to business marketing?** *   1 point
○ Facebook
○ Orkut
○ Ryze
◉ LinkedIn

9. **Which of the following resources are required to create a blog?** *   1 point
○ Domain Name
○ Hosting
○ content
◉ All the above

1.6 Rubrics

Parameter	Excellent	Very good	Good	Average
	20	15	10	5
Content and Creativity GA2	Content provides comprehensive insight, understanding, and reflective thought about the topic by building a focused argument around a specific issue or asking a new related question or making an oppositional statement supported by personal experience or related research.	Content provides moderate insight, understanding and reflective thought about the topic.	Content provides minimal insight, understanding and reflective thought about the topic.	Content shows no evidence of insight, understanding or reflective thought about the topic.
Text Layout, Use of Graphics and Multimedia GA5	Selects and includes high quality graphics and multimedia when appropriate to enhance the content's visual appeal and increase readability.	Selects and includes graphics and multimedia that are mostly high quality and enhance and clarify the content.	Selects and includes many low-quality graphics and multimedia which do not enhance the content.	Does not include any graphics, or uses only low-quality graphics and multimedia, which do not enhance the content.
Quality of Writing and Proofreading GA10,GA7	Written content is free of grammatical, spelling or punctuation errors. The style of writing facilitates communication.	Written content is largely free of grammatical, spelling or punctuation errors. The style of writing generally facilitates communication.	Written content includes some grammatical, spelling or punctuation errors that distract the reader.	Written content contains numerous grammatical, spelling or punctuation errors. The style of writing does not facilitate effective communication.
Citations	All images, media and text created by others display appropriate copyright permissions and accurate citations.	Most images, media or text created by others display appropriate copyright permissions and accurate, properly formatted citations.	Some of the images, media or text created by others does not display appropriate copyright permissions and does not include accurate, properly formatted citations.	No images, media or text created by others display appropriate copyright permissions and do not include accurate, properly formatted citations.
Publication of blog GA12	The blog is posted on student's host site.	The blog is posted on free blog site.	The blog is made into a web page.	The blog is not posted.
Likes, Shares and Comments on the blog GA06	The blog has all three of the following: Likes, shares and comments.	The blog has any two of the following: Likes, shares and comments.	The blog has received only one of the following: Likes, shares and comments.	The blog has not received either of the following: Likes, shares and comments.

Chapter 2: Prototype Development

2.1 Screenshots of Prototype Development

2.2 Tool Description

Figma is a web-based design tool that is widely used for creating high-quality user interface (UI) designs, wireframes, and prototypes. It is a collaborative design platform that enables multiple team members to work on a single design project simultaneously, making it a popular tool for remote teams.

One of the key benefits of Figma is its collaboration features. Designers and other team members can work together on a design project in real-time, making it easy to share feedback and iterate on designs quickly. Figma also supports commenting and version control, which makes it easy to keep track of changes and updates to a design project.

In the case of designing the prototype of a website, Figma offers a range of tools that can be used to create wireframes, visual designs, and interactive prototypes. Designers can use Figma's design components to create reusable UI elements such as buttons, forms, and menus, which can be quickly added to multiple pages of the website. Figma also has a feature called "Auto Layout" that enables designers to create responsive designs that adapt to different screen sizes and device types.

2.3 Screenshots of Quiz

System prototyping is becoming very popular, as it enables to understand customer ___ at an early stage of development.

1 point

○ Opinion

○ Development

◉ Requirements

○ All of the mentioned above

Clear selection

A prototyping model starts with ___ analysis.

1 point

○ Data

◉ Requirement

○ Dummy system

○ None of the mentioned above

Clear selection

In requirement analysis the ___ gather requirement and thinks about the solution of a problem.

1 point

◉ Experts

○ Executives

○ Employees

○ None of the mentioned above

Clear selection

In the ___ method, the prototype developed initially is incrementally refined on the basis of customer feedback till it finally gets accepted.

1 point

○ Rapid Throwaway Prototyping

◉ Evolutionary Prototyping

○ Incremental Prototyping

○ Extreme Prototyping

Clear selection

Prototyping is an attractive idea for ___ and large systems.   1 point

- ⦿ Complicated
- ○ Simple
- ○ Developed
- ○ All of the mentioned above

Clear selection

___ is the first and initial phase of the Prototype Model. This phase is used   1 point
to identify the problem.

- ⦿ Problem identification
- ○ Design
- ○ Testing
- ○ All of the mentioned above

Clear selection

In requirement analysis the ___ gather requirement and thinks about the   1 point
solution of a problem.

- ⦿ Experts
- ○ Executives
- ○ Employees
- ○ None of the mentioned above

Clear selection

In the ___ method, the prototype developed initially is incrementally refined   1 point
on the basis of customer feedback till it finally gets accepted.

- ○ Rapid Throwaway Prototyping
- ⦿ Evolutionary Prototyping
- ○ Incremental Prototyping
- ○ Extreme Prototyping

Clear selection

2.4 Rubrics

Parameter	Excellent (20)	Very Good (15)	Good (10)	Average (05)
Identifying Type of Prototype (Visual prototype/Functional prototype /Presentation prototype) (GA3)	Functional Prototype	Presentation prototype	Visual Prototype with little functions	Only Visual prototype
Advantages of prototyping GA4	Identification of innovative design thinking-based approach to make a prototype which is easy to implement and cost effective Interdisciplinary knowledge is applied	Identification of innovative approach to make a prototype	Try existing methods with slight modification to make a prototype	Apply existing methods and solution as it is to make a prototype
A step-by-step break-down of prototyping GA4,GA5	Apply latest Tools and technology	Apply latest Tools and technology learned in academics	Application of old techniques with slight modification	Application of old tools and techniques
The Spiral model GA8,GA10	Review and plan for next phase Use of own design	Develop next version of product Use of existing design	Objective determination and identify alternative solutions	Identify and resolve risks
Conclusion GA 11	(A)Clarify its purpose, function and appearance+(B)+(C) +(D) Use of project management tools and knowledge to conclude	(B) Improve user experience and marketability)+(C) +(D)	(C)Explore its manufacturability and make-up+(D)	(D) Solve problems before they occur

Chapter3: Mathematical Modelling and Infographics

3.1 Description of Mathematical model used

Extremely Randomized Trees, one of the algorithms in consideration during the technical review uses the concept of Entropy to decide which branch is to be expanded while generating a tree. It uses the following formula:

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

3.2 Infographics
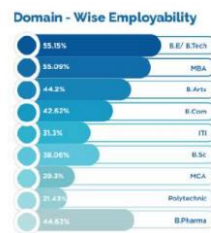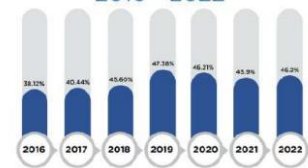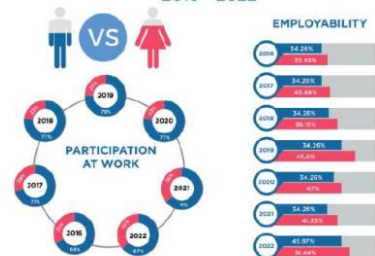
3.3 Screenshots of Quiz

1. Every mathematical model. *   1 point

○ Must be deterministic

○ Requires computer aid for its solution

◉ Represents data in numerical form

○ All of the above

2. A physical model is example of *   1 point

○ An iconic model

○ An analogue model

◉ A verbal model

○ A mathematical model

3. The qualitative approach to decision analysis relies on. *   1 point

○ Experience

○ Judgement

○ Intuition

◉ All of the above

4. The use of decision models. *   1 point

○ Is possible when the variables value is known

○ Reduces the scope of judgement & intuition known with certainty in decision-making

○ Require the use of computer software

◉ None of the above

---

5. This process of creating a mathematical model is referred to as _____ *   1 point

○ real world modeling

○ descriptive modeling

○ solving a problem

◉ scientific modeling

---

6. Operations research is the application of _____ methods to arrive   * 1 point
at the optimal Solutions to the problems

○ economical

◉ scientific

○ a and b both

○ artistic

---

7. In operations research, the ----------------------------are prepared for situations *   1 point

◉ mathematical models

○ physical models diagrammatic

○ diagrammatic models

○ All of the above

8. By constructing models, the problems in libraries increase and cannot be * 1 point
solved.

○ True

◉ False

9. In _____ Model there is a risk and uncertainty. *   1 point

○ Deterministic Models

◉ Probabilistic Models

○ Both a and b

○ None of the above

○ Other:

10. The main limitation of operations research is that it often ignores the * 1 point
human element in the production process.

◉ True

○ False

3.4 Rubrics

Marks	20	15	10	5
Parameter	**Excellent**	**Very Good**	**Good**	**Average**
Statement of Problem in real world	Students are able to identify the real-world problem that can be represent in specific mathematical model	Students are able to identify the real-world problem that can be represent in general mathematical model	Students are able to identify the real-world problem that can be represent in abstract mathematical model	Students are able to identify the real-world problem that can be represent in poor mathematical model
Technicality	Students are able to identify clear and specific mathematical variables (parameters)	Students are able to identify generalized mathematical variables (parameters) that will be directly or indirectly influenced	Students are able to identify abstract mathematical variables (parameters) that will be directly or indirectly influenced	Students are able to identify barely relevant mathematical variables (parameters)
Design and Formulation of Model	It completely enables the construction of a mathematical model using of tools required for mathematical modelling and simulation	It enables the construction of a mathematical model use of some tools required for mathematical modelling and simulation	It enables the construction of a mathematical model use of a few tools required for mathematical modelling and simulation	It enables the construction of a mathematical model without using tools required for mathematical modelling and simulation.
Presentation and Team Work	Student demonstrates full knowledge, answering all queries with explanations through the attractive infographics.	Student demonstrates partial knowledge, answering some of queries with explanations through the very good infographics.	Student is able to answer only basic queries utilization good infographics.	Student have poor knowledge; they are able to answer only few queries utilization poor infographics.
Evaluation of the Mathematical Model	Mathematical model is able to represent exact behaviour of real world problem identified and same is reflected using infographics	Mathematical model is able to represent similar behaviour of real world problem and same is reflected using infographics	Mathematical model is able to represent relevant behaviour of real world problem and same is reflected using good infographics	Mathematical model is able to represent irrelevant behaviour of real world problem and same is reflected using poor infographics

Chapter 4: Research Paper

4.1 Screenshot of Research Paper Presentation

Slide 1

IC-ICN - 2023

Title - Algorithms for Predicting Employability Potential: An overview

Authors :
Group no: C1
TE_COMP_C_13_Ashutosh Sharma
TE_COMP_C_18_Karthik Shetty
TE_COMP_C_23_Rajan Shukla
TE_COMP_C_36_Hariom Soni

Slide 2 — Index

1. Abstract
2. Introduction
3. Literature Survey
4. Algorithms Considered
5. Proposed model and Implementation
6. Prototype
7. Results and Discussion
8. Conclusion
9. References

Slide 3 — Abstract

- The demand for skilled professionals in engineering and technology is on rise
- Universities and colleges need to identify the employability of the students and the employment potential of the courses
- Using machine learning and data mining, universities can better understand the factors that impact employability
- By understanding these factors universities and colleges can make more informed decisions about their curriculum and student support services.

Slide 4 — Introduction

- Prediction and Classification have become fundamental tools with rapid growth of data
- Machine learning helps in making informed decisions by discovering the underlying patterns in data
- Success of these algorithm depends on the quality of data and efficiency of the algorithms
- Algorithms such as ERT, Apriori, LightGBM and Neural Networks have been discussed

Slide 5 — Literature Survey and Considered Algorithm

- Lang (2018) improved the Apriori algorithm based on matrix pruning and weight analysis.

Input data → Weight analysis → Matrix pruning

Simple matrix building using the transaction data:

- Transaction 1: {A, B, C, D}
- Transaction 2: {A, C, D, E}
- Transaction 3: {A, C, D, E}
- Transaction 4: {B, D, E, F}

	A	B	C	D	E	F
1	1	1	1	1	0	0
2	1	0	1	1	1	0
3	1	0	1	1	1	0
4	0	1	0	1	1	1

Slide 6 — Literature Survey and Considered Algorithm

- In a paper published by Sathyabhama Institute of Science and Technology, 2021 it was concluded that for same dataset CatBoost algorithm resulted in highest accuracy
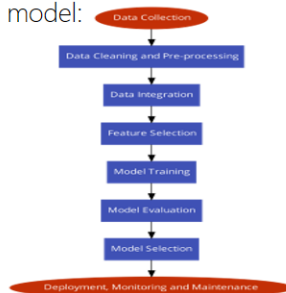
MODEL	CABOOST
MAE	0.03
MSE	0.001
R2 SCORE	0.89
ACCURACY	0.95

MODEL	LINEAR REGRESSION
MAE	0.04
MSE	0.003
R2 SCORE	0.84
ACCURACY	0.93

Slide 7 — Literature Review and Considered Algorithms

- Geurts, P., Ernst, D. & Wehenkel proposed a new algorithm named Extremely Randomized Trees, a variant of Random forest algorithm. This algorithm introduces an additional level of randomness to the tree construction.
- "LightGBM: A Highly Efficient Gradient Boosting Decision Tree" (2017) : This paper introduces a new decision tree based algorithm.
- Chen, Li, & Zhu in 2020 wrote a comparative study of gradient boosting frameworks. This paper compared GDBT, XGBoost and LightGBM.

Slide 8 — Proposed model:

Data Collection → Data Cleaning and Pre-processing → Data Integration → Feature Selection → Model Training → Model Evaluation → Model Selection → Deployment, Monitoring and Maintenance

Results and Discussion

- CatBoost algorithm resulted in higher accuracy compared to Linear Regression when effectively implemented
- The use of matrix pruning and weight analysis reduces the need for repeated scanning of transactional databases increasing the efficiency
- SHAP values measures each feature's contribution to model's prediction increasing its understandability
- ERT is better than RF and GBDT in various parameters like speed, memory requirements, accuracy of the result and changes in the dataset size

11

Conclusion

- Use of prediction and classification algorithms is crucial for achieving accurate and efficient results
- Apriori, ERT, LightGBM, and ANNs, have shown promising results in various benchmarks and applications
- It is essential to select appropriate algorithms based on specific data
- With rapid growth in data generated, prediction and classification algorithms will continue to play a significant role in various fields

12

References

- Mishra Shakti and Mishra Anjali, "Prediction of Admissions and Jobs in Technical Courses with Respect to Demographic Location Using Multi-Linear Regression Model"
- A.Sivasangari, V.Shivani, Y.Bindhu, Prediction Probability of Getting an Admission into a University using Machine Learning, Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
- Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).
- Lang, Zhenhong. 2018. "The improved Apriori algorithm based on matrix pruning and weight analysis." AIP Publishing. https://aip.scitation.org/doi/abs/10.1063/1.5033777.
- Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42, (2006). . https://doi.org/10.1007/s10994-006-6226-1
- Lang, Zhenhong. 2018. "The improved Apriori algorithm based on matrix pruning and weight analysis." AIP Publishing. https://aip.scitation.org/doi/abs/10.1063/1.5033777.

Thank You

Authors :
Group no: C1
TE_COMP_C_13_Ashutosh Sharma
TE_COMP_C_18_Karthik Shetty
TE_COMP_C_23_Rajan Shukla
TE_COMP_C_36_Hariom Soni

4.2 Research Paper

Algorithms for Predicting Employability Potential: Overview

Mr. Ashutosh Sharma
Department of Computer Engineering
Thakur College of Engineering & Technology
(Student)
Kandivali, Mumbai, India
sharma.ashutosh1010@gmail.com

Mr. Karthik Shetty
Department of Computer Engineering
Thakur College of Engineering & Technology
(Student)
Kandivali, Mumbai, India
Karthikshetty270902@gmail.com

Mr. Rajan Shukla
Department of Computer Engineering
Thakur College of Engineering & Technology
(Student)
Kandivali, Mumbai, India
rajan93992@gmail.com

Mr. Hariom Soni
Department of Computer Engineering
Thakur College of Engineering & Technology
(Student)
Kandivali, Mumbai, India
pratham.soni25@gmail.com

Dr. Sheetal Rathi
Professor,
Department of Computer Engineering
Thakur College of Engineering & Technology
Kandivali, Mumbai, India
sheetal.rathi@thakureducation.org

Abstract- The demand for skilled professionals in the engineering and technology industries in India has been on the rise, leading to an increase in competition for admission to technical courses and job opportunities. For the same purpose universities and colleges need to identify the employability of the students and the employment potential of different engineering and technical courses to increase the chances of students getting better jobs. This paper aims to investigate different machine learning algorithms and data mining techniques which are used to analyze the educational data available and predict the employment potential of different technical courses. By utilizing these techniques, universities and colleges can better understand the factors that impact employability, and make data-driven decisions about which courses to offer and which skills to prioritize in their curriculum. The findings of this paper will provide valuable insights into the usage of machine learning to predict the employment potential of technical courses, and help universities and colleges make more informed decisions about their curriculum and student support services.

Keywords: Employability, employment potential, machine learning algorithms, data mining, data-driven decisions

I. INTRODUCTION

In recent years, with the rapid growth of data, prediction classification has become a fundamental tool for various fields of study. It plays a vital role in discovering the underlying patterns in the data and making informed decisions. However, the success of prediction classification is heavily dependent on the quality of the data and the efficiency of the algorithms used. To achieve accurate and efficient prediction classification, it is essential to have clean and relevant data and to use appropriate algorithms that can handle the complexity of the data. In this review paper, we explore some of the most promising algorithms for prediction classification, including the improved Apriori algorithm, Extremely

Randomized Trees (ERT), LightGBM, and Artificial Neural Networks.

II. LITERATURE REVIEW

The literature review comprises papers that provide insights into prediction classification and machine learning techniques in educational data analytics. The review will be presented in a paragraph format, highlighting the research topic, methodology and key findings of each paper.

A recent study, "Prediction of Admissions and Jobs in Technical Courses with Respect to Demographic Location Using Multi-Linear Regression Model" used a predictive model to examine the impact of demographic location on admission and job opportunities in the engineering and technology industries. They proposed a multiple linear regression-based analysis that predicts the number of admission and jobs and using this data the employment potential of technical courses will be determined. Their model predicted using the features college rank, cutoff rank, placement percentage, accreditation and the number of companies visiting, the number of partner companies, quality of placement training, and average salary package for admission and placement respectively. [2]

The model proposed by the Sathyabama Institute of Science and Technology Chennai, India uses Machine Learning algorithms such as Linear regression, random forest algorithms, and CatBoost algorithms independently to analyze the graduate student data to predict admission chances in universities worldwide. For the data set, they considered the academics score as well as the scores of the competitive exams were considered as important. Because the admission cutoffs are dynamic and change every year assumptions were made to create a base cutoff for the known universities based on the admission history. [3] Similarly, assumptions based upon the competition in the history the students applied and students got selected the algorithm was provided with a full set of concrete conditions. Through the model, it was concluded that for the same data set the CatBoost algorithm resulted in the highest accuracy in the result. The following table shows the accuracy of the Linear regression algorithm and CatBoost algorithm. [3]

TABLE: 2.2. Linear Regression result

Model	Linear Regression
MAE	0.04
MSE	0.003
R2 Score	0.84
Accuracy	0.93

TABLE: 2.1. CatBoost result

Model	CatBoost
MAE	0.03
MSE	0.001
R2 Score	0.89
Accuracy	0.95
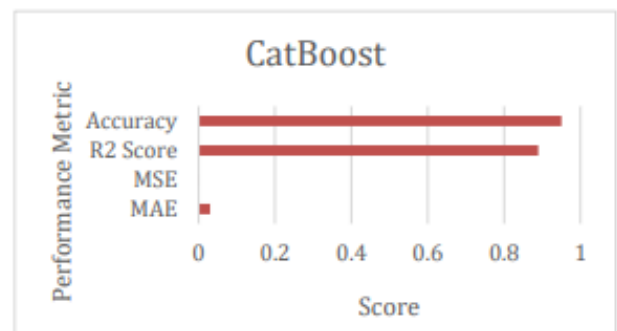
Based on the survey results, [4] CatBoost stands out as a promising machine learning tool for Big Data applications. It is automatic handling of categorical data, along with its superior performance in comparison to other gradient-boosting decision trees (GBDT) algorithms, makes it a suitable choice for many applications. Specifically, it is well-suited for datasets that are heterogeneous and contain categorical features. However, the survey [4] also found that the performance of CatBoost is sensitive to hyper-parameter settings, including the maximum number of iterations, the maximum depth of decision trees, and the maximum number of combinations of categorical features.

A 2006 study under the name "Extremely randomized trees" by Geurts P., Ernst D. & Wehenkel [6] proposes

an algorithm called Extra-Trees that uses a combination of random subspace and totally random cut-point selection to improve the performance of decision tree ensembles. The optimal value of one of the algorithm's main parameters, K, depends on the problem specifics, but default values are near-optimal for most datasets. The algorithm reduces variance and increases bias, and its models are continuous and piecewise multi-linear. Extra-Trees provide near-optimal accuracy and good computational complexity for high-dimensional problems, while totally randomized trees are faster and provide independent tree structures. The authors suggest potential improvements for handling other types of attributes and reducing bias.

"Understanding random forests: From theory to practice." by Louppe, Gilles presented that the theory of random forests involves constructing multiple decision trees by selecting random subsets of the data and features, and combining their predictions to make a final prediction. [7] This can be done using various programming libraries, such as scikit-learn in Python. The scikit-learn library has a Random Forest Classifier class that can be used to create and train a random forest model, as well as make predictions on new data. The hyperparameters, such as the number of trees and the maximum depth of each tree, can be tuned to optimize the model's performance.

The paper "Consistent individualized feature attribution for tree ensembles" by Lundberg, Erion, and Lee presents a novel method for computing feature importance scores in tree ensembles that addresses the shortcomings of existing methods. [8] The authors introduce a new method, called SHAP (SHapley Additive exPlanations), which relies on a game-theoretic approach to compute feature importance scores that are both accurate and consistent across different prediction outcomes. The paper provides extensive experimental results that demonstrate the effectiveness of the SHAP method in a range of applications, including healthcare, finance, and computer vision. The authors also provide an open-source Python library that implements the SHAP method, making it widely accessible to practitioners.

The paper by Chen, Li, and Zhu (2020) presents a comparative study of five gradient boosting frameworks, namely XGBoost, LightGBM, CatBoost, NGBoost, and Sklearn-GBM. [13] The study evaluates these frameworks based on their accuracy, speed, and memory usage for a variety of datasets and learning tasks. The authors found that the newer frameworks, such as LightGBM and CatBoost, generally outperformed the more established frameworks, XGBoost and Sklearn-GBM, in terms of both accuracy and speed. They also found that NGBoost, which is based on probabilistic modeling, is promising for tasks that require uncertainty quantification. Overall, the study provides insights into the strengths and weaknesses of different gradient boosting frameworks and can be used to inform the selection of an appropriate framework for a given task.

In the paper "A Highly Efficient Gradient Boosting Decision Tree", Ke et al. introduced LightGBM, a gradient-boosting framework that uses a novel technique called Gradient-based One-Side Sampling (GOSS) to reduce the number of data instances in the training set. [14] The authors claim that this approach achieves a significant reduction in training time and memory usage compared to other gradient-boosting frameworks, while still maintaining high accuracy. Ke et al. provide experimental results showing that LightGBM outperforms other popular gradient boosting frameworks on several datasets, including the Microsoft Learning to Rank dataset and the Higgs boson dataset.

III. ALGORITHMS

Before working with the algorithm, the main criterion is to have a clean set of data. The values that the data represents must be clear in understanding and look relevant to the data analyzer. The finding process must be so reliable that any relevant data must not be obscure in any way. It is common to use the Knowledge discovery in database (KDD) process to identify valid, novel, useful, and understandable patterns from large and complex datasets.[1] This method consists of sub-process including Data Selection, Data preparation, and transformation, Data Mining, and Evaluation. Following this comes to the process of Test and Training the ascertain data from the mining process.

1. Improved Apriori algorithm based on matrix pruning and weight analysis

The Apriori algorithm's flaws are identified as the repeated scanning of databases leading to excessive consumption of space and time and the dependence on

minimum support, resulting in suboptimal association rules. The proposed algorithm addresses these flaws through matrix pruning and weight analysis. This approach reduces the number of repeated scans and improves the efficiency of data correlation mining, making it a more effective algorithm for association rule mining. Association rules require that all non-empty subsets of frequent item sets must also be frequent, and any non-frequent item set cannot be a subset of frequent item sets. The algorithm works by scanning the transactional database and generating a Boolean transaction matrix. The support counting of each item set is compared to the minimum support count to generate frequent 1-item sets. The weight of each item and transaction is then calculated using formulas, and the support counting of item weight is calculated to gain information about the item in each transaction. The weighted value of candidate k-item sets is then compared to the minimal support to generate frequent k-item sets, which are pruned by intersecting with L'k. The algorithm continues to generate candidate (k+1)-item sets until all frequent item sets are formed. This approach reduces the need for repeated scanning of the transactional database, which improves the efficiency of the algorithm.

Table 3.1 Comparison of the Four Algorithms Runtime Performance [5]
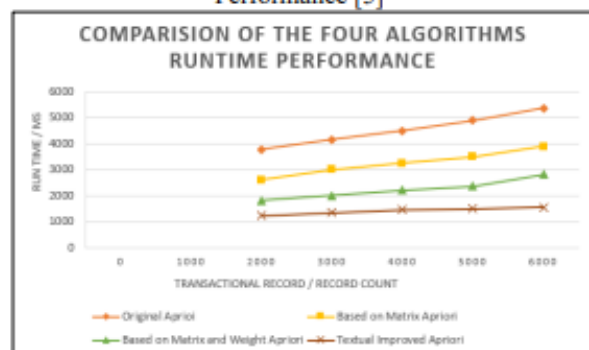
Overall, the improved Apriori algorithm has significant theoretical and practical value for data mining in the era of big data. [5]

2. Extremely Randomized Trees (ERT) using SHAP (SHapley Additive exPlanations) for feature importance

A popular ensemble learning approach for regression and classification is Random Forest. It applies the bagging concept, which involves randomly dividing the training data into a large number of subsets and fitting a decision tree to each subset. By merging the outcomes of various decision trees, the Random Forest approach can generate forecasts that are more accurate and dependable.

Extremely Randomized Trees (ERT) is an extension of the RF algorithm that uses a different method for selecting the splits. ERT selects splitting point at random within a predefined range, rather than searching for the optimal split point based on the impurity measure. The use of random subsamples and random feature selection helps to reduce the variance of the model and prevent overfitting [7]. ERT is also robust against the changing size and noise in the data. The table below compares the performance of ERT with other popular ensemble learning algorithms. The comparison uses iris, wine, glass and yeast dataset.

Table 3.2 Comparison of ERT with other ensemble learning algorithms [6]

Algorithm	Classification Accuracy	Regression RMSE
ERT	95.9%, 98.1%, 72.4%, 62%	0.041, 0.042, 0.065, 0.174
Random Forest	94.7%, 97.2%, 70.7%, 58.3%	0.049, 0.055, 0.068, 0.197
AdaBoost	95.8%, 97.2%, 69.1%, 50.2%	0.047, 0.064, 0.071, 0.262
Bagging	94.7%, 97.2%, 69.1%, 58.1%	0.052, 0.058, 0.072, 0.198

Direct impurity is a generally used measure of feature importance, however it does not account for the contribution of each feature. SHAP values, in contrast, more accurately reflect human intuition, consistently assign feature importance, and better retrieve influential features. We transform SHAP values in tree ensembles into a workable replacement for previous feature significance methods [8] by introducing the first polynomial time approach for them. The difference between the expected output of the model with the feature included and the predicted output of the model with the feature deleted, averaged over all potential subsets of features, is the SHAP value for a specific feature. In order to account for the interactions between the features, it measures each feature's contribution to the model's prediction.

3. LightGBM

LightGBM is an open-source gradient boosting framework that is designed to efficiently handle datasets and deliver accurate predictions. A histogram-based approach is used by LightGBM to split data and generate decision trees. This makes it efficient when used for computing split points and reduces memory usage [15]. Generally speaking, LightGBM and XGBoost are recognised to be quicker and more effective than one another, especially when working with smaller datasets. The training process is accelerated by LightGBM using a histogram-based method, which can be very helpful when working with datasets that include a limited number of unique feature values. As opposed to LightGBM, XGBoost is a more conventional gradient boosting technique that makes use of pre-sorted sparse data structures and has the potential to be memory-efficient.

Table 3.3 Comparison of XGBoost, LightGBM and CatBoost on various parameters.[9] [10] [11] [12]

| Feature	XGBoost	LightGBM	CatBoost
Tree construction	Level-wise	Leaf-wise	Leaf-wise
Histogram Based	No	Yes	Yes
Categorical features	One-hot encoding	Native support	Native support
GPU acceleration	Yes	Yes	Yes
Parallel learning	Yes	Yes	Yes
Multi-class	SoftMax	One-vs-all	Ordered boosting
Missing values	Yes	Yes	Yes
Speed	Medium	Fast	Medium-Fast
Memory Usage	Medium	Low	Medium-Low
Regularization	L1,L2	L1, L2	L1,L2, Random
Embedding Support	No	Yes	Yes

4. NEURAL NETWORKS

Artificial neural networks (ANNs) are a family of statistical learning models inspired by biological neural networks, and they are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. ANNs are presented as systems of interconnected "neurons" that exchange messages with each other. The connections between neurons have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. Backpropagation is a common method of training artificial neural networks used in conjunction with an optimization method such as gradient descent. The method calculates the gradient of a loss function with respect to all the weights in the network and updates the weights to minimize the loss function. Backpropagation is a supervised learning method that requires a known, desired output for each input value in order to calculate the loss function gradient. The backpropagation learning algorithm can be divided into two phases: propagation and weight update. During propagation, the input is forwarded through the neural network to generate the output activations, and then the output activations are backward propagated through the neural network using the training pattern target to generate the deltas of all output and hidden neurons. During weight update, for each weight-synapse, the output delta and input activation are multiplied to get the gradient of the weight, and a ratio of the gradient is subtracted from the weight.[8] This process is repeated until the performance of the network is satisfactory. The learning rate, or the ratio of the gradient, influences the speed and quality of learning. A higher learning rate makes the neuron train faster, while a lower learning rate makes the training more accurate. Five steps were considered for the implementations including the- Designing a Neural Network, Creating/Gathering training Data sets; Thermometer Encoding; Training the Neural Network; Testing the Neural Network against Random data [8]

For the designing step, A neural network has been designed using Backpropagation algorithm which is a supervised learning algorithm. Backpropagation works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally. [13] The next few steps include finding the appropriate data set and then before feeding to the neural network, the data is then encoded into binary format to

make it easier to work with. Supervised training method is both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. A limiting value is set so that the training stops reaching the limit. A learning rate is initialized which is set to either low (if the dataset is variable and large) or high (if the dataset is small). Once the training part is completed, the network is given random sets of data as input. The network will classify the given input into the appropriate group/class based on how effectively the network is trained. The network will be 80- 90% accurate in classifying the random sets of data. [13]

Fig 3.4. Flowchart for data mining using the Backpropagation method. [8]
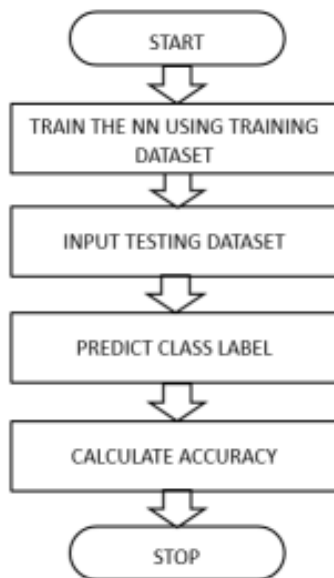
IV.CONCLUSION

In conclusion, the use of prediction classification algorithms is crucial for achieving accurate and efficient results when working with big data. The algorithms discussed in this literature review, including the improved Apriori algorithm, ERT, LightGBM, and ANNs, have shown promising results in various benchmarks and applications. However, it is essential to select the appropriate algorithm based on the specific data and problem at hand. With the rapid growth of big data, prediction classification algorithms will continue to play a significant role in various fields, including healthcare, finance, and marketing.

VI.REFERENCES:

[1] Rojanavasu, P. (2019). Educational Data Analytics using Association Rule Mining and Classification. 2019 Joint International Conference on Digital Arts, Media, and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer, and Telecommunications Engineering

[2] Mishra Shakti and Mishra Anjali, "Prediction of Admissions and Jobs in Technical Courses with Respect to Demographic Location Using Multi-Linear Regression Model"

[3] A.Sivasangari, V.Shivani, Y.Bindhu, Prediction Probability of Getting an Admission into a University using Machine Learning, Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)

[4] Hancock, J.T., Khoshgoftaar, T.M. CatBoost for big data: an interdisciplinary review. J Big Data 7, 94 November 2020.

[5] Lang, Zhenhong. 2018. "The improved Apriori algorithm based on matrix pruning and weight analysis."AIPPublishing.
https://aip.scitation.org/doi/abs/10.1063/1.5033777.

[6] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42, (2006). .
https://doi.org/10.1007/s10994-006-6226-1

[7] Louppe, Gilles. "Understanding random forests: From theory to practice." arXiv preprint arXiv:1407.7502 (2014).

[8] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).

[9] Chen, H., Li, J., & Zhu, X. (2020). A comparative study of gradient boosting frameworks. arXiv preprint arXiv:2006.11287.

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[11] CatBoost. (n.d.). CatBoost - open-source gradient boosting library. Retrieved from https://catboost.ai/

[12] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 6638-6648).

[13] Lu H, Setiono R, Liu H. Effective data mining using neural networks. IEEE Trans Knowl Data Eng. 2016;8(6):957-61.

[14] Bekir Karlik and A. Vehbi Olgac International Journal of Artificial Intelligence And Expert Systems (IJAE), Volume (1): Issue (4) 111 Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks J. D. Poston and W. D. Horne, "Discontiguous OFDM considerations for dynamic spectrum access in idel TV channels," in Proc. IEEE DySPAN, 2005.

[15] "LightGBM: A Highly Efficient Gradient Boosting Decision Tree" (2017) by G. Ke et al

4.3 Screenshots of Quiz

1. Research is *   1 point

○ Searching again and again

○ Finding solution to any problem

◉ Working in a scientific way to search for truth of any problem

○ None of above

2. Which of the following is the first step in starting the research process? *   1 point

○ Searching sources of information to locate problem.

○ Survey of related literature

◉ Identification of problem

○ Searching for solutions to the problem

3. Reading center's, story hours, exhibitions and reading to literature are   * 1 point
forms of ...

○ Extension service

◉ Service of a public library

○ Both

○ None

4. An appropriate source to find out descriptive information is.................. . *   1 point

○ Bibliography

○ Directory

◉ Encyclopedia

○ Dictionary

---

5. A literature review is best described as: *   1 point

○ A list of relevant articles and other published material you have read about your topic, describing the content of each source

○ An internet search for articles describing research relevant to your topic criticising the methodology and reliability of the findings

◉ An evaluative overview of what is known about a topic, based on published research and theoretical accounts, which serves as a basis for future research or policy decisions

○ An essay looking at the theoretical background to your research study

---

6. A research intends to explore the result of possible factors for the organization of effective mid-day meal interventions. Which research method will be most appropriate for this study?   * 1 point

◉ Descriptive survey method

○ Historical method

○ Ex-post facto method

○ Experimental method

---

7. Which of the following options are the main tasks of research in modern society?   * 1 point

○ To learn new things

○ To keep pace with the advancement in knowledge

◉ To systematically examine and critically analyze the investigations/sources with the objective

○ All of the above

8. The conclusions/findings of which type of research cannot be generalized to other situations? * 1 point

○ Casual Comparative Research

○ Historical Research

○ Descriptive Research

◉ Experimental Research

9. What is the major attribute of Correlation Analysis? * 1 point

○ Association among variables

○ Difference among variables

○ Regression among variables

◉ Variations among variables

10. How to judge the depth of any research * 1 point

○ Someone who is a hard worker

○ Possesses post-graduation degree

◉ Has studied research methodology

○ Possesses thinking and reasoning ability

4.4 Rubrics

Marks	20	15	10	5
Parameter	Excellent	Very Good	Good	Average
Statement of Problem in real world	Students are able to identify the real-world problem that can be represent in specific mathematical model	Students are able to identify the real-world problem that can be represent in general mathematical model	Students are able to identify the real-world problem that can be represent in abstract mathematical model	Students are able to identify the real-world problem that can be represent in poor mathematical model
Technicality	Students are able to identify clear and specific mathematical variables (parameters)	Students are able to identify generalized mathematical variables (parameters) that will be directly or indirectly influenced	Students are able to identify abstract mathematical variables (parameters) that will be directly or indirectly influenced	Students are able to identify barely relevant mathematical variables (parameters)
Design and Formulation of Model	It completely enables the construction of a mathematical model using of tools required for mathematical modelling and simulation	It enables the construction of a mathematical model use of some tools required for mathematical modelling and simulation	It enables the construction of a mathematical model use of a few tools required for mathematical modelling and simulation	It enables the construction of a mathematical model without using tools required for mathematical modelling and simulation.
Presentation and Team Work	Student demonstrates full knowledge, answering all queries with explanations through the attractive infographics.	Students demonstrates partial knowledge, answering some of queries with explanations through the very good infographics.	Students are able to answer only basic queries utilization good infographics.	Student have poor knowledge; they are able to answer only a few queries utilization poor infographics.
Evaluation of the Mathematical Model	Mathematical model is able to represent exact behaviour of real world problem identified and same is reflected using infographics	Mathematical model is able to represent similar behaviour of real world problem and same is reflected using infographics	Mathematical model is able to represent relevant behaviour of real world problem and same is reflected using good infographics	Mathematical model is able to represent irrelevant behaviour of real world problem and same is reflected using poor infographics