****

**A**

**PROJECT REPORT**

**ON**

**“STUDENT BEHAVIOUR ANALYSIS USING UNSUPERVISED APPROACH”**

**Submitted in partial fulfilment of the Award of the degree of**

**MASTER OF COMPUTER SCIENCE**

**By**

**TURPU MEGHANA**

**117220504036**

**MSc (Computer Science) II Year IV Semester**

**Under the Guidance of**

**Mrs.C.Jyothi Sree**

**Assistant professor , Department of Computer Science**

****

**[Estd: 1954]**

**DEPARTMENT OF COMPUTER SCIENCE (PG)**

**RBVRR WOMEN’S COLLEGE(Autonomous)**

**Affiliated to Osmania University**

**Reaccredited by NAAC with ‘B++’ Grade**

**Hyderabad-500027, Telangana State**

**[2020-2022]**

**ORGANIZATION PROFILE**

****

**COMPANY OVERVIEW LEADER STATEMENT**

*“Innovation distinguishes between a leader and a follower”. Do not follow where the path may lead. Go instead where there is no path and leave a trail.”*

Pantech focuses on providing Technical Training & Design Solutions. Our Technical Training Domains includes Machine Learning, Deep Learning, IoT, Embedded Systems, Robotics, Android Development, Open CV, Python, VLSI, Image Processing, Network Security, Power Electronics, Power Systems, Renewable Energy, Big data and Data Science. The design and implementation of these technologies, however, is typically done independently. As a result, each one operates on a separate platform.

**Vision**

"To Gain Global Leadership in providing Technological Solutions through Sustained Innovation."

**Mission:**

“Our mission is to provide the highest level of information and professional services for members and their clients worldwide” .

**Core Values :**

Our 6 core values are derived from our stewardship quality.

• **Integrity** - Honesty in how we deal with our clients, each other and with the world.

• **Candor** - Be open and upfront in all our conversations. Keep clients updated on the real situation. Deal with situations early; avoid last minute surprises.

• **Service** - Seek to empower and enable our clients. Consider ourselves successful not when we deliver our client’s final product but when the product is launched and meets success.

• **Kindness** - Go the extra mile. Speak the truth with grace. Deliver more than is expected or promised.

**• Competence** - Benchmark with the best in the business. Try new and better things. Never rest on laurels. Move out of comfort zones. Keep suggesting new things. Seek to know more.

• **Growth** - Success is a journey, not a destination. Seek to multiply/increase what we have - wealth, skills, influence, and our client’s business.

**History & Objectives:**

Pantech Group of Companies founded 17 years ago, has been complementing it’s academic activity at undergraduate, graduate and doctoral levels. Salute its dedicated R&D team on these seminal accomplishments, including setting up its own campus across Major Cities in India, Establishment of a dedicated Team of Professionals for Article Publications in Various Journals and much more with in such a short span of time Pantech has been successful in inculcating technical education cum services for more than a decade. It is with that leveraging experience, technical zenith and updations in the technology on industrial applications and with a intention to make the process of Digital E Learning simplified, PANTECH E LEARNING was launched. The prime Intention was to make the students cum staffs or the participants in general to become technology savvy, creative and be an active contributor to the development of technology solutions. Free access to technical materials and sharing of knowledge and experience gained throughout the 17 years of unparalled journey in the technical services sector was the prime objective behind this initiative.

**Main Objectives:**

• Commitment to quality, constant innovation & exceeding expectations.

• Always in pursuit of technological updations & transfer of knowledge to the staff and student community.

• Pantech R&D covers everything from basic engineering to advanced research in electronics, electrical & software domains.

**Our Services:**

**➢ R&D Design**

• Customized Product Design and Manufacture for Customers.

• Projects Design & Installation of Project Proto types in colleges.

• Design, Testing & Manufacturing of Lab Equipments to the Engineering Colleges.

• Manufacturing & sales of Project Kits.

**➢ Online / Offline Training Services**

• On Campus Technical Trainings @ College premises.

• In House Trainings – Short Term & Long-Term course Offerings Course with Placements.

• Two / Three Days Workshop Programs & Hands on Sessions.

• Faculty Development programs (FDP) & STTPs.

• Internships / In Plant Trainings.

• Workshop and the Internship for both the faculties and the students.

**➢ Project Solutions**

• Guidance for Projects for Final Years / Pre Final Students of Engineering & Diploma Stream.

• Research Project Solutions to Colleges as a whole & to corporates Conducting Faculty Development Programs & Hands on Sessions in the project domains.

• Demo of the Lab Kits to the Staff of Engineering Colleges.

• Working on global market through ecommerce.

**Our Associations - Education & Training**

• **IETE- MUMBAI** - The Institution of Electronics and Telecommunication Engineers (IETE) is India's leading recognised professional society devoted to the advancement of Science and Technology of Electronics, Telecommunication & IT.The IETE is the National Apex Professional body of Electronics and Telecommunication, Computer Science and IT Professionals and is a GOVT of INDIA Certified Scientific Institute Research Organisation ( SIRO )

• **NITK-STEP- NITK** - Science & Technology Entrepreneurs' Park (NITK-STEP), formerly KREC-STEP, was established on 31st August 1994 in KREC campus. It is situated in an independent location in the complex of NITK about 23 K.M. from Mangalore city, adjacent to N.H. 66 and adjoining the Arabian Sea beach. Out of 16500 sft. built up area, 8500 sq.ft. consists of Entrepreneurs' Block and the remaining area comprises of computer laboratory, conference hall and administrative block. It is an entrepreneurial driven self-sustaining body.

• **IIPM - KANSBAHAL** - Indian Institute for Production Management, popularly known as IIPM Kansbahal is a distinguish corporate promoted Techno-Managerial Institute of the country, which has always set a new benchmarking in the field of Training, Education, Research and CBM which is only replicated by others. It is still leading in Training, Education and Research & CBM sectors through its unmatched expertise, experience for over three decades and dedicated team players. It's in early 80's when business and industry had just realized that “Training is not a cost, it is an investment” and education sector was feeling the germane of global standard.

|  |
| --- |
| Ph. No :27564660  Fax No.:040-27564296  E-mail: rbvrrwc@rediffmail.com  Website: www.rbvrrwomenscollege.net. |

****

**[Estd 1954]**

**RAJA BAHADUR VENKATA RAMA REDDY**

**WOMEN’S COLLEGE**

**(AUTONOMOUS)**

**(Affiliated to Osmania University, Reaccredited by NAAC with ‘B++’ Grade)**

**NARAYANGUDA, HYDERABAD -27 T.S.**

**CERTIFICATE**

This is to certify that the project report entitled **“STUDENT BEHAVIOUR ANALYSIS USING UNSUPERVISED APPROACH”** being submitted by

**Ms. Turpu Meghana** bearing the Hall Ticket No: **117220504036** in partial fulfilment for the award of the degree of Master in Computer Science to the Osmania University, Hyderabad is a record of benefited work carried out under my guidance and supervision.

**Mrs .C.Jyothi Sree**

Assistant Professor

Department of Computer Science

RBVRR Women’s College(Autonomous)

Narayanguda , Hyderabad**.**

|  |
| --- |
| Ph. No :27564660  Fax No.:040-27564296  E-mail: rbvrrwc@rediffmail.com  Website: www.rbvrrwomenscollege.net. |

****

**[Estd 1954]**

**RAJA BAHADUR VENKATA RAMA REDDY WOMEN’S COLLEGE**

**(AUTONOMOUS)**

**(Affiliated to Osmania University, Reaccredited by NAAC with ‘B++’ Grade )**

**NARAYANGUDA, HYDERABAD -27 T.S**

**CERTIFICATE**

This is to certify that **Ms.Turpu Meghana** bearing the Hall Ticket No: **117220504036** studying MSC (Computer Science) II Year IV Semester has successfully completed Project work titled “**STUDENT BEHAVIOUR ANALYSIS USING UNSUPERVISED APPROACH IN PANTECH”** organisation for the academic year **2021-2022**.

**Internal Examiner External Examiner**

**DECLARATION**

I hereby declare that the dissertation entitled “**STUDENT BEHAVIOUR ANALYSIS USING UNSUPERVISED APPROACH”** submitted for the Master’s Degree in my original work and the dissertation has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles.

**Place:** Hyderabad **Name:** Turpu Meghana

**Date: Hall Ticket No:** 117220504036.

**ACKNOWLEDGEMENT**

Its give me immense pleasure to express my deep sense of gratitude to our guide **Ms.Sreelekha** Madam developer has been kind enough to guide us in planning of this project, encouragement timely guidance and giving valuable advices. I am thankful to staff of **PANTECH e-Learning Organisation** in execution of the project.

I thank **Mrs.T.VAMSHI MOHANA,Associate Professor,Head Department of computer Science** and my internal guide **Mrs.C.JYOTHI SREE**,**Assistant Proforessor,Department of Computer Science,RBVRR Women’s College(Autonomous),Narayanguda,Hyderabad.**

This project would not have completed without her enormous help and worthy experience. Whenever I was in need, she was there behind me. We are obliged and grateful to my guide for her sagacious guidance in all respects and for her valuable in each and every stage of this work, which helped me in completing this work successfully.

**Name:** Turpu Meghana

**Hall Ticket No:** 117220504036.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Name Of The Content** | **Page No** |
|  | **ABSTRACT** |  |
| 1 | **INTRODUCTION** | 1 |
|  | 1.1 Problem Statement | 1 |
|  | 1.2 Motivation | 1 |
|  | 1.3 Objective | 2 |
| 2 | **LITERATURE SURVEY** |  |
|  | 2.1 A Predictive System Informed by Students’ Similar Behaviour | 3 |
|  | 2.2 Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course | 3-4 |
|  | 2.3 Learning the Number of Clusters in Self Organizing Map | 4 |
|  | 2.4 The impact of online learning on students’ course outcomes | 4-5 |
|  | 2.5 Comparing Student Performance: Online versus Blended versus Face-to-Face | 5-6 |
|  | 2.6 Existing System | 6 |
| 3 | **SYSTEM ANALYSIS & REQUIREMENTS** |  |
|  | 3.1 Hardware Requirements | 7 |
|  | 3.2 Software Requirements | 7 |
|  | 3.3 Proposed System | 7 |
|  | 3.4 Functions & Methods | 8-9 |
| 4 | **SOFTWARE DESIGN** |  |
|  | 4.1 System Architecture Diagram | 10 |
|  | 4.2 Data Flow Diagram | 11-12 |
|  | 4.3 UML Diagram | 13 |
|  | 4.3.1 Use Case Diagram | 13-14 |
|  | 4.3.2 Activity Diagram | 14-15 |
| 5 | **INTRODUCTION TO TECHNOLOGIES** |  |
|  | 5.1 Logistic Regression | 16 |
|  | 5.2 K Nearest Classifier | 16 |
|  | 5.3 Machine Learning Overview | 17-18 |
|  | 5.4 Python Overview | 18 |
|  | 5.5 Tensor Flow | 19 |
|  | 5.6 Anaconda Navigator | 19 |
|  | 5.7 Numpy | 20 |
|  | 5.8 Html | 20 |
| 6 | **IMPLEMENTATION** |  |
|  | 6.1 Pseudocode | 21 |
|  | 6.1.1 Code | 21 |
|  | 6.1.2 EDA | 21-22 |
|  | 6.1.3 Pre-processing | 23 |
|  | 6.1.4Univariate Selection | 24 |
|  | 6.1.5 Plot the heat map | 24 |
|  | 6.1.6 Splitting and Classification | 24 |
|  | 6.1.7 Logistic Regression | 24 |
| 7 | **EXPERIMENTAL RESULTS** | 25-37 |
| 8 | **CONCULSION & FUTURE SCOPE** |  |
|  | 8.1 Conclusion | 38 |
|  | 8.2 Future scope | 38 |
| 9 | **REFERENCES** | 39-41 |
|  | List of Figures | i |
|  | List of Tables | ii |
|  | List of Charts | iii |
|  | Acronyms and Abbreviations | iv |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Name of the Figure** | **Page No** |
| 4.1 | Architecture Diagram | 10 |
| 4.2 | Data Flow Diagram | 11 -12 |
| 4.3.1 | Use Case Diagram | 14 |
| 4.3.2 | Activity Diagram | 15 |
| 7.3 | System Screenshots | 34-37 |
| 7.3.1 | Step-1 | 34-35 |
| 7.3.2 | Step-2 | 35 |
| 7.3.3 | Step-3 | 36-37 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Name of the Table** | **Page No** |
| 7.1.1 | Discribes the dataset | 25 |
| 7.1.2 | Machine learning | 26-27 |

**LIST OF CHARTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Name of the Chart** | **Page No** |
| 7.2.1 | Pivot Table | 28 |
| 7.2.2 | Simple Plots Correlation | 29-30 |
| 7.2.3 | Complex Plots | 31-32 |
| 7.2.4 | Machine Learning | 33 |

i

**ACROYNYMS & ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Acronym** | **Abbreviation** |
| 1 | SOM | Self-Organizing Map |
| 2 | RF | Random Forest |
| 3 | KNN | K-Nearest Neighbours |
| 4 | LR | Logistic Regression |
| 5 | SVM | Support Vector Machine |
| 6 | DT | Decision Tree |
| 7 | ML | Machine Learning |
| 8 | UML | Unified Modelling Language |
| 9 | DFD | Data Flow Diagram |
| 10 | LMSs | Online Learning Management System |
| 11 | UNIR | Universidad Internacional de La Rioja |
| 12 | MOOC | Massive Open Online Course |

ii

**ABSTACT**

An accurate analysis of user behaviour in online learning environments is a useful means of early follow up of students, so that they can be better supported to improve their performance and achieve the expected competences. However, that task becomes challenging due to the massive data that learning management systems store and categories. With the COVID-19 pandemic still on-going, face-to-face learning settings have migrate into online and blended ones, meaning an increase of online students and teachers in need for a tailored and effective support to their needs. A novel unsupervised clustering technique based on the Self-Organizing Map (SOM) artificial neural network model is used in this research to analysis 1,709,189 records of online students enrolled from 2015 to 2019 at Universidad Internacional de La Rioja (UNIR), a fully online Higher Education institution. SOM performs a precise and diverse user clustering based on those records. Results highlight that specific clusters are linked to the intake average profile at the university, with a clear relation between user interaction and a higher performance. Further, results show that, out of a targeted desk research compared to the analysis in this paper, face-to-face and online settings are connected through the methodological approach beyond the technology-based environment, which presents a similar behaviour in both contexts.

**Keywords**: Artificial neural networks, data science applications in education, distance education and online learning, pattern analysis, self-organizing map (SOM), student behaviour, unsupervised learning.

**CHAPTER– 1**

# **1.INTRODUCTION**

From a technological point of view, online training is practical and easy to implement. Online learning management systems (LMSs) are widely accessible, as are open-source tools (e.g., Sakai and Moodle). Thus, any higher education institution can install these systems in an agile manner and begin uploading content and registering users. At the methodological and pedagogical levels, it is somewhat more complicated. These contexts involve developing content and implementing a face-to-face model or a remote, purely online model – or a blended model that includes face-to-face and online learning. The means of delivery is key for tailoring the methodology, interaction, and performance. The current context, with the pandemic still underway and vaccines gradually rolling out, is characterized by severe mobility restrictions by country and region, which has in a few months driven the migration to online learning that has been slowly evolving over the last 20 years. However, all the complexity of an academic structure and community cannot immediately be transferred from traditional teaching models to the online model . Adaptation requires rigorous analysis and meticulous planning. Therefore, it is necessary to identify systems that support the successful design and implementation of online learning and teaching processes.

* 1. **PROBLEM STATEMENT**

The problem statement is to analysis the student behaviour based on their performance.

* 1. **MOTIVATION**

A wealth of recent experiments by Dickinson and Balleine (2002) has been devoted to the motivational control of conditioned behaviour, revealing a rich and intricate tapestry of effects. A large body of literature suggests that motivational states affect at least some classes of instrumental actions only indirectly. According to Adams and Dickinson (1981) specifically, motivational states affect moderately trained instrumental actions by determining the incentive value of the outcome of the behaviour.

## 

## **OBJECTIVES**

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it. The technique has been applied in the study of mathematics and logic since before Aristotle (384–322 B.C.), though analysis as a formal concept is a relatively recent development.

**CHAPTER– 2**

# **2.LITERATURE SUREVY**

**2.1 TITLE**: A Predictive System Informed by Students’ Similar Behaviour

**AUTHORS**: Daniel Burgos

It is quite complex to adapt instruction to student needs in view of online education owing to the ensuing communication disconnection in such learning environments. Decision support schemes offer assistance by automatically gathering students’ data and forwarding them to the tutor in the appropriate perspective, in order to predict their behaviour and implement some action beforehand to avert or promote the final upshot. This study shows of a decision support scheme known as u-Tutor that is centred on the similarity computation between learners in the past, and how it was used in a real-case scenario. For this case study, this tool has been utilized by two real courses comprising of 392 learners alongside academic faculty, as of 2015 to 2019. The analysis offered focuses on 3 research areas: (1) perceived usefulness, (2) usability of the tool and (3) success rate of classification. From the acquired data, it can be seen that the teaching group managed to offer excellent approximations for those learners who eventually managed to pass the course, whereas u-Tutor seemed to be an early warning for learners at risk, indicating its capacity as a tutors’ supportive tool.

**KEYWORDS**: Recommended System, Background Similarities, Student Behaviour.

**LIMITATIONS**: Limitations of the study.

**2.2 TITLE**: Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course

**AUTHORS:** Youngjin Lee and University of Kansas

This study investigated whether clustering can identify different groups of students enrolled in a massive open online course (MOOC). This study applied self-organizing map and hierarchical clustering algorithms to the log files of a physics MOOC capturing how students solved weekly homework and quiz problems to identify clusters of students showing similar problem-solving patterns. The usefulness of the identified clusters was verified by examining various characteristics of students such as number of problems students attempted to solve, weekly and daily problem completion percentages, and whether they earned a course certificate. The findings of this study suggest that the clustering technique utilizing self-organizing map and hierarchical clustering algorithms in tandem can be a useful exploratory data analysis tool that can help MOOC instructors identify similar students based on a large number of variables and examine their characteristics from multiple perspectives.

**KEYWORDS**: Massive open online Course, Self organization map.

LIMITATIONS: Number of problems students attempted to solve.

**2.3 TITLE**: Learning the Number of Clusters in Self Organizing Map

**AUTHORS**: Guenael Cabanes and Younes Bennani

The Self-Organizing Map (SOM: Kohonen (1984, 2001)) is a neuro-computational algorithm to map high-dimensional data to a two-dimensional space through a competitive and unsupervised learning process. Self-Organizing Maps differ from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This unsupervised learning algorithm is a popular nonlinear technique for dimensionality reduction and data visualization.

**KEYWORDS**: SOM, Student behaviour.

**LIMITATIONS:** Various approaches have been proposed to solve the problem.

**2.4 TITLE**: Theimpact of online learning on students’ course outcomes

**AUTHORS**: Di Xu and Shanna Smith Jaggars

Using a large administrative dataset from a statewide system including 34 community and technical colleges, the authors employed an [instrumental variable](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/instrumental-variables) technique to estimate the impact of online versus face-to-face course delivery on student course performance. The travel distance between each student's home and college campus served as an instrument for the likelihood of enrolling in an online section of a given course. In addition, college-by-course fixed effects controlled for within- and between-course selection bias. Analyses yield robust negative estimates for online learning in terms of both course persistence and course grade, contradicting the notion that there is no significant difference between online and face-to-face student outcomes—at least within the community college setting. Accordingly, both two-year and four-year colleges may wish to focus on evaluating and improving the quality of online coursework before engaging in further expansions of online learning.

**KEYWORDS**: Face-to-face course, Recommended system

**LIMITATIONS**: Time consumption is more for loading the input dataset.

**2.5 TITLE**: Comparing Student Performance: Online versus Blended versus Face-to-Face

**AUTHORS**: Larson, David K.; Sung and Chung- Hsien

The purpose of this research was to perform a three way comparison of delivery modes for an introductory Management Information Systems course to determine if there existed a difference in student success among the delivery modes. The research compares student exam and final grade results in this class that was taught by the same instructor using face-to-face, blended and online delivery modes. An Analysis of Variance test was used on the exam and final grade data to determine if a significant difference existed. Additionally, a discussion of this class in relation to student satisfaction, learning effectiveness and faculty satisfaction is presented. This research demonstrates that there is no significant difference among delivery modes. Additionally, blended and online modes for this class do very well when measuring student satisfaction, learning effectiveness and faculty satisfaction.

**KEYWORDS:** Distance Learning, Management Information Systems, Student Attitudes.

**LIMITATIONS**: Accuracy level is not good .

**2.6 EXISTING SYSTEM**

The previous predictive models only focused on using the student’s demographic data like gender, age, family status, family income and qualifications. In addition to the study related attributes including the homework and study hours as well as the previous achievements and grades. These previous work were only limited to provide the prediction of the academic success or failure, without illustrating the reasons of this prediction. Most of the previous researches have focused to gather more than 40 attributes in their data set to predict the student’s academic performance. These attributes were from the same type of data category whether demographic, study related attributes or both, that lead to lack of diversity of predicting rules.

**DISADVANTAGES OF EXISTING SYSTEM:**

* As a result, these generated rules did not fully extract the knowledge for the reasons behind the student’s dropout.
* Apart from the previously mentioned work, there were previous statistical analysis models from the perspective of educational psychology that conducted a couple of studies to examine the correlation between the mental health and the academic performance.
* The type of the recommendations was too brief, they missed illustrating the methodologies to apply them.

**Algorithm**: K Nearest Classifier and Logistic regression.

**CHAPTER – 3**

**3.SYSTEM ANALYSIS & SPECIFICATION**

**3.1 HARDWARE REQUIRMENTS**

* OS – Windows 7, 8 and 10 (32 and 64 bit)
* RAM – 4GB

**3.2 SOFTWARE REQUIREMENTS**

* Anaconda navigator software tool
* jupyter notebook
* python language

**3.3PROPOSED SYSTEM**

The proposed framework firstly focuses on merging the demographic and study related attributes with the educational psychology fields, by adding the student’s psychological characteristics to the previously used data set (i.e., the students’ demographic data and study related ones). After surveying the previously used factors for predicting the student’s academic performance, we picked the most relevant attributes based on their rationale and correlation with the academic performance. An Unsupervised Ensemble Clustering used in this research.

#### **ADVANTAGES OF PROPOSED SYSTEM :**

The proposal aims to analyse student’s demographic data, study related details and psychological characteristics in terms of final state to figure whether the student is on the right track or struggling or even failing. In addition to extensive comparison of our proposed model with the other previous related models

**Algorithm**: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K Nearest Neighbours (KNN).

**3.4 FUNCTIONS & METHODS**

* **DATA COLLECTION**

Data used in this paper is a set of student details in the school records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

* **DATA PRE-PROCESSING**

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

1. **Formatting**
2. **Cleaning**
3. **Sampling**

**1.Formatting:** The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

**2.Cleaning:** Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonym zed or removed from the data entirely.

**3.Sampling:** There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

* **FEATURE EXTRATION**

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

* **EVALUATION MODEL**

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation to avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

**CHAPTER – 4**

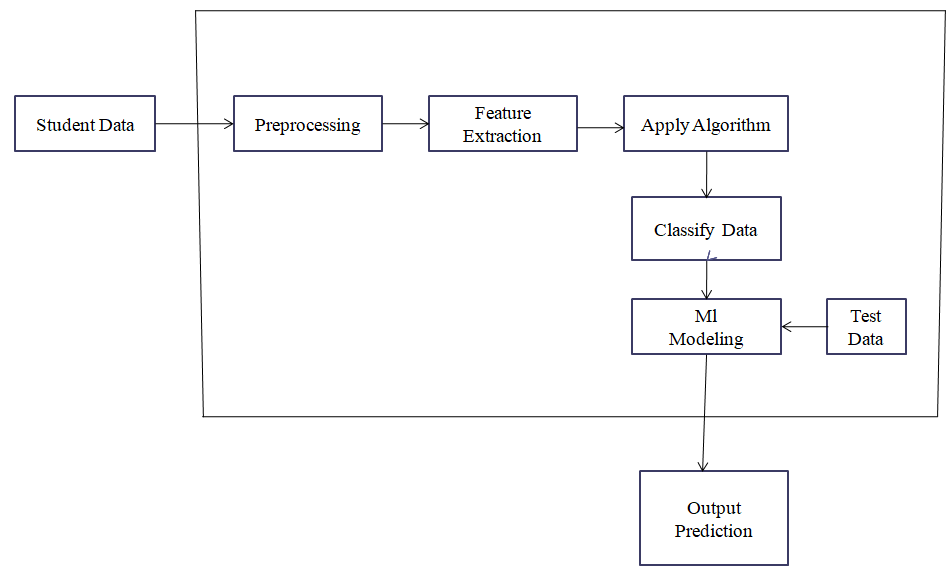
**4.SOFTWARE DESIGN**

Software design is a process to transform user requirements into some suitable form, which helps the programmer in software coding and implementation. Software design is the first step in SDLC (Software Design Life Cycle), which moves the concentration from problem domain to solution domain. It tries to specify how to fulfil the requirements mentioned in SRS (Software Requirement Specification).

## **4.1 SYSTEM ARCHITECTURE DIAGRAM**

## The software architecture diagram is a visual presentation of all of the aspects that constitute a system, either in part or whole. It is a depiction of a set of concepts that comprise architecture, such as its principles, components, and materials. It is also a system diagram used to abstract the general layout of the software system as well as the interactions, limitations, and limits between parts.

Also, an architecture diagram is a network map used to describe the general structure of a software program as well as the interactions, restrictions, and limits between elements. It is a significant tool since it offers a broader picture of the computer underlying physical installation as well as its development plan.



**Fig4.1.1: System Architecture diagram for student behaviour analysis**

**4.2DATA FLOW DIAGRAM**

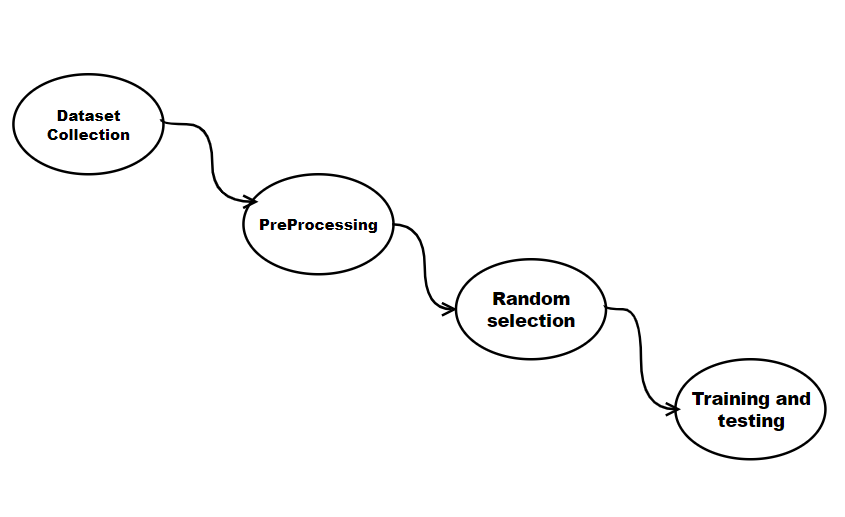
The Data Flow Diagram is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

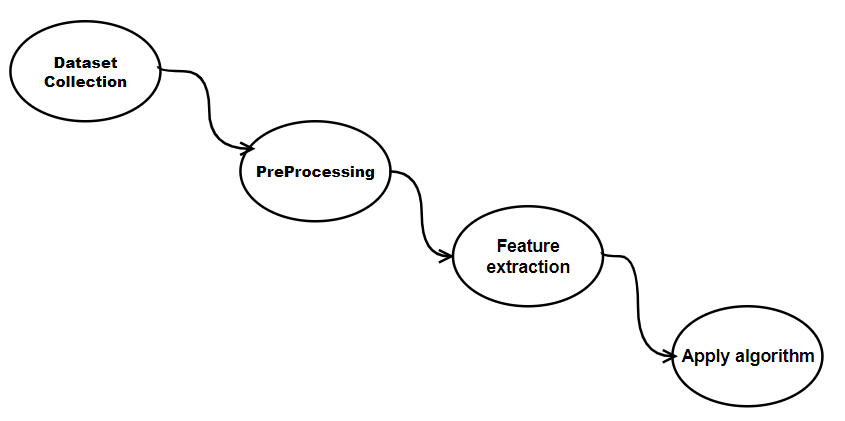
Data Flow Diagram shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

Data Flow Diagram is also known as bubble chart. A Data Flow Diagram may be used to represent a system at any level of abstraction. Data Flow Diagram may be partitioned into levels that represent increasing information flow and functional detail.

**LEVEL 0**

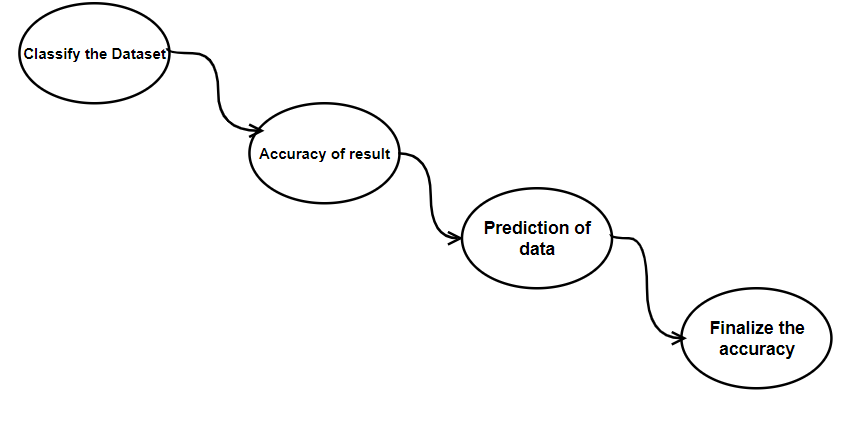
 **Fig 4.2.1 Data flow diagram with level 0**

**LEVEL 1**



**Fig 4.2.2 Data flow diagram with level 1**

**LEVEL 2**

****

**Fig 4.2.3 Data flow diagram with level 2**

## 

## **4.3 UML DIAGRAMS**

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of objectoriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modelling and other non-software systems.

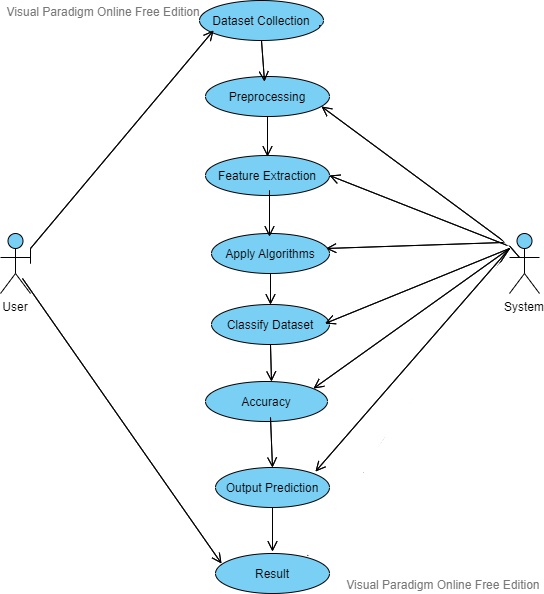
**GOALS:**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modelling language.
5. Support higher level development concepts such as collaborations, frameworks, patterns and components.

### **4.3.1 USE CASE DIAGRAM**

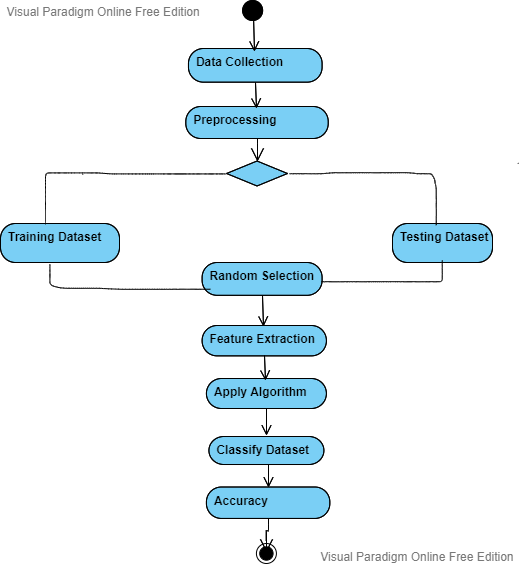
A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor.



**Fig 4.3.1 Use Case Diagram of Analysis and prediction of the student performance**

### **4.3.2 ACTIVITY DIAGRAM**

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



**Fig 4.3.2 Activity Diagram of Analysis and prediction of the student performance**

## 

## **CHAPTER-5**

## **5.INTRODUCTION TO TECHNOLOGIES**

## 

## **5.1 LOGISTIC REGRESSION**

It is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it.

**How logistic regression works**

**Logistic regression** uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).

**Advantages of using logistic regression:-**

* Logistic regression is easier to implement, interpret, and very efficient to train.
* It makes no assumptions about distributions of classes in feature space.
* It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
* It not only provides a measure of how appropriate a predictor (coefficient size)

Is, but also its direction of association (positive or negative)

**5.2 K NEAREST CLASSIFIER**

The K in the name of this classifier represents the k nearest neighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors.

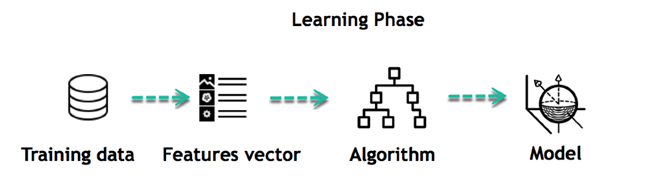
**5.3 MACHINE LEARNING OVERVIEW**

Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data to produce accurate results.

Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and Bayesian predictive modelling. The machine receives data as input, use an algorithm to formulate answers.

## **How does Machine learning work?**

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation.



For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model.

The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

**5.4 PYTHON OVERVIEW**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted:** Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive:** You can actually sit at a Python prompt and interactwith the interpreter directly to write your programs.
* **Python is Object-Oriented:** Python supports Object-Oriented style ortechnique of programming that encapsulates code within objects.
* **Python is a Beginner's Language:** Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Python’s features include:

* Easy-to-use
* Easy-to-maintain
* Easy-to-read
* Interactive mode
* GUI programming etc.,

**5.5 TENSOR FLOW**

The most famous deep learning library in the world is Google's TensorFlow.

Tensorflow architecture works in three parts:

* Pre-processing the data
* Build the model
* Train and estimate the model

It is called Tensorflow because it takes input as a multi-dimensional array, also known as **tensors**. You can construct a sort of **flowchart** of operations (called a Graph) that you want to perform on that input. The input goes in at one end, and then it flows through this system of multiple operations and comes out the other end as output.

This is why it is called TensorFlow because the tensor goes in it flows through a list of operations, and then it comes out the other side.

**5.6 ANACONDA NAVIGATOR**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, mac OS and Linux.

## **Why use Navigator?**

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions.

The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

**5.7 NUMPY**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object.

**5.8 HTML**

HTML (HyperText Markup Language) is the code that is used to structure a web page and its content. For example, content could be structured within a set of paragraphs, a list of bulleted points, or using images and data tables.

HTML is used to create electronic documents (called pages) that are displayed on the World Wide Web. Each page contains several connections to other pages called hyperlinks. Every web page you see was written using one version of HTML

**CHAPTER-6**

# **6.IMPLEMENTATION**

**6.1 PSEUDOCODE**

**6.1.1 CODE**

import warnings

warnings.filterwarnings('ignore')

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.metrics import accuracy\_score,confusion\_matrix

df=pd.read\_csv('xAPI-Edu-Data.csv')

df.head()

df.shape

df.info()

df.dropna()

df.isnull().sum()

**6.1.2 EDA**

sns.countplot(x="gender", order=['F','M'], data=df, palette="Set1")

plt.show()

sns.countplot(x="gender", order=['F','M'], hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

df['NationalITy'].value\_counts(normalize=True).plot(kind='bar')

plt.show()

df['PlaceofBirth'].value\_counts(normalize=True).plot(kind='bar')

plt.show()

sns.countplot(y="NationalITy", data=df, palette="muted")

plt.show()

sns.countplot(y="NationalITy", hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

sns.countplot(x="Relation", order=['Mum','Father'], data=df, palette="Set1")

plt.show()

sns.countplot(x="Relation", order=['Mum','Father'], hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

sns.countplot(x="StageID", data=df, palette="muted")

plt.show()

sns.countplot(x="StageID", hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

sns.countplot(x="GradeID", data=df, palette="muted")

plt.show()

sns.countplot(x="GradeID", hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

plt.subplot(1,2,1)

sns.countplot(x="SectionID", order=['A','B','C'], data=df, palette="muted")

plt.subplot(1,2,2)

sns.countplot(x="SectionID", order=['A','B','C'], hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

plt.subplot(1,2,1)

sns.countplot(y="Topic", data=df, palette="muted")

plt.subplot(1,2,2)

sns.countplot(y="Topic", hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

sns.countplot(x="ParentschoolSatisfaction", data=df, palette="muted")

plt.show()

sns.countplot(x="ParentschoolSatisfaction", hue="Class", hue\_order=['L','M','H'], data=df, palette="muted")

plt.show()

plt.figure(figsize=(8, 8))

sns.countplot('Class', data=df)

plt.title('Balanced Classes')

plt.show()

**6.1.3 PRE-PROCESSING**

from sklearn import preprocessing

le=preprocessing.LabelEncoder()

df['LGender'] = le.fit\_transform(df['gender'])#.values.reshape(-1,1).ravel())

df['LNationalITy'] = le.fit\_transform(df['NationalITy'])

df['LPlaceofBirth'] = le.fit\_transform(df['PlaceofBirth'])

df['LStageID'] = le.fit\_transform(df['StageID'])

df['LGradeID'] = le.fit\_transform(df['GradeID'])

df['LSectionID'] = le.fit\_transform(df['SectionID'])

df['LTopic'] = le.fit\_transform(df['Topic'])

df['LSemester'] = le.fit\_transform(df['Semester'])

df['LRelation'] = le.fit\_transform(df['Relation'])

df['LParentschoolSatisfaction'] = le.fit\_transform(df['ParentschoolSatisfaction'])

df['LParentAnsweringSurvey'] = le.fit\_transform(df['ParentAnsweringSurvey'])

df['LStudentAbsenceDays'] = le.fit\_transform(df['StudentAbsenceDays'])

df['LClass'] = le.fit\_transform(df['Class'])

df.head(1)

df=df.drop(["gender"],axis=1)

df=df.drop(["NationalITy"],axis=1)

df=df.drop(["PlaceofBirth"],axis=1)

df=df.drop(["StageID"],axis=1)

df=df.drop(["GradeID"],axis=1)

df=df.drop(["SectionID"],axis=1)

df=df.drop(["Topic"],axis=1)

df=df.drop(["Semester"],axis=1)

df=df.drop(["Relation"],axis=1)

df=df.drop(["ParentAnsweringSurvey"],axis=1)

df=df.drop(["StudentAbsenceDays"],axis=1)

df=df.drop(["ParentschoolSatisfaction"],axis=1)

df=df.drop(["Class"],axis=1)

df.head()

df.to\_csv('data.csv')

**6.1.4 UNIVARIATE SELECTION**

from sklearn.feature\_selection import SelectKBest

from sklearn.feature\_selection import chi2

x=df.iloc[:,df.columns !='LClass']

y=df.iloc[:,df.columns =='LClass']

bestfeatures = SelectKBest(score\_func=chi2, k=10)

fit = bestfeatures.fit(x,y)

dfscores = pd.DataFrame(fit.scores\_)

dfcolumns = pd.DataFrame(x.columns)

featureScores = pd.concat([dfcolumns,dfscores],axis=1)

featureScores.columns = ['Specs','Score']

featureScores.nlargest(10,'Score')

corr = df.corr()

**6.1.5 PLOT THE HEATMAP**

sns.heatmap(corr,

xticklabels=corr.columns,

yticklabels=corr.columns)

**6.1.6 SPLITTING AND CLASSIFICATION**

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test=train\_test\_split(x, y, test\_size=0.2)

**6.1.7 LOGISTIC REGRESSION**

from sklearn.linear\_model import LogisticRegression

lr=LogisticRegression()

lr.fit(x\_train,y\_train)

predict1=lr.predict(x\_test)

model1=accuracy\_score(y\_test,predict1)

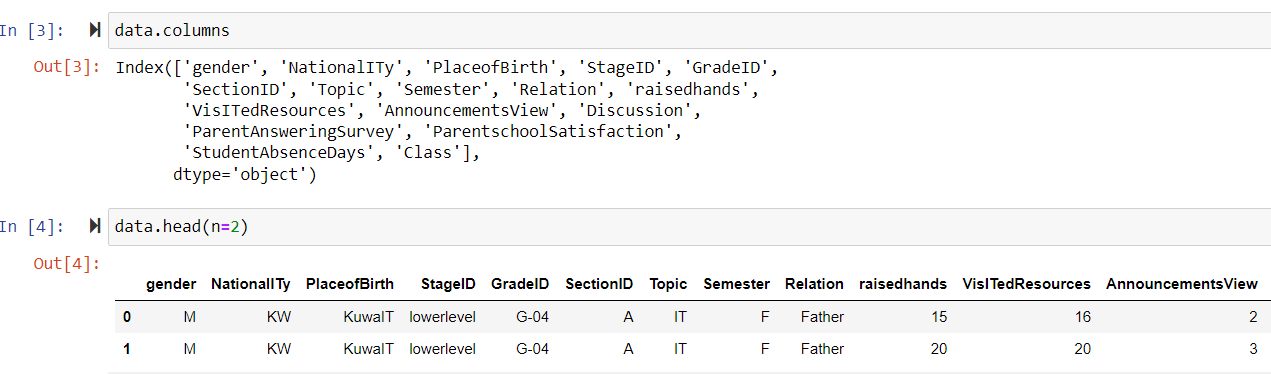
print(model1)

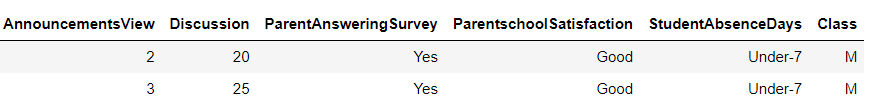
plt.show()

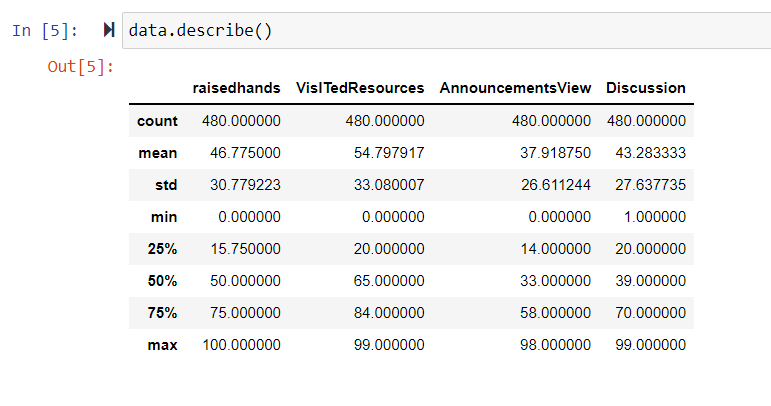
**CHAPTER – 7**

# **7.EXPERIMENTAL RESULTS**

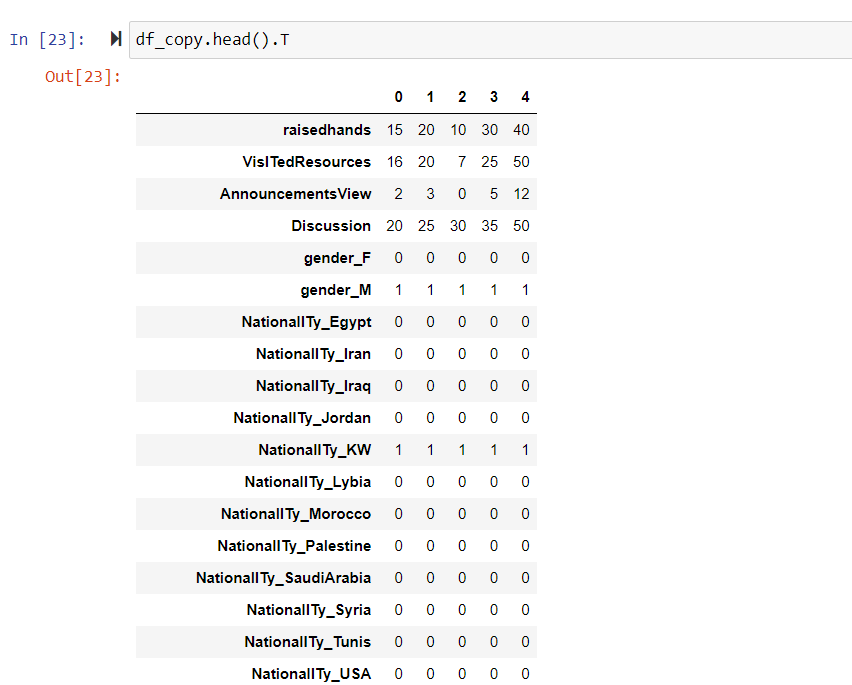
**7.1 LIST OF TABLES**

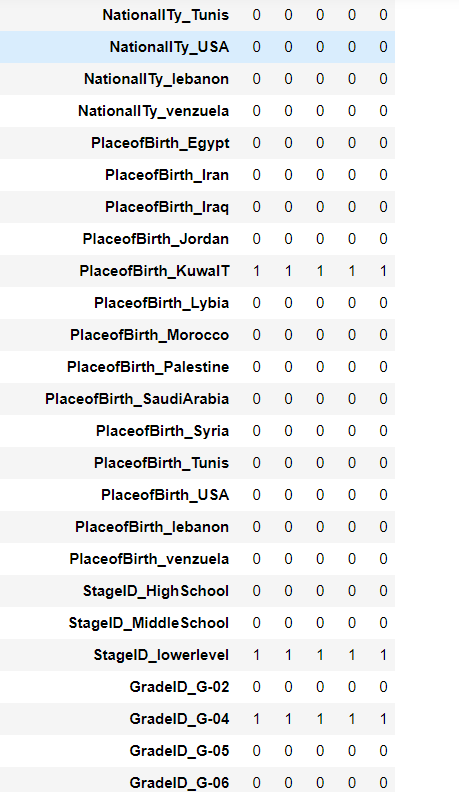
****

****

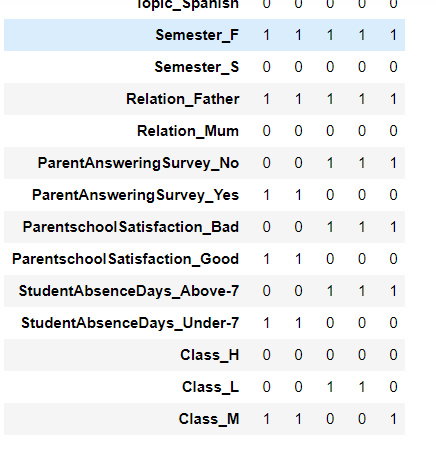
****

**Fig-7.1.1: Describe the Dataset**

****

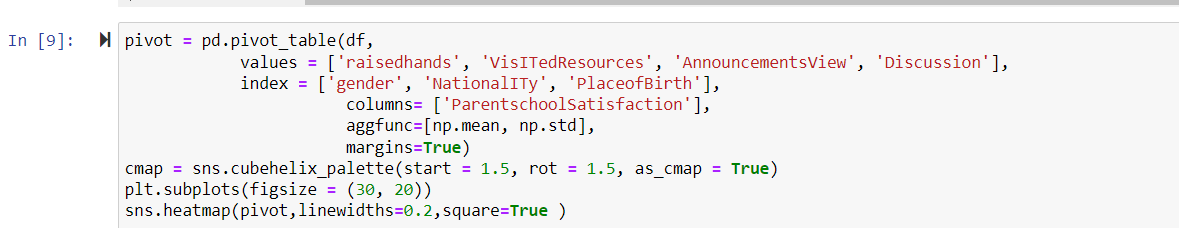
****

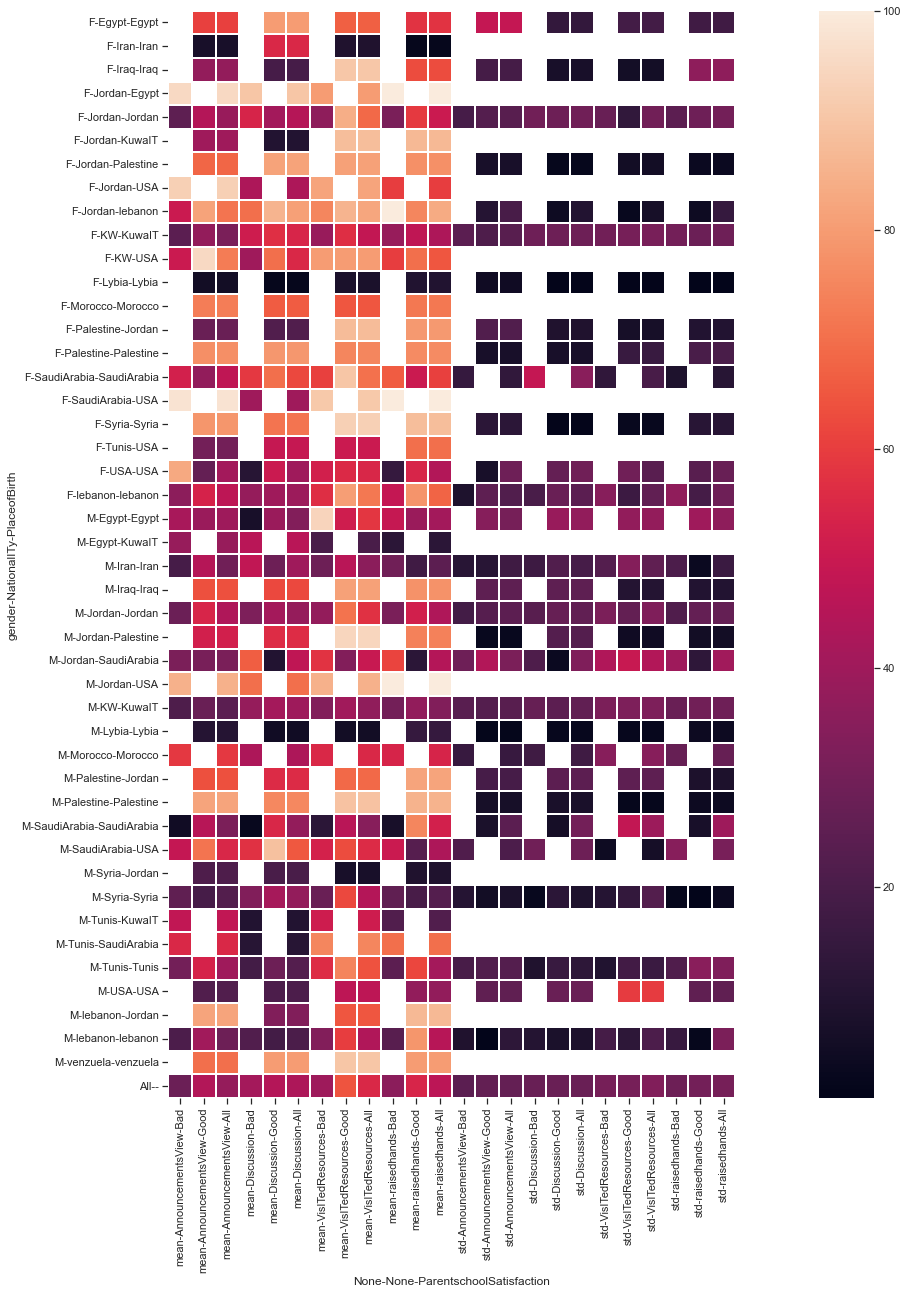
****

****

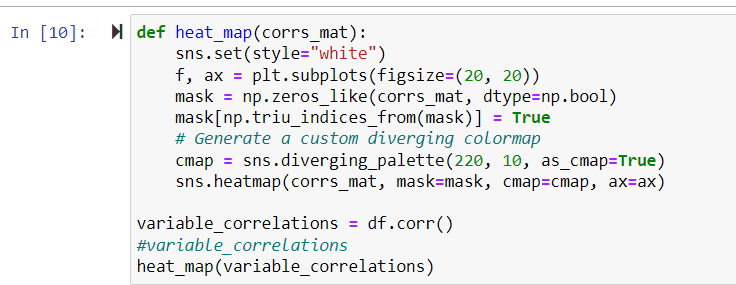
**Fig-7.1.2: Machine Learning**

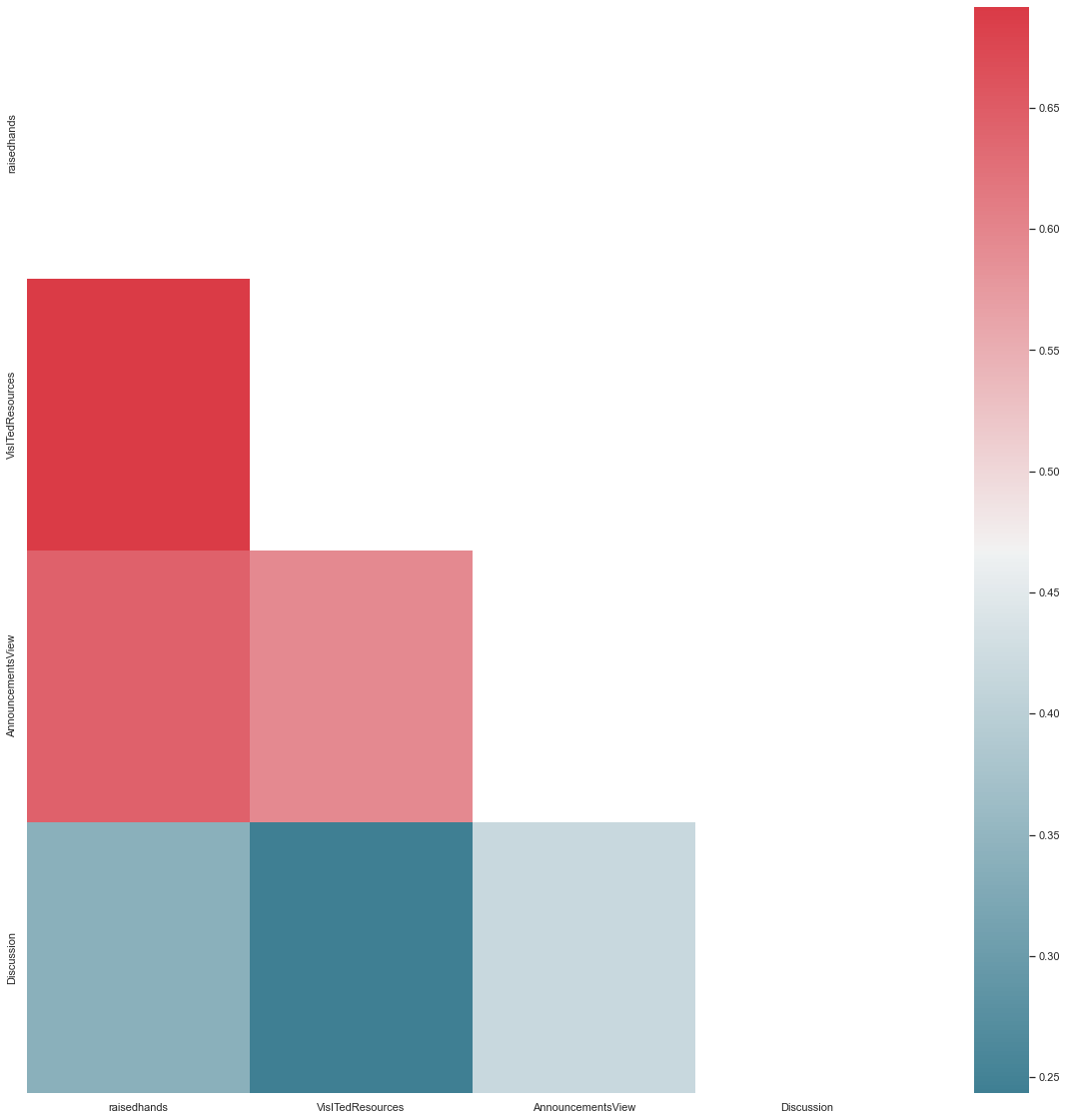
**7.2 LIST OF CHARTS**

****

****

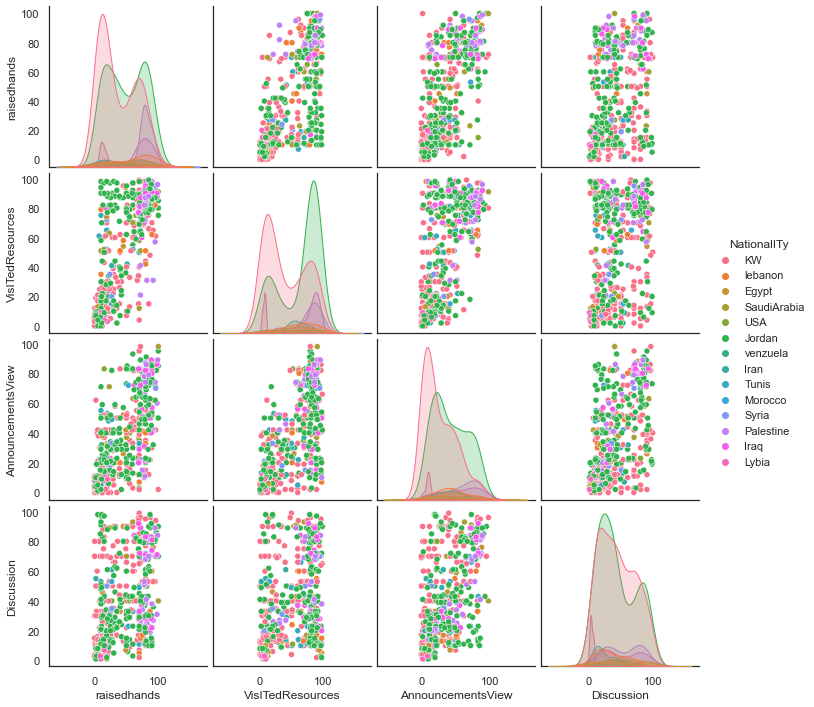
**Fig-7.2.1: Pivot Table**





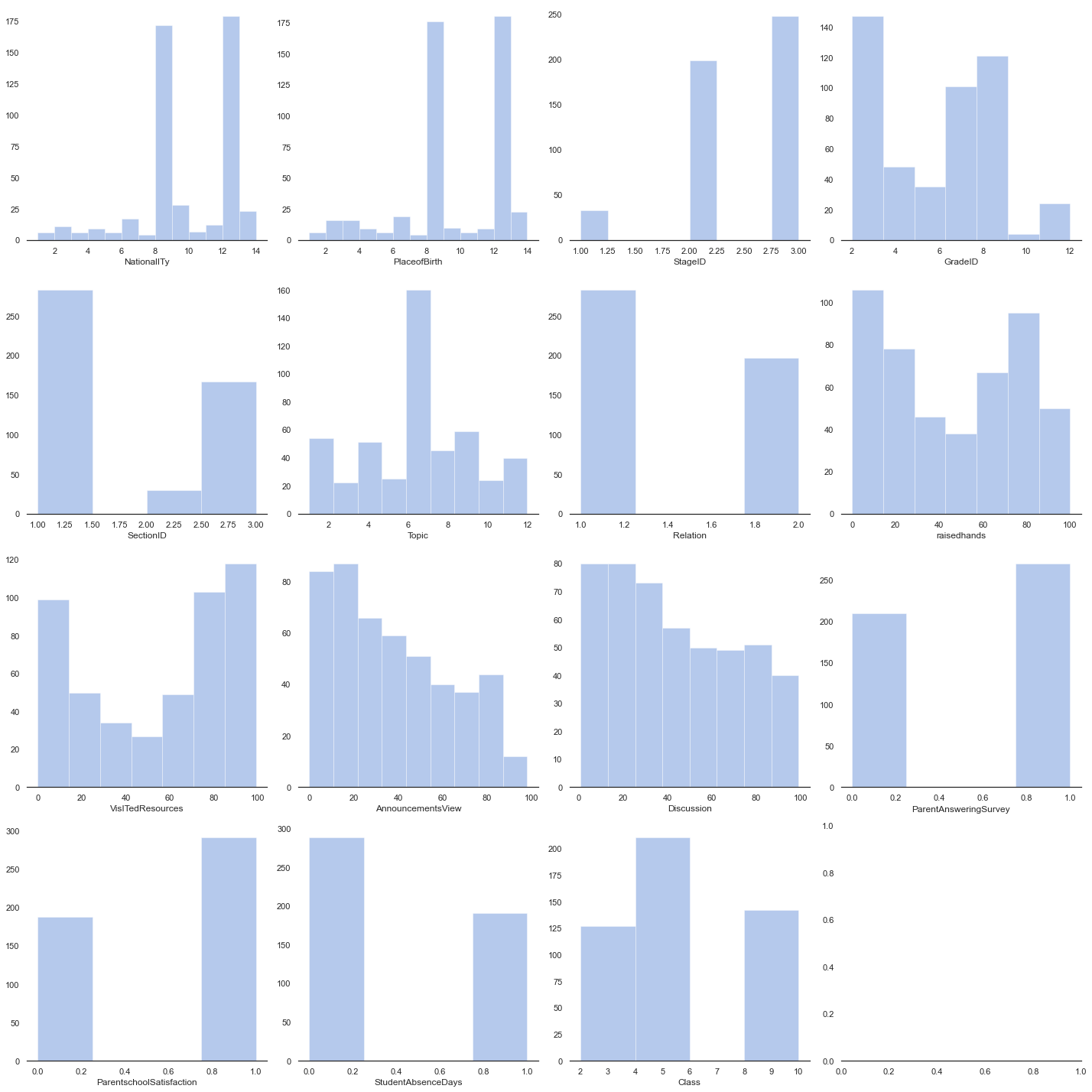
**Fig-7.2.2: (a)Simple Plots Correlation**

C:\Users\TOSHIBA\Pictures\OP 7.PNG



**Fig-7.2.2: (b)Simple Plots Correlation**

****

****

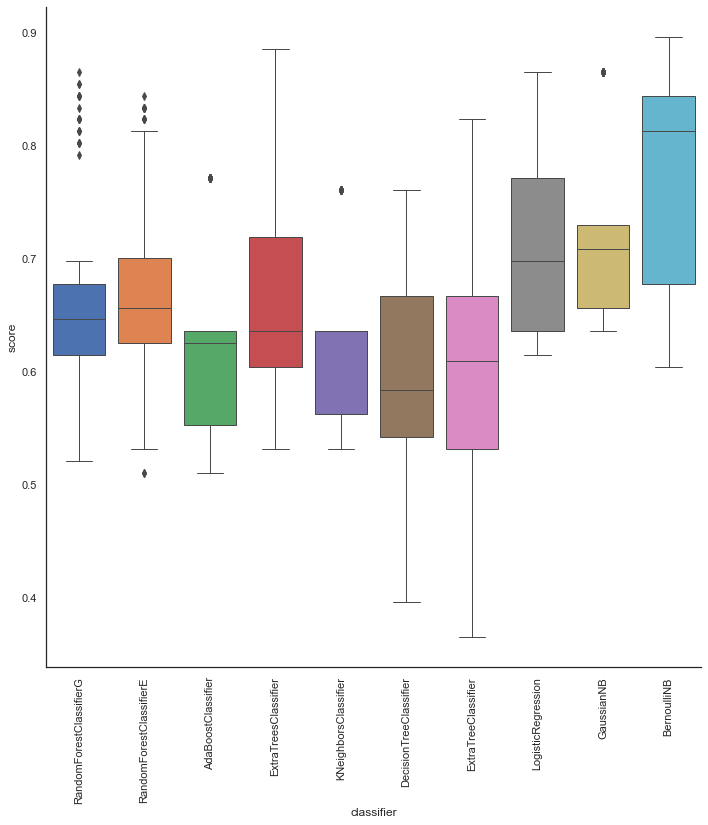
**Fig-7.2.3: (a)Complex Plots**

# C:\Users\TOSHIBA\Pictures\OP 9.PNG

# C:\Users\TOSHIBA\Downloads\OP 10.jfif

**Fig-7.2.3: (b)Complex Plots**

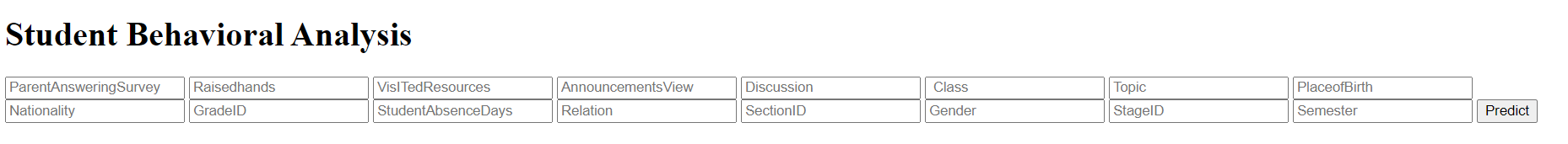
****

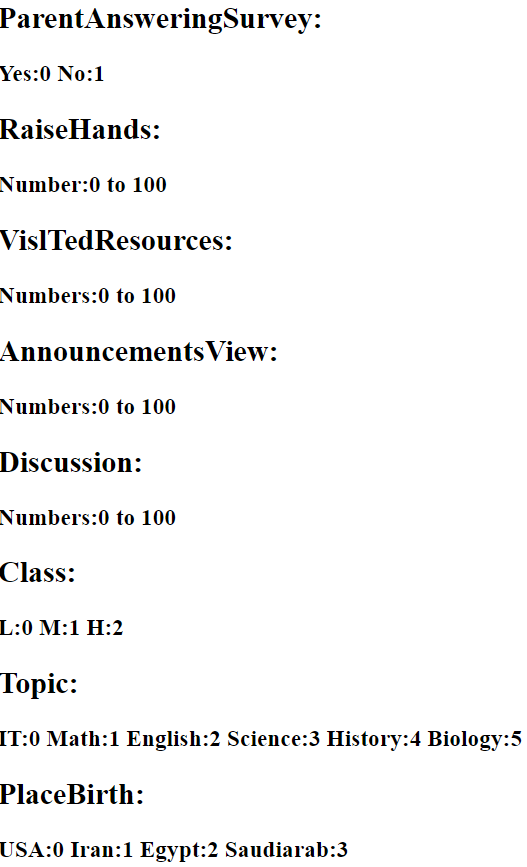
****

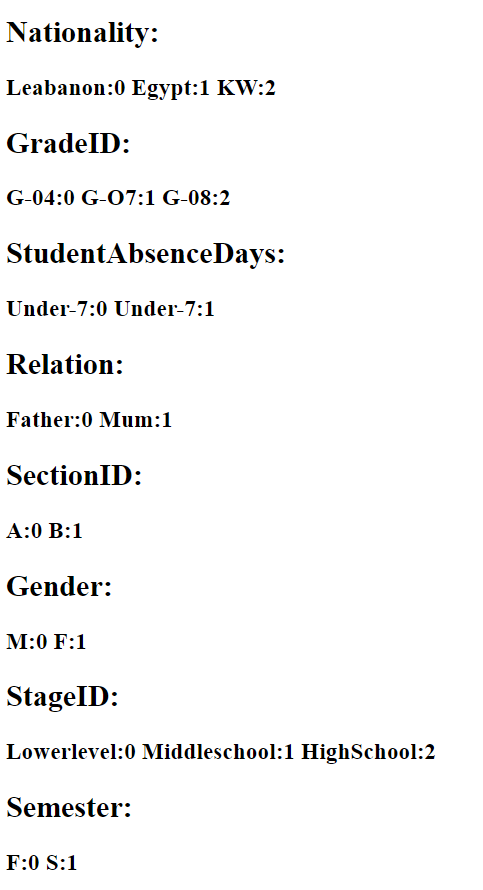
**Fig-7.2.4: Machine Learning**

**7.3 SYSTEM SCREENSHOTS**

**7.3.1: Step-1**

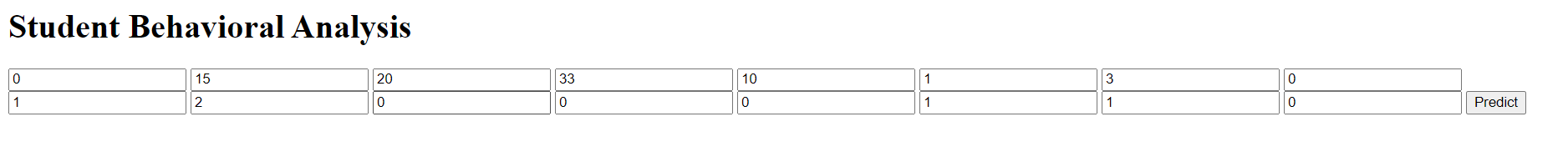
****

****

****

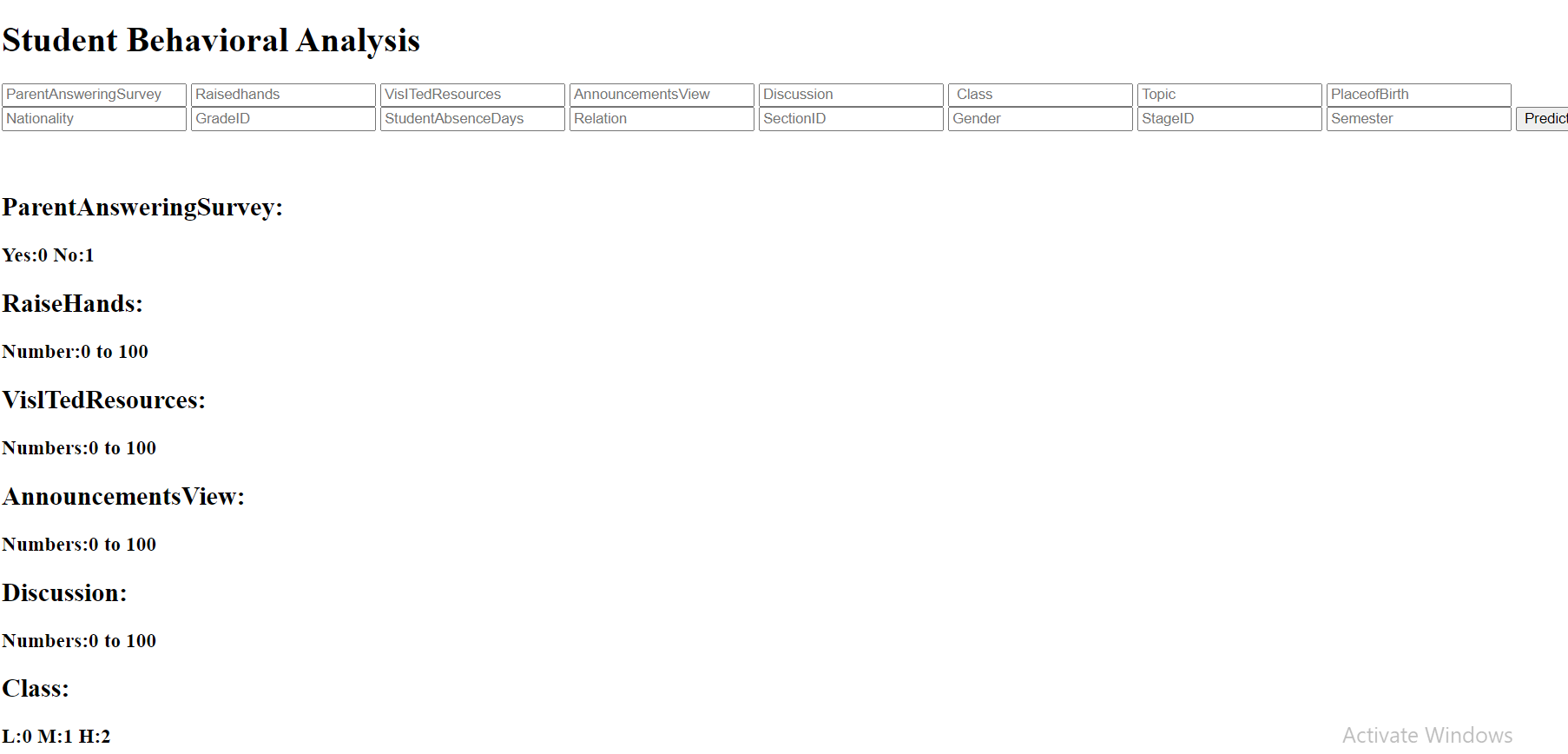
In step-1 we assign the values to the ParentAnsweringSurvey,RaiseHands, VislTedResources, AnnouncementViews, Discussion, Class, Topic, PlaceBirth, Nationality, GradeID, StudentAbsenceDays, Relation, SectionID, Gender, StageID and Semester.

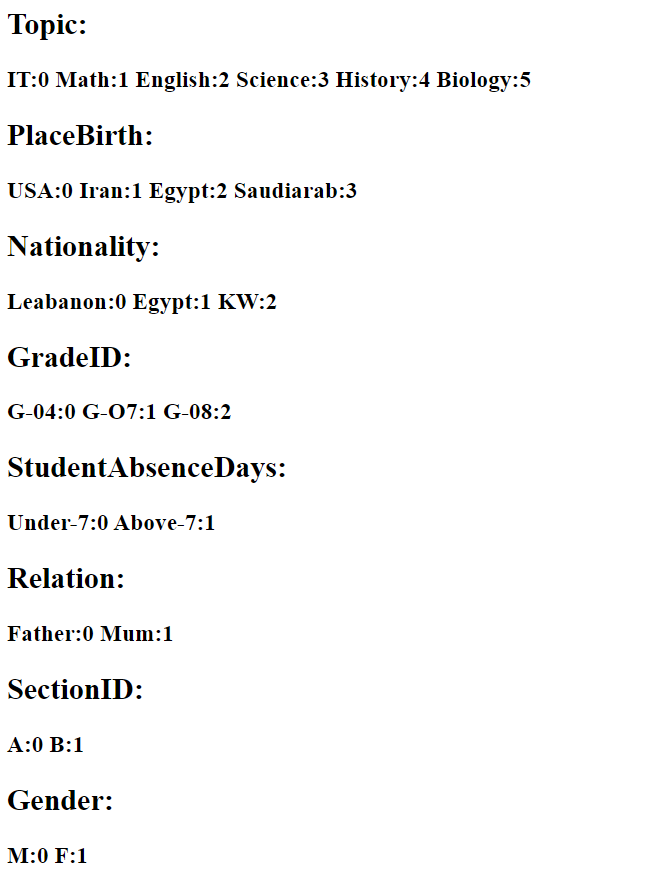
**7.3.2: Step-2**

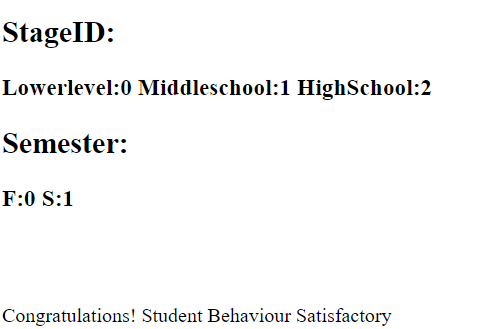
****

In step-2 we predict the assigning values.

**7.3.3: Step-3**

****

****

****

In step-3 we got the final output like **“Congratulations! Student Behaviour Satisfactory”.**

# **CHAPTER-8**

**8.CONCLUSION AND FUTURESCOPE**

## **8.1 CONCLUSION**

Finally, performance analysis for students are a major problem. It is important that they are countered. The work reported in this thesis indicates the machine learning techniques with supervised learning algorithms to understand the performance of algorithm with respect to student records where we analyses the performance of student and categorized it into three classes as high , average, low with the accuracy of 64% .

## **9.2 FUTURE SCOPE**

In the future we provide some technical solution by improve the efficiency of student performance .The user interaction model could be derived for giving the record of student dynamically and it could give staff an alert message about those students who are having low performance . We could build the prediction using Neural Network and can expect improvised results. We can add non- academic attributes along with academics attributes.

# 

# **CHAPTER-9**

# **9.REFERENCES**

* [1] R. Huang, A. Tlili, H. Wang, Y. Shi, C. J. Bonk, J. Yang, and D. Burgos, ‘‘Emergence of the online-merge-offline (OMO) learning wave in the postCOVID-19 Era: A pilot study,’’ Sustainability, vol. 13, no. 6, p. 3512, Mar. 2021.
* [2] M. Hughes and C. Hagie, ‘‘The positive and challenging aspects of learning online and in traditional face-to-face classrooms: Astudent perspective,’’ J. Spec. Educ. Technol., vol. 20, no. 2, p. 52, 2005.
* [3] D. Burgos, A. Tlili, and A. Tabacco, ‘‘Education in a crisis context: Summary, insights and future,’’ in Radical Solutions for Education in a Crisis Context (Lecture Notes in Educational Technology), A. Burgos, D. Tlili, and A. Tabacco, Eds. Singapore: Springer, 2021, pp. 3–10.
* [4] F. Simanca, R. Gonzalez, L. Rodriguez, and D. Burgos, ‘‘Personalized tutoring model through the application of learning analytics phases,’’ IEEE Latin Amer. Trans., vol. 18, no. 1, pp. 7–15, Jan. 2020.
* [5] F. Simanca, R. Gonzalez Crespo, L. Rodríguez-Baena, and D. Burgos, ‘‘Identifying students at risk of failing a subject by using learning analytics for subsequent customised tutoring,’’ Appl. Sci., vol. 9, no. 3, p. 448, 2019.
* [6] A. Horspool and C. Lange, ‘‘Applying the scholarship of teaching and learning: Student perceptions, behaviours and success online and face-to-face,’’ Assessment Eval. Higher Educ., vol. 37, no. 1, pp. 73–88, Feb. 2012.
* [7] D. Burgos, ‘‘A predictive system informed by students’ similar behaviour,’’ Sustainability, vol. 12, no. 2, p. 706, Jan. 2020.
* [8] P. Euzent, T. Martin, P. Moskal, and P. D. Moskal, ‘‘Assessing student performance and perceptions in lecture capture vs. face-to-face course delivery,’’ J. Inf. Technol. Educ., Res., vol. 10, pp. 295–307, Jan. 2011.
* [9] D. K. Larson and C.-H. Sung, ‘‘Comparing student performance: Online versus blended versus face-to-face,’’ Online Learn., vol. 13, no. 1, pp. 31–42, Feb. 2019.
* [10] A. Lopez, S. A. Gómez, D. Martín, and D. Burgos, ‘‘A framework for a semiautomatic competence valuation,’’ in Radical Solutions and eLearning: Practical Innovations and Online Educational Technology, D. Burgos, Ed. Singapore: Springer, 2020, pp. 215–236.
* [11] D. Xu and S. S. Jaggars, ‘‘The impact of online learning on students’ course outcomes: Evidence from a large community and technical college system,’’ Econ. Educ. Rev., vol. 37, pp. 46–57, Dec. 2013.
* [12] A. Driscoll, K. Jicha, A. N. Hunt, L. Tichavsky, and G. Thompson, ‘‘Can online courses deliver in-class results? A comparison of student performance and satisfaction in an online versus a face-to-face introductory sociology course,’’ Teach. Sociol., vol. 40, no. 4, pp. 312–331, Oct. 2012.
* [13] N. Kemp and R. Grieve, ‘‘Face-to-face or face-to-screen? Undergraduates’ opinions and test performance in classroom vs. online learning,’’ Frontiers Psychol., vol. 5, p. 1278, Nov. 2014.
* [14] M. Utari, B. Warsito, and R. Kusumaningrum, ‘‘Implementation of data mining for drop-out prediction using random forest method,’’ in Proc. 8th Int. Conf. Inf. Commun. Technol. (ICoICT), Jun. 2020, pp. 1–5.
* [15] S. Z. Selim and M. A. Ismail, ‘‘K-means-type algorithms: A generalized convergence theorem and characterization of local optimality,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-6, no. 1, pp. 7–81, Jan. 1984.
* [16] F. Nielsen, ‘‘Hierarchical clustering,’’ in Introduction to HPC with MPI for Data Science. Cham, Switzerland: Springer, 2016, pp. 195–211.
* [17] H.-P. Kriegel, P. Króger, J. Sander, and A. Zimek, ‘‘Density-based clustering,’’ Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 3, pp. 231–240, May 2011.
* [18] G. Xu, Y. Zong, and Z. Yang, Applied Data Mining. Boca Raton, FL, USA: CRC Press, 2013.
* [19] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, ‘‘A survey of clustering with deep learning: From the perspective of network architecture,’’ IEEE Access, vol. 6, pp. 39501–39514, 2018.
* [20] W. Wang, Y. Huang, Y. Wang, and L. Wang, ‘‘Generalized autoencoder: A neural network framework for dimensionality reduction,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2014, pp. 496–503