CHAPTER -1

INTRODUCTION

Education is a key factor for accomplishing an indelible economic progress. Student's academic performance centres on multiple aspects, making the analysis quite challenging. In later years, there has been an increase in the rate of interest and concern of people in the usage of data mining for educational and academic purposes [5]. Data mining portrays rising and upcoming areas of researches in education and it contains certain discrete requirements which other field's lack. In this project, the performance analysis of students have been discussed. The goal of this project is to provide a comparative study of student's performance through different algorithms [2].

There are many varying levels of school quality across India, as well as many different factors affecting student performance. There are certain questions arises while we talk about the student performance analysis such as

1. Have you ever wondered how different factors in your place of education actually affect your performance?

2. Have you wondered what specific things were causing you to perform a certain way academically?

3. As university students ourselves, we realized that there are many levels of influence, ranging from the students themselves, to academic advisors, to policymakers.

4. How can we bring all of these different groups together and provide an easy solution for all sides to boost education levels?

There are many studies in the learning field that investigated the ways of applying machine learning techniques for various educational purposes. One of the focuses of these studies is to identify high-risk students, as well as to identify features which affect the performance of students [3].

The study conducted by Kotsiantis is one of the initial studies which investigated application of machine learning techniques in distance learning for dropout prediction. The most significant contribution by this study was that it was a pioneer and carved the path for several such studies. While machine learning algorithms had been previously implemented in several settings, this was perhaps the first time that these techniques were applied to an academic environment.

Bhardwaj and Pal conducted a study in India, Faizabad to determine factors that most heavily affected student performance. They used Bayesian Classification for their study. The study by Erkan Er was based upon Kotsiantis as well as other similar studies. It concluded that Naive Bayes indeed performed better than any other machine learning algorithm. However, the crucial contribution of this study was that time-invariant features may be detrimental to the machine learning process, and hence are better left out of the study entirely. He also concluded that "Instead of demographic characteristics of students, using initial attendance and homework grades produces better prediction rate at earlier stages"[1].

With the wide usage of computers and internet, there has recently been a huge increase in publicly available data that can be analysed. Be it online sales information, website traffic, or user habits, data is generated every day. Such a large amount of data present both a problem and an opportunity. The problem is that it is difficult for humans to analyse such large data. The opportunity is that this type of data is ideal for computers to process, because it is stored digitally in a well-formatted way, and computers can process data much faster than humans [20].

The concept of machine learning is something born out of this environment. Computers can analyse digital data to find patterns and laws in ways that is too complex for a human to do. The basic idea of machine learning is that a computer can automatically learn from experience. Although machine learning applications vary, its general function is similar throughout its applications. The computer analyses a large amount of data, and finds patterns and rules hidden in the data [1]. These patterns and rules are mathematical in nature, and they can be easily defined and processed by a computer. The computer can then use those rules to meaningfully characterize new data. The creation of rules from data is an automatic process, and it is something that continuously improves with newly presented data. It’s focuses on supervised learning, more specifically predictive analytics, which is the process of using machine learning to predict future outcomes. Predictive analytics has a wide range of applications, such as fraud detection, analysing population trends, or understanding user behaviour [18].

The specific focus of this report is education. The aim is to predict student performance. Data about students is used to create a model that can predict whether the student is successful or not, based on other properties. First, the training data set is taken as input. There are two different data sets, containing different types of information. These data sets are in tabular format, where each row represents a student and each column, or variable, contains certain information about a student such as age, gender, family background or medical information. In addition, a column representing the success of the student is used as the variable that the algorithm is trying to predict. The algorithm creates a model, which is a function that outputs success or failure of the student, using other variables as input.

This report evaluates the effectiveness of different machine learning algorithms and methods. While algorithms that are used in creating predictive models are numerous, this thesis focuses on three of them, which are linear regression, decision trees, and naïve Bayes classification [2]. It is also measures the improvement made by feature engineering, which refers to modifying the data to make it more suitable for machine learning

Statement of Problem and Hypothesis:

Inductive machine learning is the process of learning from examples (instances), a set of rules, or more generally speaking a concept or a classifier that can be used to generalize to new examples. Inductive learning can be loosely defined for a two-class problem as the following. Let c be any Boolean target concept that is being searched for.

- Given a classifier L and a set of instances X for which c is defined over, train L on X to estimate c. The instances X which L is trained on are known as training examples and are made up of ordered pairs <x, c(x)>, where x is a vector of attributes (which have values), and c(x) is the associated classification of the vector x. L's approximation of c is its hypothesis h [3].

- In an ideal situation after training L on X, h equals c, but in reality a classifier can only guarantee a hypothesis h, such that it fits the training data. Without any other information we assume that the hypothesis, which fits the target concept on the training data, will also fit the target concept on unseen examples [11].

- A confusion matrix is presented in the table, which shows the type of classification errors a classifier can make for the two-class case. Thus, the breakdown of a confusion matrix is as follows: a is the number of positive instances correctly classified, b is the number of positive instances misclassified as negative, c is the number of negative instances misclassified as positive, d is the number of negative instances correctly classified.

Actually, the most well-known classifier criterion is its prediction accuracy. The prediction accuracy (denoted as acc) is commonly defined over all the classification errors that are made and it is calculated as:

$$\mathbf{acc = (a + d) / (a + b + c + d)}$$

In machine learning techniques classification speed is also in many cases a crucial property that is demanded by the classifier. This efficiency criterion is less often considered, but arises from the requirement that a classifier should use only reasonable amounts of time and memory for training and application provided by Gaga in the year 1996 [4].

In order to predict student's performance, the application of 6 most common machine learning techniques namely Decision Trees (given by Murthy, 1998), Neural Networks (given by Mitchell, 1997), Naive Bayes algorithm (given by Domingos and Pazzani (1997), Instance-Based Learning Algorithms (given by Aha, 1997), Logistic Regression (given by Long, 1997) and Support Vector Machines (given by Burges, 1998) are used [2]. In the next sub-section, we will briefly describe these supervised machine learning techniques. A detailed description can be found in Kotsiantis (2002).

CHAPTER-2

SYSTEM ANALYSIS

The word ―SYSTEM covers a very broad spectrum of concepts. This is derived from a Greek word *"system",* which means an organized relationship among the functioning units or components. System analysis, then, is the process of gathering and interpreting facts, diagnosing problems and using the information to recommend improvements to the system. In brief we can say that analysis specifies what the system should do. Design states how to accomplish the objective.

System analysis is concerned with becoming aware of the problem, identifying the relevant and decisional variables.

A detailed study of the process must be made by various techniques like interviews, questionaries' etc. The data collected by these sources must be scrutinized to arrive to a conclusion. The conclusion is an understanding of how the system functions. This system is called existing system. Now the existing system is subjected to close study and problem areas are identified. The designer now functions as a problem solver and tries to sort out the difficulties that the enterprise faces. The solutions are given as proposals. The proposal is then weighed with the existing system analytically and the best one is selected. The proposal is presented to the user for an endorsement by the user. The proposal is reviewed on user request and suitable changes are made. This is loop that ends as soon as the user is satisfied with proposal.

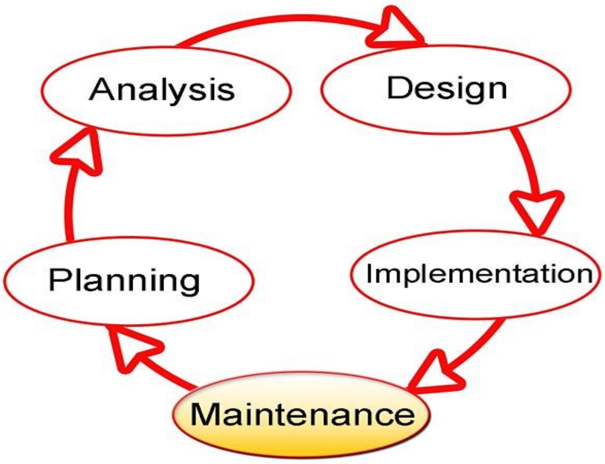


Figure 2.1: System Analysis

So, as mentioned in the figure 2.1, we can see that the System Analysis is a cyclic process which contain the various components.

These are carried out in the particular steps, which are as follows:

1. Planning
2. Analysis
3. Design
4. Implementation
5. Maintenance

2.1. PRELIMINAR INVESTIGATION

The purpose of preliminary investigation is to evaluate project requests. It is the process of collecting the information that help to evaluate the merit of the project request and make an informed judgment. Preliminary study is problem solving activity that requires intensive communication between the system users and system developers.

In this phase I conducted the following investigations:

2.1 REVIEWING ORGANISATIONL DOCUMENTS: -

During this phase we observed the various documents of the - Offline Communication within an Organization. We reviewed the following documents:

1. Communication report
2. Activities of user's report
3. Download Mail to get Information
4. Send Information to the user schedule by Admin

2.2 PROPOSED SYSTEM

The aim of proposed system is to develop a system which can provide the mental status of the students with greater accuracy. The proposed system can overcome all the limitations of the existing system. The system provides greater accuracy. The existing system has several disadvantages and many more difficulties to work well. The proposed system tries to eliminate or reduce these difficulties up to some extent. The proposed system will help the user to reduce the mental conflict. The proposed system helps the user to work user friendly and he can easily do his jobs without time lagging.

Expected Advantages of proposed System

The system is very simple in design and to implement. The system requires very low system resources and the system will work in almost all configurations. It has got following features:

1. Ensure data accuracy‘s.
2. Proper control of the higher officials.
3. Reduce the damages of the machines.
4. Minimize manual data entry.
5. Minimum time needed for the various processing.
6. Greater efficiency.
7. Better service.
8. User friendliness and interactive.
9. Minimum time required.

2.3 FEASIBILITY STUDY

Feasibility is the determination of whether or not a project is worth doing. This type of study determines if a project can and should be taken. A feasibility study is carried out to select the best system that meets performance requirements. A feasibility study of a system proposal is according to its workability, which is the impact on the organization, ability to meet their user needs and effective use of resources. Thus when a new application is proposed it normally goes through a feasibility study before it is approved for development. The document provides the feasibility of the project that is being designed and lists various areas that were considered very carefully during the feasibility study of this project such as Technical, Economic and Operational feasibility. To check whether the proposed system is worth making or not I conducted the feasibility study in which I studied the following aspects.

2.3.1. TECHNICAL FEASIBILTY

The system must be evaluated from the technical point of view first. The assessment of this feasibility must be based on an outline design of the system requirement in the terms of input, output, program and procedure. Having identified an outline system, the investigation must go to suggest the type of equipment, required method developing the system, of running the system once it has been designed.

Technical issues raised during the investigation are:

1 Does the existing technology sufficient for the suggested one?

2 Can the system expand if developed?

The project should be developed such that the necessary functions and performance are achieved within the constraints. The project is developed within latest technology. Through the technology may become obsolete often some period of time, due to the fact that newer version of same software supports older versions, the system may still be used. So there are minimal constraints involved with this project. The system has been developed using java the project is technically feasible for development.

Here we are concerned with specifying the equipment‘s and software that will successfully satisfy the requirements of the system. While studying the technical feasibility I observed the following aspects:

1 There is adequate and regular power supply.

2 There is easy availability of intranet network and also client of the system can avail the connection with minimum cost.

3 Specified H/W like processor of about 1GHz with a 256MB RAM and 4 S/W like R/Python and database are already available with the client.

2.3.2. OPERATIONAL FEASIBILTY

Here we consider the human aspect of the organization. This test of feasibility asks if the system will work when it is developed and installed. In this stage I observed the aspect. Everyone welcomes the new system and there was hardly any resistance because of ease of access and user friendly environment. The existing staff is skilled enough and is computer literate to handle the new system. It was perceived that the system, when launched, will do no harm to the existing business. After observing the above aspects, it was concluded that the system is operationally feasible.

2.3.3. ECONOMICAL FEASIBILTY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require.

Here we are concerned with the economical aspect of overall system. A system that can be developed technically and that will be used if installed, must still be profitable for the organization.

Here we estimated the following costs:

1. One-time development cost
2. One-time H/W and S/W cost
3. Periodic maintenance cost
4. After analysis it was found that this cost was less than the current system cost.
5. It also provides the following features:
6. Ease of access
7. Fast processing
8. Fast retrieval of information

CHAPTER -3

Aims and Objective

The basic idea of machine learning is that a computer can automatically learn from experience. Although machine learning applications vary, its general function is similar throughout its applications.

This report focuses on supervised learning, more specifically predictive analytics, which is the process of using machine learning to predict future outcomes. Predictive analytics has a wide range of applications, such as fraud detection, analysing population trends, or understanding user behaviour.

- The aim is to predict if the student has passed the exam (provided test) or not by looking at the other variables (the column of the table).

- In this case, the column "Passed" is called the dependent variable, and every other variable is called the independent variable. In the "Passed" column, "1" means student has passed the exam and "0" means failure in the exam.

- By applying a machine learning algorithm to this data, a function can be created, also known as the prediction model that gives the value for the dependent variable as output, and takes every other variable as input [1].

- The act of creating a prediction model from previously known data is called training, and such data is called the training data or a training set. After the model is created, it must be applied to another data set to test its effectiveness [3].

- Data used for such purpose is called test data or test set.

- The reason for using two different sets is to ensure that the model is flexible enough to be used on data sets other than the one it was built with. Otherwise, the problem of overfitting may occur, which is when a model is accurate with its original data set, but performs poorly on other data sets, because it is overly complicated [2].

- A common method to avoid overfitting is to divide the input data set into training and test sets.

1. Review of literature:

The study conducted by Kotsiantis is one of the initial studies which investigated application of machine learning techniques in distance learning for dropout prediction. The most significant contribution by this study was that it was a pioneer and carved the path for several such studies. While machine learning algorithms had been previously implemented in several settings, this was perhaps the first time that these techniques were applied to an academic environment. Bhardwaj and Pal conducted a study in India, Faizabad to determine factors that most heavily affected student performance. They used Bayesian Classification for their study.

The study by Erkan Er was based upon Kotsiantis' as well as other similar studies. It concluded that Naive Bayes indeed performed better than any other machine learning algorithm. However, the crucial contribution of this study was that time-invariant features may be detrimental to the machine learning process, and hence are better left out of the study entirely. He also concluded that "Instead of demographic characteristics of students, using initial attendance and homework grades produces better prediction rate at earlier stages." Student retention is an important issue in education [18].

While intervention programs can improve retention rates, such programs need prior knowledge of student’s performance. That is where performance prediction becomes important. The usage of machine learning to predict either the student performance or the student dropout is a commonly found subject in academic literature. Dropout prediction in virtual learning, or e-learning is a particularly common focus in such studies, due to both high dropout rates and easily available data [1]. Areas outside of virtual learning are also common contexts where dropout or performance predictions are used for research. The purpose of the research of these studies varies. In some of them, the aim is to find the best method for prediction. In others, the aim is simply to evaluate whether machine learning is a viable approach for predicting student dropout or performance.

One study evaluating the effectiveness of machine learning for dropout prediction was done at the Eindhoven University of Technology. Basic methodology was to build multiple prediction models using different machine learning methods, such as CART, BayesNet, and Logit. Then, prediction results of different models were compared in terms of their effectiveness. Most successful model was built by using the J48 classifier. A similar study was made by researchers from three different universities in India. A data set of university students was analysed by different algorithms, after which precision and recall values of the predictions were compared. The ADT decision tree model provided the most accurate results [19].

However, predicting student performance instead of student dropouts is more related with this pattern, and there are examples of such studies as well. One of these studies, made in the Hellenic Open University, analysed the usage of machine learning in distance education. Genetic algorithms and decision trees were used to build a predictive model, and the results were compared in terms of accuracy [21]. Most accurate results were provided by the GATREE i.e. genetically evolved decision trees model given by Kalles and Pierrakeas in the year 2006.

Another study about performance prediction was made at the University of Jordan. A data set of students from different countries was used. In addition to using individual machine learning methods, the researchers also applied ensemble methods, and compared the results between them. Decision trees provided the best results. Another area that the researchers focused on were behavioural features. A model was built with and without these features. It was found that the inclusion of behavioural features improved the prediction results [1].

The last study reviewed here was also about performance prediction. It was done at the University of Minho, Portugal. The data set contained information about whether the student had passed the exam in the subjects of math and Portuguese language. Decision trees, random forest, neural networks, and support vector machines were used. These methods were compared in terms of accuracy. Another comparison was made between a data set that included the past exam results and the one that did not. Inclusion of the past grades resulted in an improved performance.

The pattern is similar in most of these studies. First, different algorithms are applied to a data set to build prediction models. Then, predictions made by these models are compared using common evaluation criteria, such as accuracy, precision, and recall. Feature selection is also a commonly compared criteria. However, what these studies are missing is a more comprehensive comparison between distinct approaches such as method selection and feature engineering. This is the part where this algorithm can introduce a new approach [16]. By comparing the effectiveness of different processes used in machine learning, this pattern can provide insight into the more efficient ways to improve predictions in student performance.

In the paper “Data Mining Approach for predicting Student performance” published by E. Osmanbegovic et al. in the year 2012 they had applied three supervised machine learning algorithms to predict the success of a student in a course. They also predicted the performance of the learning methods where it was found that Naïve Bayes classifier had outperformed in prediction decision tree and neural network methods. In the year 2016 Snehal Kekane et al. published a paper titled “Automatic Student Performance Analysis and Automatic Student Performance Analysis and Monitoring” in which they proposed a system which will display in one single click the results of student performance by the user which will not only will initiate automation and also help in reducing manual efforts of the staff [2].

In the year 2015 titled “Predicting Academic Performance of Students using Data Mining Technique” by Indhu U Priya`1. et al. proposed system which focuses on the development of a tool for predicting the performance of a student for which pre-processing, classification and clustering techniques were used. Indhu U Priya et. al. published a paper titled “Predicting

Academic Performance of Students Using Data Mining Technique “in the year 2015 in which they proposed a system that focuses on the development of a tool for predicting the performance of a student [2].

In the paper titled “Student Performance Analysis System (SPAS)” published in 2015 by Chew Li SA Et Al few features were employed in the proposed framework during the design and implementation phase. Interface of the user and performance prediction were few of them which made sure that the objectives are achieved. In the year 2013, V. Ramesh et al. published a paper titled “Predicting Student Performance- a Statistical and Data Mining Approach” where they investigated a survey cum experimental methodology which was adopted to generate a database and was constructed from a primary and a secondary source [1]. It proved that Multilayer Perceptron (MLP) was considered to be the most appropriate classifier for prediction of student’s performance.

The other methods used were Naive Bayes, Multi-Layer Perception, SMO, J48, REP Tree algorithms. In the year 2007, a paper titled “Improving Student Performance in Public Primary Schools in Developing Countries: Evidence from Indonesia.” by Daniel Suryadarm, investigated the correlates of student performance in mathematics and dictation tests among schoolchildren in Indonesia. It finds a significant non-monotonic concave relationship between pupil-teacher ratio and student's mathematics performance. Ordinary Least Squares (OLS) And Data Regression Analysis were the techniques used [2]. The paper titled “Academic self-efficacy and first year college student performance and adjustment.” published by Martin M. Chemers in the year 2001, investigated that Self-efficacy can be considered an elementary feature in showing powerful relationships to academic progress and personal adjustment of the college students in the first year. A structural equation modelling (SEM) approach was used to test the adequacy of the hypothesized model.

In 2016 A. Bermejo Garcia published a paper titled “Student Performance Analysis” where the analysis of the performance of a student was done with the help of techniques such as Machine Learning. Besides the above algorithms, the tools and techniques used were Logistic Regression along with Support Vector Machine.

2. Existing methodology/Architecture:

➢ Existing Algorithm-

Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). There are different classification methods [7].

Bayesian classification is an algorithm that is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input [1]. Bayes rule or Bayes theorem is-

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

A more recent development in classification is that of artificial neural networks. These networks are modelled after the human neural system (hence the name), and have proven to be as powerful, if not more, as any other algorithm. While implementations may be complex, these networks are capable of understanding non-linear patterns in data.

A detailed description of the algorithms provided by Kotsiantis where he compared five algorithms, viz. Decision Trees, Naive Bayes algorithm (Bayesian networks), 3-NN (kNN), RIPPER (Rule Learning) and WINNOW (Perceptron based neural networks). This study was composed of two experimental stages, training and testing. During these stages, number of attributes was increased step-by-step [7]. For example, while only demographic data was included in the first step, performance attributes were added in the next step. Five algorithms were tested for each these subsequent steps and then they were compared. This comparative study helped in narrowing down candidates for our own application.

However, classification of data into binary groups seems insufficient. The primary goal of this study was only detecting at-risk students instead of determining performance levels of students. Classifying students according to their performances in different levels (e.g. poor, average, good, excellent, etc.) might be more useful. In this way, instructors can provide more adaptive feedback for each student [10].

- **Existing Methodology Feature-**

Bhardwaj and Pal conducted a study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance [17].

In the present study, those variables whose probability values were greater than 0.70 were given due considerations and the highly influencing variables with high probability values. These features were used for prediction model construction. For both variable selection and prediction model construction, the publishers have used MATLAB [3].

It is found that the second high potential variable for students' performance is their living location, and the third high potential variable for students' performance is medium of teaching. In Uttar Pradesh the mother tongue language of students is Hindi. Hence, students tend to be more comfortable in Hindi and other languages, than in the English language.

- **Existing Methodology Uniqueness-**

The study conducted by Erkan Er proved valuable in confirming the uniqueness of the proposed application. His work concluded that all current applications of machine learning in an academic setting were to predict dropout rates in a distance learning program [3]. There is perhaps no application that attempts to predict the absolute performance of the student. If one does exist, it has not been published yet.

3. Need for Proposal:

The Higher Education department in the HR development ministry puts out an annual survey called the **All India Survey on Higher Education (AISHE)**. It is supply-side and numerical. Based on the recent surveys, 65% of students who participate in an examination or the other secures a poor result despite taking enough time to appositely prepare for these tests. It therefore becomes extremely necessary to get to know the key responsible elements required to upgrade the performance of scholars [12].

All India Educational Survey (AIES) is conducted periodically. It's main purpose and objective is to collect, collate and communicate information of the nation's overall progress in the academics sector. There has been an improvement from 19.4 % in 2010-11 to 25.2 % in 2016-17, which is a significant achievement. However, AIES 2017 stated that the scenario is pretty staggering and a lot needs to be done. India, being a developing country has set an aggressive aim of attaining 30 % GER in higher education by 2020 [20].

Educational institutions will have to ramp up its efforts and get serious about what goes on in its organizations. For this lot of analysis needs to be done, and that is why this research have come up with this introspection about the various factors that are responsible for the performance of a student.

CHAPTER 4

REQUIREMENTS

Hardware and Software tools used:

4.1 Hardware:

The hardware configurations of the desktop used are processor speed of 3.20 GHz, RAM of size 4.00GB or more and system type of 64 bit operating system with a x64 based processor.

4.2 Language Used-

- **Python-**

  Python is an interpreted high-level programming language for general-purpose programming. It has a design philosophy that emphasizes code readability. Scikit-learn, a python machine learning library was also use along with matplot, a plotting library. Package of Pandas has been used, which is a python package and provides fast, flexible data structures which manipulates numerical tables as well as time series. An active internet connection for data recovering over network was also required [16].

  Python is a simple, easy to learn, powerful, high level and object-oriented programming language.

  Python is an interpreted scripting language also. Guido Van Rossum is known as the founder of python programming.

  Our Python tutorial includes all topics of Python Programming such as installation, control statements, Strings, Lists, Tuples, Dictionary, Modules, Exceptions, Date and Time, File I/O, Programs, etc. There are also given Python interview questions to help you better understand the Python Programming.

  Python Introduction

  Python is a general purpose, dynamic, high level and interpreted programming language. It supports Object Oriented programming approach to develop applications.

It is simple and easy to learn and provides lots of high-level data structures. It is easy to learn yet powerful and versatile scripting language which makes it attractive for Application Development.

Python's syntax and *dynamic typing* with its interpreted nature, makes it an ideal language for scripting and rapid application development.

Python supports multiple programming pattern, including object oriented, imperative and functional or procedural programming styles.

Python is not intended to work on special area such as web programming. That is why it is known as multipurpose because it can be used with web, enterprise, 3D CAD etc.

We don't need to use data types to declare variable because it is dynamically typed so we can write a=10 to assign an integer value in an integer variable.

Python makes the development and debugging *fast* because there is no compilation step included in python development and edit-test-debug cycle is very fast.

Python 2 vs. Python 3

In most of the programming languages, whenever a new version releases, it supports the features and syntax of the existing version of the language, therefore, it is easier for the projects built in the language to switch in the newer version. However, in the case of python, the two versions python 2 and python are very much different from each other.

The two differences which can be pointed out here is

1. Python 2 uses **print** as a statement and used as print "something" to print some string on the console. On the other hand, Python 3 uses **print** as a function and used as print("something") to print something on the console.
2. Python 2 uses the function raw input () to accept the user's input. It returns the string representing the value which has typed by the user. To convert it into the integer, we need to use the into () function in python. On the other hand, Python 3 uses input () function which automatically interpreted the type of input entered by the user. However, we can cast this value to any type by using primitive functions (int (), str (), etc.).
3. In python 2, the implicit string type is ASCII whereas, in python 3, the implicit string type is Unicode.

4. Python 3 doesn't contain the range () function of python 2. The xrange () is the variant of range () function which returns a xrange object that works similar to Java iterator. The range () returns a list for example the function range (0,3) contains 0, 1, 2.
5. There is also a small change made in Exception handling in python 3. It defines a keyword **as** which is necessary to be used. We will discuss it in Exception handling section of this tutorial.

Python Applications-

Python is known for its general purpose nature that makes it applicable in almost each domain of software development. Python as a whole can be used in any sphere of development.

Here, we are specifying applications areas where python can be applied.

1) Web Applications

We can use Python to develop web applications. It provides libraries to handle internet protocols such as HTML and XML, JSON, Email processing, request, beautiful Soup, Feed parser etc. It also provides Frameworks such as Django, Pyramid, Flask etc. to design and develop web based applications. Some important developments are: PythonWikiEngines, Pocoo, PythonBlogSoftware etc.

2) Desktop GUI Applications

Python provides Tk GUI library to develop user interface in python based application. Some other useful toolkits wxWidgets, Kivy, pyqt that are useable on several platforms. The Kivy is popular for writing multitouch applications.

3) Software Development

Python is helpful for software development process. It works as a support language and can be used for build control and management, testing etc.

4) Scientific and Numeric

Python is popular and widely used in scientific and numeric computing. Some useful library and package are SciPy, Pandas, IPython etc. SciPy is group of packages of engineering, science and mathematics.

5) Business Applications

Python is used to build Bussiness applications like ERP and e-commerce systems. Tryton is a high level application platform.

6) Console Based Application

We can use Python to develop console based applications. For example: **IPython**.

7) Audio or Video based Applications

Python is awesome to perform multiple tasks and can be used to develop multimedia applications. Some of real applications are: TimPlayer, cplay etc.

8) 3D CAD Applications

To create CAD application Fandango is a real application which provides full features of CAD.

9) Enterprise Applications

Python can be used to create applications which can be used within an Enterprise or an Organization. Some real time applications are: OpenErp, Tryton, Picalo etc.

10) Applications for Images

Using Python several applications can be developed for image. Applications developed are: VPython, Gogh, imgSeek etc.

There are several such applications which can be developed using Python.

- **RStudio-**

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, and debugging and workspace management. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Pre-packaged distributions of RStudio Desktop are available for Windows, macOS, and Linux. RStudio is available in open source and commercial editions and runs on the desktop (Windows, macOS, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE and SLES) [8].

RStudio and its team have contributed too many R packages. These include:

- Tidyverse – R packages for data science, including ggplot2, dplyr, tidyr, and purrr
- Shiny – An interactive web technology
- RMarkdown – Insert R code into markdown documents
- knitr – Dynamic reports combining R, TeX, Markdown & HTML
- packrat – Package dependency tool
- devtools – Package development tool

Features of R-

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R –

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

As a conclusion, R is world's most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications. This tutorial will teach you R programming along with suitable examples in simple and easy steps.

Who uses R?

- The Consumer Financial Protection Bureau uses R for data analysis
- Statisticians at John Deere use R for time series modelling and geospatial analysis in a reliable and reproducible way.
- Bank of America uses R for reporting.
- R is part of technology stack behind Four square's famed recommendation engine.
- ANZ, the fourth largest bank in Australia, using R for credit risk analysis.
- Google uses R to predict Economic Activity.

- Mozilla, the foundation responsible for the Firefox web browser, uses R to visualize Web activity.

4.3 HTML

Hypertext Markup Language was developed by a consortium of organizations, the World Wide Web consortium (W3C). HTML is a set of tags and elements used to create Web pages. Web pages are documents that you view on the Web. Web pages are stored as files with the extension .htm or .html. HTML is a web based language. It is very easy to learn and use.

4.4 JAVA SCRIPT

Java script client-side, you can use a scripting language for creating interactive interfaces on the Web.

4.5 CSS

CSS stands for Cascading Style Sheet. It defines a way how the information is presented by all the browsers on the web. A style sheet is a set of rules that controls the formatting of HTML elements on one or more web pages. Thus, the appearance of a Web page can be changed by changing the style sheet associated with it. There is no need to make detailed changes within the Web page to change how it looks. As one style sheet can be used for a whole web site, it normally means that the overall size of the website is smaller and downloads required for each page can be decreased by up to 40%.

4.6 BACK-END: RDBMS

The proposed system falls under RDBMS (Relational Database Management System) category. I have adopted JSP as front end for the software and **MySQL** as back end. JSP is at present one of the most popular development platform for Web based system in both side Client-Side and Server-Side.

4.6.1 MySQL

It is at present the most reliable and secure RDBMS tool. MySQL works to efficiently manage its resource, a database of information, among the multiple clients requesting and sending data in the network. MySQL has many important features that make it not only an exceptional database management system but also an excellent database server choice for client/server database computing. So the overall system will prove too reliable, secure and efficient for the organization.

MySQL, a Relational Database Management System (RDBMS), works to efficiently manage its resource, a database of information, among the multiple clients requesting and sending data in the network. **MySQL** has many important features that make it not only an exceptional database management system but also an excellent database server choice for client/server database computing. Here are some of these features:

1. Its supports Microsoft 2000 and higher version
2. It has client/server features developers can use to minimize the network traffic between client and server. Therefore, application based on **MySQL** can deliver the performance that users require to the productive in their jobs.
3. It maintains database security using a system of access controls. This simply means that at an administrator's discretion, the administrator creates registered database user and then grants them the privileges to perform specific database operation and use specific data.

CHAPTER-5

PROPOSED METHODOLOGY/ARCHITECTURE

A student’s performance can be evaluated on the basis of student’s learning outcomes. These outcomes can be in the form of various assessments. They give us essential information about what a pupil is learning and about the extent to which the teaching goals are being met. However, grading shall be the most optimum technique. Grading is the application of the standardized measurements of varying levels of achievement. A grade is mainly classified as two types: Grade Point Average (GPA) and Cumulative Grade Point Average (CGPA). GPA is calculated by taking the number of grade points a student earned in a given period of time [13].

While there are various recognized correlating factors related to student’s academic progress, here in this investigation This paper has scrutinized some basic key elements like, age, sex, family size, parent’s cohabitation status, mother and fathers education etc.

Based on the above key factors clustering techniques are used to analyze the performance of the student and check how they vary from one another depending on the data set variables. The data has been taken from the UCI Machine learning repository [3]. It has been collected from the records of the two Portuguese schools using schools records and questionnaire. The dataset consists of attributes namely-Age, Sex, Grade point for three different semester, Fathers Job, Mothers job, Extracurricular activities, romantic activities, travel time, number of absences of student and alcohol consumption as mentioned above. The data consists of 394 students. The next step is pre-processing in which any data with null values has been dropped. In the third step after pre-processing the data has been selected according to the requirements of the particular query. Then the K Means and Mean shift clustering have been applied in python. Then results so obtained have been discussed. A graph is plotted between GPA secured by a student and the above factors one by one. The clustering tool will help us to know what the factors are and is it directly or inversely proportional to the performance of the student [19].

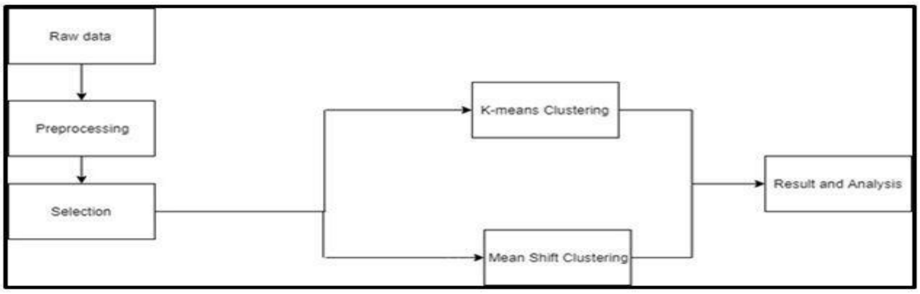


Fig 5.1: Data Flow diagram for the proposed framework

5.1 Algorithms to be used during project:

5.1.1 Naïve Bayes Algorithm:

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c) [3]. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [3].

Assumption:

The fundamental Naive Bayes assumption is that each feature makes an:

- Independent
- Equal contribution to the outcome. With relation to our dataset, this concept can be understood as:
- We assume that no pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being 'Rainy' has no effect on the winds. Hence, the features are assumed to be **independent**.

- Secondly, each feature is given the same weight (or importance). For example, knowing only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

Bayes' Theorem-

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$\mathbf{P\ (A/B) = P\ (B/A)\ P\ (A)\ /\ P\ (B)}$$

where A and B are events and P (B) != 0.

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.

- P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

5.1.2 Decision Tree:

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner [15].

Decision Tree Algorithm Pseudocode: -

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

Assumptions while creating Decision Tree: -

The below are the some of the assumptions we make while using Decision tree [1]:

- At the beginning, the whole training set is considered as the root**.**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Decision Trees follow Sum of Product (SOP) representation.

It’s a sum of product representation. The Sum of product (SOP) is also known as Disjunctive Normal Form. For a class, every branch from the root of the tree to a leaf node having the same class is a conjunction (product) of values, different branches ending in that class form a disjunction (sum) [2].

The primary challenge in the decision tree implementation is to identify which attributes we need to consider as the root node and each level. Handling this is know the attributes selection. We have different attributes selection measure to identify the attribute which can be considered as the root note at each level.

The popular attribute selection measures:

- Information gain
- Gini index

Attributes Selection

If dataset consists of **“n”** attributes, then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can’t solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criterion like information gain, gini index etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e., the attribute with a high value (in case of information gain) is placed at the root.

While using information Gain as a criterion, we assume attributes to be categorical, and for Gini index, attributes are assumed to be continuous.

Information Gain

By using information gain as a criterion, we try to estimate the information contained by each attribute. We are going to use some points deducted from information theory. To measure the randomness or uncertainty of a random variable X is defined by Entropy.

For a binary classification problem with only two classes, positive and negative class.

- If all examples are positive or all are negative then entropy will be zero i.e., low.
- If half of the records are of positive class and half are of negative class then entropy is one i.e., high [5].

By calculating entropy measure of each attribute we can calculate their information gain**.** Information Gain calculates the expected reduction in entropy due to sorting on the attribute. Information gain can be calculated [3].

5.1.3 Neural Network:

An Artificial Neural Network**,** often just called a neural network, is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data. There is no single formal definition of what an artificial neural network is. Generally, it involves a network of simple processing elements that exhibit complex global behaviour determined by the connections between the processing elements and element parameters [12]. Artificial neural networks are used with algorithms designed to alter the strength of the connections in the network to produce a desired signal flow. Neural networks are also similar to biological neural networks in that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned. The term “neural network” usually refers to models employed in statistics, cognitive psychology and artificial intelligence [1]. Neural network models which emulate the central nervous system are part of theoretical neuroscience and computational neuroscience.

The idea of how neural networks work-

Recently there has been a great buzz around the words "neural network" in the field of computer science and it has attracted a great deal of attention from many people. But what is this all about, how do they work, and are these things really beneficial?

Essentially, neural networks are composed of layers of computational units called neurons, with connections in different layers. These networks transform data until they can classify it as an output. Each neuron multiplies an initial value by some weight, sums results with other values coming into the same neuron, adjusts the resulting number by the neuron's bias, and then normalizes the output with an activation function.

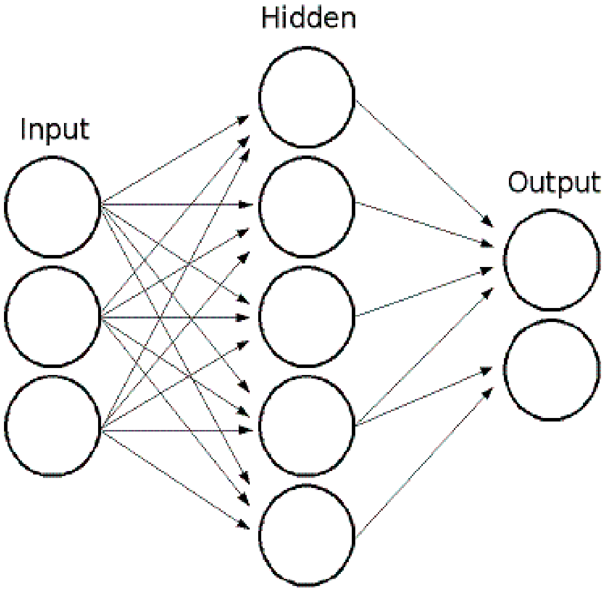


Figure 5.2: Neural Network representation with Hidden elements

The neural network are composed of the layers of computational unit, which contains some input elements and the output elements. The input section pass the message to the Hidden section then it will provide the output results as shown in the figure 5.2.

Iterative learning process

A key feature of neural networks is an iterative learning process in which records (rows) are presented to the network one at a time, and the weights associated with the input values are adjusted each time. After all cases are presented, the process is often repeated. During this learning phase, the network trains by adjusting the weights to predict the correct class label of input samples.

Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained. The most popular neural network algorithm is the backpropagation algorithm.

Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights (described in the next section) are chosen randomly. Then the training (learning) begins.

The network processes the records in the "training set" one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record.

This process occurs repeatedly as the weights are tweaked. During the training of a network, the same set of data is processed many times as the connection weights are continually refined.

So what's so hard about that?

One of the challenges for beginners in learning neural networks is understanding what exactly goes on at each layer. We know that after training, each layer extracts higher and higher-level features of the dataset (input), until the final layer essentially makes a decision on what the input features refer to. How can it be done?

Instead of exactly prescribing which feature we want the network to amplify, we can let the network make that decision. Let's say we simply feed the network an arbitrary image or photo and let the network analyse the picture. We then pick a layer and ask the network to enhance whatever it detected. Each layer of the network deals with features at a different level of abstraction, so the complexity of features we generate depends on which layer we choose to enhance.

Popular types of neural networks and their usage

In this post on neural networks for beginners, we'll look at auto encoders, convolutional neural networks, and recurrent neural networks.

Auto encoders-

This approach is based on the observation that random initialization is a bad idea and that pre-training each layer with an unsupervised learning algorithm can allow for better initial weights. Examples of such unsupervised algorithms are Deep Belief Networks. There are a few recent research attempts to revive this area, for example, using variation methods for probabilistic auto encoders.

They are rarely used in practical applications. Recently, batch normalization started allowing for even deeper networks, we could train arbitrarily deep networks from scratch using residual learning. With appropriate dimensionality and sparsity constraints, auto encoders can learn data projections that are more interesting than PCA or other basic techniques.

Let's look at the two interesting practical applications of auto encoders:

• In data demonising a denoising auto encoder constructed using convolutional layers is used for efficient denoising of medical images.

A stochastic corruption process randomly sets some of the inputs to zero, forcing the denoising auto encoder to predict missing (corrupted) values for randomly selected subsets of missing patterns.

• Dimensionality reduction for data visualization attempts dimensional reduction using methods such as Principle Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE). They were utilized in conjunction with neural network training to increase model prediction accuracy. Also, MLP neural network prediction accuracy depended greatly on neural network architecture, pre-processing of data, and the type of problem for which the network was developed.

Convolutional Neural Networks

Convent's derive their name from the "convolution" operator. The primary purpose of convolution in the case of a Convent is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. Convent's have been successful in such fields as:

- **Identifying faces**

In the identifying faces work, they have used a CNN cascade for fast face detection. The detector evaluates the input image at low resolution to quickly reject non-face regions and carefully process the challenging regions at higher resolution for accurate detection.

- **Self-driving cars**

In the self-driving cars project, depth estimation is an important consideration in autonomous driving as it ensures the safety of the passengers and of other vehicles. Such aspects of CNN usage have been applied in projects like NVIDIA's autonomous car.

CNN's layers allow them to be extremely versatile because they can process inputs through multiple parameters. Subtypes of these networks also include deep belief networks (DBNs). Convolutional neural networks are traditionally used for image analysis and object recognition. And for fun, a link to use CNNs to drive a car in a game simulator and predict steering angle.

Recurrent Neural Networks-

RNNs can be trained for sequence generation by processing real data sequences one step at a time and predicting what comes next. Here is the guide on how to implement such a model.

Assuming the predictions are probabilistic, novel sequences can be generated from a trained network by iteratively sampling from the network's output distribution, then feeding in the sample as input at the next step. In other words, by making the network treat its inventions as if they were real, much like a person dreaming.

- **Language-driven image generation**

Can we learn to generate handwriting for a given text? To meet this challenge a soft window is convolved with the text string and fed as an extra input to the prediction network. The parameters of the window are output by the network at the same time as it makes the predictions, so that it dynamically determines an alignment between the text and the pen locations. Put simply, it learns to decide which character to write next.

- **Predictions**

A neural network can be trained to produce outputs that are expected, given a particular input. If we have a network that fits well in modelling a known sequence of values, one can use it to predict future results. An obvious example is Stock Market Prediction.

Applying Neural Networks to Different Industries-

Neural networks are broadly used for real world business problems such as sales forecasting, customer research, data validation, and risk management.

Marketing-

Target marketing involves market segmentation, where we divide the market into distinct groups of customers with different consumer behaviour.

Neural networks are well-equipped to carry this out by segmenting customers according to basic characteristics including demographics, economic status, location, purchase patterns, and attitude towards a product. Unsupervised neural networks can be used to automatically group and segment customers based on the similarity of their characteristics, while supervised neural networks can be trained to learn the boundaries between customer segments based on a group of customers.

Retail & Sales

Neural networks have the ability to simultaneously consider multiple variables such as market demand for a product, a customer's income, population, and product price. Forecasting of sales in supermarkets can be of great advantage here.

If there is a relationship between two products over time, say within 3–4 months of buying a printer the customer returns to buy a new cartridge, then retailers can use this information to contact the customer, decreasing the chance that the customer will purchase the product from a competitor.

Banking & Finance

Neural networks have been applied successfully to problems like derivative securities pricing and hedging, futures price forecasting, exchange rate forecasting, and stock performance. Traditionally, statistical techniques have driven the software. These days, however, neural networks are the underlying technique driving the decision making.

Medicine

It is a trending research area in medicine and it is believed that they will receive extensive application to biomedical systems in the next few years. At the moment, the research is mostly on modelling parts of the human body and recognising diseases from various scans.

Conclusion

Perhaps NNs can, though, give us some insight into the "easy problems" of consciousness: how does the brain process environmental stimulation? How does it integrate information? But, the real question is, why and how is all of this processing, in humans, accompanied by an experienced inner life, and can a machine achieve such a self-awareness?

It makes us wonder whether neural networks could become a tool for artists—a new way to remix visual concepts—or perhaps even shed a little light on the roots of the creative process in general.

All in all, neural networks have made computer systems more useful by making them more human. So next time you think you might like your brain to be as reliable as a computer, think again—and be grateful you have such a superb neural network already installed in your head!

I hope that this introduction to neural networks for beginners will help you build your first project with NNs.

5.1.4 Logistic Regression:

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical) [4]. A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

What is Logistic Regression?

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

Derivation of Logistic Regression Equation

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression. In fact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression etc.) which included logistic regression as a special case.

The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Important Points

1. GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
2. The dependent variable need not to be normally distributed.
3. It does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
4. Errors need to be independent but not normally distributed.

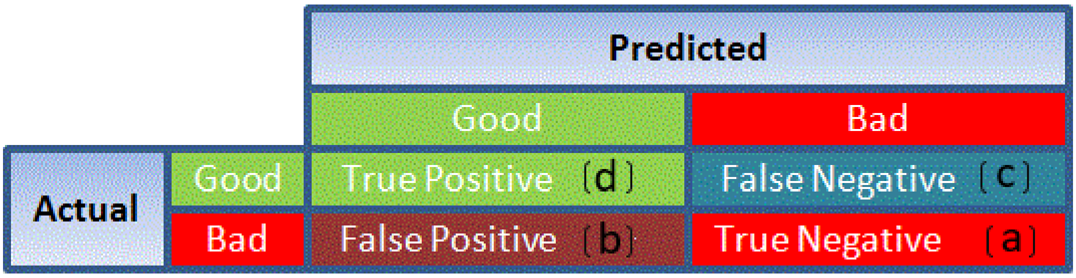
Performance of Logistic Regression Model-

To evaluate the performance of a logistic regression model, we must consider few metrics. Irrespective of tool (SAS, R, Python) you would work on, always look for:

1. **AIC (Akaike Information Criteria)** – The analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

2. **Null Deviance and Residual Deviance** – Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

3. **Confusion Matrix:** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. Figure 5.3 has shown how the confusion matrix actually works, this is how it looks like:



		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

Figure 5.3: Confusion Matrix

You can calculate the **accuracy** of your model with:

$$\frac{\text{True Positive} + \text{True Negatives}}{\text{True Positive} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

From confusion matrix, Specificity and Sensitivity can be derived as illustrated below:

$$\left.\begin{aligned}\text{True Negative Rate (TNR), specificity} &= \frac{A}{A+B}\\ \text{False Positve Rate (FPR), } 1-\text{specificity} &= \frac{B}{A+B}\end{aligned}\right\}\text{sum to 1}$$

$$\left.\begin{aligned}\text{True Positive Rate (TPR), sensitivity} &= \frac{D}{C+D}\\ \text{False Negative Rate (FNR)} &= \frac{C}{C+D}\end{aligned}\right\}\text{sum to 1}$$

5.1.5 K Nearest Neighbours:

5.1.5 K Nearest Neighbours: K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique [4].

Algorithm: -

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function.
If K = 1, then the case is simply assigned to the class of its nearest neighbour. It should also be noted that all three distance measures i.e. Euclidean, Manhattan and Minkowski are only valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset [8].

There are various algorithms apart from the above mentioned algorithms which will be used for the prediction and classification in the project for the better success ratio. Few algorithms are as follows:

- Multi-Layer Perception (MLP)
- Sequential Minimal Optimization (SMO)
- J48 Decision Tree
- REP Tree algorithms etc.

CHAPTER 6

IMPLEMENTATION

6.1 STANDARDIZATION OF THE CODING/CODE EFFICIENCY

In my project I have used various controls like command buttons, hover buttons and text box and similar other elements from various classes. Standardization of coding is a very good technique which should be followed while coding the project. Standardization of coding has various advantages over non-standard code such as:

- Standard code can be easily understood
- It is easy to debug
- It is easy to modify
- It is easy to upgrade

In this, project, much care has been given to in developing the standard program code. For example, each java class name start with uppercase letter and each method starts with a lowercase letter, but other words of method except first one starts with uppercase letter. All JSP files are stored in a folder named jsp, all bean classes are stored in a folder named beans with in SRC package. And all the contents of project code files have been stored according to a proper directory structure**.**

Code of the software is said to be efficient, if the complexity of all types is minimum. In the code of the developed software, I have tried to minimize the space and time complexity, so the code is efficient.

The complexities can be minimized in three ranges. They are called as organized simplicity, disorganized complexity and organized complexity.

Organized complexity is methodologically undeveloped in the sense that neither analytical nor statistical methods are adequate for dealing with systems that fit into it.

Computational complexity: -

Computational complexity is a characterization of time and space requirements for solving a problem by a particular algorithm. Either of these requirements is usually expressed in terms of a single that represent the size of the problem. Although computational complexity has been predominantly studied in terms of the time it takes to perform a computation, the amount of computer memory is required frequently just as important. This requirement is usually called the space requirement. It is expressed in the terms of a space complexity function, analogous to the time complexity function.

Since the modular approach is used to design the software. This approach uses the object oriented design methods. Hence the code is optimized due to the above reason. Because of object oriented programming, the features like modularity and reusability can be achieved in the software.

Reusability means – the programmer modifies the program ‘s functionality by replacing the old elements or objects with new objects or by simply plugging new objects into the application. General instruction requires no modification because specific implementation details reside within the object.

Optimization of coding-

Optimization of code is very important to produce the better quality of software. Without optimization it is very difficult to debug the syntactical and logical errors present in the code. If the code will be optimized, then it becomes trouble-free to find out the errors is stipulated time. Every step has been taken to optimize the coding. The efforts have been made to modularized the whole working of the software by which it will be easy to locate the errors is time saving manner. Any modification in the requirement of users can also be implemented if the code of the software will be optimized.

For the purpose of code optimization, the existing code has been re used, rather than doing all the coding from scratch. For example, all the bean classes and java classes are kept in source package and these are imported on the JSP pages whenever required.

6.2 ERROR HANDALING -

Handling of errors in this project is done by using try and catch statement for the various operations like insertion of data, updating of data etc. Here I am providing the code of different operations with data:

An Exception occurs when a program encounters any unexpected problems. Such as running out of memory or attempting to read from a file that no longer exists or ORACLE connections not found etc. These problems are not necessarily caused by a programming error but they mainly occur because of violation of assumption that we might have made about the execution environment. When a program encounters an exception the default behaviours is to throw the exception which generally translates to abruptly, terminating the program after displaying an error message. But this is not a characteristic of a robust application.

The best way is to bindle the exception situations if possible, gracefully recover from them so that program will not terminate abnormally, instead it will show proper error message. This is called ―exception handling‖.

In the ―Offline Communication Withing an Organizationl project, try, catch, finally and throw keywords has been used to handle the exception. And also, it has been tried to keep the condition of occurring an error using proper use of condition statements and displaying appropriate error message:

6.3 PARAMETERS CALLING/PASSING

Passing parameters from one web page to another is a very common task in Web development. There are many situations in which we need to pass data from one Web page to another. There are many techniques available for this purpose. The most common techniques used for this purpose are url rewriting, hidden form filled, session management, set Attribute () method of

HTTP request etc. In this project ―Offline Communication within an Organization‖**,** parameter passing has been done using url re-writing, session and setting attribute in request. However, every method of parameter passing its own limitations and scopes. Using url rewriting has disadvantage that it the passing parameters would be shown in the address bar, so it can‘t be used for passing sensitive data. Setting request attribute for passing parameter has limitation that parameter will only be available to the requested page and not elsewhere. Http Session is used when passing parameters specific to the client Session has number of advantages over the url re-writing and also some disadvantages. The comparison of these is described below-

1. Session works on server side, while url re-writing work on client side.
2. The information or data stored in query string in url re-writing is visible to everyone.

   But in Session it is hidden and can ‘t be viewed easily.
3. Query String in url re-writing can store only a piece of information but in Session we can store the more and more data.
4. The Query String speed never falls as the load increase because it stores a piece of information. But on the other hand Session increase congestion as the loads increase.

6.4 VALIDATION CHECKS

For every system, validations play a very important role while accepting the inputs from the users. This is because, the data being input by the user is further used to keep track of the various activities and their accounts.

In this the user inputs are validated because that data is then further used for generation of reports, for verification etc.

In this system two languages are used for validation checks.

Java Script is used for validating the user inputs for client site verification. Such as if a user inputs the email id in data field they have to input appropriate email id or if a user inputs name then they can ‘t use number in name.

The main validations that are done as follows:

1) All the screens have a similar look and feel. They all have the almost same color combinations in its background. This provides a better user interface to the users.
2) Whenever a page needs to be refreshed, only a required portion needs to be updated, instead of the whole page. This helps in fast refreshing of the page.
3) Whenever a user logs on to the website through the Login, his/her rights and privileges are checked and then he/she is allowed to work under the permitted rights only.
4) The data entered by the users is validated before the saving or retrieval of the record.
5) The user, except that administrator, is not allowed to message directly to the other user.
6) Password field contains at least 6 character long data, which stored in the database in the encrypted form.
7) If any data entered by users are wrong, then it will show an error message without proceeding any process.
8) Numeric field only contains numbers like age, mobile no, enrolment no, session, experience, day, month, year field. If any user entered alphabets, then it will display an error message to acknowledge the user that he/she is given a wrong entry.
9) After logout, you will have to re-sign-in for using your account.
10) Alphabetical field only contains only alphabets like first name, last name. If any user entered number, then it will display an error message to acknowledge the user that he/she is given a wrong entry.

11) Software Engineering- A Practitioner ‘s Approach, Roger S. Pressman; McGraw-Hill International Edition.

Thus, we have tried to make this system very secured and reliable by putting a number of validation checks in it. The future versions of this software are supposed to have more extended validations checks based on varied client needs.

CHAPTER-7

TESTING

The development of software systems involves a service of production activities where opportunities for injection of human fallibilities are enormous. Errors may begin to occur at the very inception of the process where the objectives may be erroneously or imperfectly specified, as well as later design and development stages. Because of human inability to perform and communicate with perfection, software companies are accompanied by a quality assurance activity.

Testing is to determine errors in a software code. It is crucial element of software quality assurance and represents the ultimate review of specifications, design and coding. The increasing visibility of a software as a system element and the attendant ―costs‖ associated with a software failure are motivating forces for well-planned through testing. Usually software development organizations expend between 30 to 40 percent of total project effort on testing. Our goal is to design a series of test cases that have a high likelihood of finding errors.

To test the software, there are so many testing techniques which provide systematic guidance for designing tests that exercise the internal logic of software components and exercise the input and output domains of the program to uncover errors in program function, behavior and performance.

If testing is conducted successfully, it will uncover errors in the software. As the secondary benefits, testing demonstrates that software functions appear to be working according to specification, that behavioural and performance requirements appear to have been met

Testing Principles: -

1. All tests should be traceable to customer requirements.
2. Tests should be planned long before testing begins.
3. Tests should begin with ―in the small‖ and progress toward testing ―in the large‖
4. Exhaustive testing is not possible.
5. To be more effective an independent third party should conduct testing.

7.1 TESTING TECHNIQUES

Software design is a critical element of software quality assurance and represents the ultimate review of specification, design and code generation. Once source code has been generated, software must be tested to uncover as many as errors as possible before delivery to the customer. Our goal is to design a series of test cases that have a high likelihood of finding errors. To test the software, there are so many testing techniques which provide systematic guidance for designing tests that exercise the internal logic of software components and exercise the input and output domains of the program to uncover errors in program function, behaviour and performance.

If testing is conducted successfully, it will uncover errors in the software. As the secondary benefits, testing demonstrates that software functions appear to be working according to specification, that behavioural and performance requirements appear to have been met. The software can be tested by one of the two ways: -

A. Knowing the specified function that a product has been designed to perform, tests can be conducted that demonstrate each function is fully operational while at the same time searching for errors in each function.
B. Knowing the internal working of the product, tests can be conducted to ensure that internal operations are performed according to specifications and all internal components have been adequately exercised.

The first approach is called white – box testing and the second, black – box testing.

7.1.1. WHITE BOX TESTING

White box testing is a test case design method that uses the control structural of the procedural design to derive test cases. It is also called glass box testing. Using this method, we can derive test cases that-

a) Guarantee that all independent paths within a module have been exercised at least once.
b) Exercise all logical decisions on their true and false sides Executes all loops at their boundaries and within their operational bounds.
c) Execute all loops at their boundaries and within their operational bounds.
d) Exercise internal data structures to ensure their validity.

What is White Box Testing?

The purpose of any security testing method is to ensure the robustness of a system in the face of malicious attacks or regular software failures. White box testing is performed based on the knowledge of *how* the system is implemented. White box testing includes analysing data flow, control flow, information flow, coding practices, and exception and error handling within the system, to test the intended and unintended software behaviour. White box testing can be performed to validate whether code implementation follows intended design, to validate implemented security functionality, and to uncover exploitable vulnerabilities.

White box testing requires access to the source code. Though white box testing can be performed any time in the life cycle after the code is developed, it is a good practice to perform white box testing during the unit testing phase.

7.1.2. BLACK BOX TESTING

Black – box testing focuses on the functional requirements of the software i.e. it enables the software engineer to derive sets of input conditions that will fully exercise all functional requirements for a program. It is also called behavioural testing.

1. It attempts to find errors in the following areas:
2. Incorrect or missing functions
3. Interface errors
4. Errors in data structures or external database access
5. Performance errors
6. Initialization errors

This software is developing as a product to be used by many customers, it is impractical to perform formal acceptance tests with each one. So out software product builders will use a process called Alpha and Beta testing to uncover errors that only the end-user seems able to find.

Advantages and Disadvantages: -

Advantages of Black Box Testing -

- more effective on larger units of code than glass box testing
- tester needs no knowledge of implementation, including specific programming languages
- tester and programmer are independent of each other
- tests are done from a user's point of view
- will help to expose any ambiguities or inconsistencies in the specifications
- test cases can be designed as soon as the specifications are complete

Disadvantages of Black Box Testing-

1. Only a small number of possible inputs can actually be tested, to test every possible input stream would take nearly forever
2. Without clear and concise specifications, test cases are hard to design
3. There may be unnecessary repetition of test inputs if the tester is not informed of test cases the programmer has already tried
4. May leave many program paths untested
5. Cannot be directed toward specific segments of code which may be very complex (and therefore more error prone)
6. most testing related research has been directed toward glass box testing

A. Alpha Testing: -

The Alpha test is conducted at the developer 's site by a customer. The software is used in a natural setting with the developer ― looking over the shoulder‖ of the user and recording errors and usage problem. Alpha test is conducted in controlled environment.

- Beta Testing: -The Beta-test is conducted at one or more customer sites by the end-user of the software. Unlike Alpha testing, the developer is generally not present. Therefore, the Beta-test is a ―live‖ application of the software in an environment that cannot be controlled by the developer. The customer records all problems (real or imagined) that are encountered during Beta testing and reports there to the developer at regular intervals. As a result of problems reported during Beta-tests, software engineers make modifications and then prepare for release of the software product to the entire customer base.

In this project we will performed incremental testing in which components and subsystem of the system are tested separately before integrating them to form the system from system testing.

7.2 TESTING STRATEGIES

Designing effective test cases is important but so is the strategy we use to execute them. A strategy for software test case design methods that result in the successful construction of software. The strategy provides a road map that describes the steps to be conducted as a part of testing.

There are a number of testing strategies, which have the following generic characteristics: -

1. Testing begins at the component level and works ―outward‖ toward the integration of the entire computer–based system.
2. Different testing techniques are approximate at different points in time.

3. Testing is conducted by the developer of the software and (for large projects) an independent test group.
4. Testing and debugging are different activities, but debugging must be accommodated in any testing strategy.

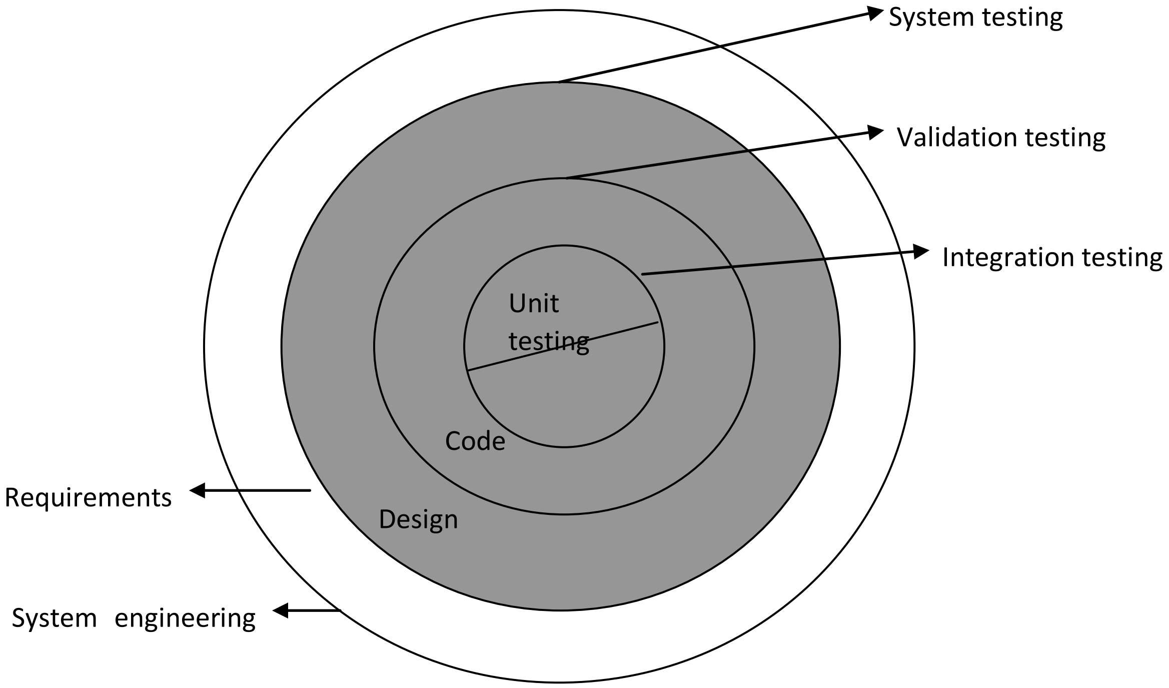


Figure 7.1: Testing techniques

7.3 Testing Strategy

Initially, system engineering defines the role of software and leads to software requirement analysis where the information domain, function, behaviours, performance, constraints and validation criteria for software are established. Moving inward along the spiral, we come to the design and finally to coding. There are a number of testing strategies as shown in the figure 7, which are given below: -

A. Unit testing: -

In the unit testing interfaces, local data structures, boundary conditions, independent paths, error-handling paths are tested. Test cases should be design to uncover errors due to erroneous computations, incorrect comparisons, or improper control flow. For this purpose, basis path and loop testing is done. After source level code has been developed, reviewed and verified for correspondence to component level design, unit test case design begins. In unit test application
‗drivers ‘are developed which are programs, accept test case data, passes such data to the component to be tested and prints relevant results. ‗Stubs ‘are also developed which serve to replace modules, that are subordinate the component to be tested.

Six Rules of Unit Testing: -

1. Write the test first
2. Never write a test that succeeds the first time
3. Start with the null case, or something that doesn't work
4. Don't be afraid of doing something trivial to make the test work
5. Loose coupling and testability go hand in hand
6. Use mock objects

Limitations of unit testing: -

1. Testing, in general, cannot be expected to catch every error in the program. The same is true for unit testing. By definition, it only tests the functionality of the units themselves. Therefore, it may not catch integration errors, performance problems, or other system-wide issues. Unit testing is more effective if it is used in conjunction with other software testing activities.
2. Like all forms of software testing, unit tests can only show the presence of errors; it cannot show the absence of errors.
3. Software testing is a combinatorial problem. For example, every Boolean decision statement requires at least two tests: one with an outcome of "true" and one with an outcome of "false". As a result, for every line of code written, programmers often need 3 to 5 lines of test code. Therefore, it is unrealistic to test all possible input combinations for any non-trivial piece of software without an automated characterization test generation tool such as JUnit Factory used with Java code or many of the tools listed in List of unit testing frameworks.
4. To obtain the intended benefits from unit testing, a rigorous sense of discipline is needed throughout the software development process. It is essential to keep careful records, not only of the tests that have been performed, but also of all changes that have been made to the source code of this or any other unit in the

software. Use of a version control system is essential. If a later version of the unit fails a particular test that it had previously passed, the version-control software can provide a list of the source code changes (if any) that have been applied to the unit since that time.

5. It is also essential to implement a sustainable process for ensuring that test case failures are reviewed daily and addressed immediately. If such a process is not implemented and ingrained into the team's workflow, the application will evolve out of sync with the unit test suite—- increasing false positives and reducing the effectiveness of the test suite.

B. **Integration testing: -**

Integration testing is systematic technique for constructing the program structure while at the same time conducting the tests to uncover errors associated with interfacing. The objective is to take unit tested components and build a program structure that has been dictated by design. There are two types of integration – Bottom up integration and Top down integration. Regression and smoke testing are done in integration testing strategy.

C. **Validation testing: -**

Next step is the validation testing where requirements established as part of software requirements analysis are validated against the software that has been constructed. At the culmination of integration testing, software is completely assembled as a package, interfacing errors has been uncovered and corrected, and a final series of software tests i.e. validation testing begins.

D. **System testing: -**

Finally, we arrive at system testing where the software and other system elements are tested as a whole. System testing verifies that all elements mesh properly and that overall system function / performance is achieved. Ultimately software is incorporated with other system elements and a series of system integration and validation tests are conducted.

7.4 DEBUGGING AND CODE IMPROVEMENT

Since the modular approach is used to design the software. This approach uses the object oriented design methods. Hence the code is improved due to the above reason. Because of object oriented programming, the features like modularity and reusability can be achieved in the software.

In this project we have used various controls like command buttons, text box and similar other elements from various class.

The process of testing gives symptoms, and a program ‘s failure is a clear symptom of the presence of the error. After getting a symptom, we begin an investigation to localize the error, that is to find out which module or interface is causing it. Then that section of the code is to be studied to determine the cause of the problem. This process is called _Debugging ‘. Hence, debugging is the activity of locating and correcting errors.

The following errors has been debugged during the creation of the project-

Table 7.1 Errors Debugged

S. No.	Bug Description	Cause of Bug	Time required to remove it
1.	Update error	Database Connectivity Problem	2 Hours
2	Retrieve error	Database Error	20 minute
3	Database Insert Error	Data Type not match	1 Hours
4.	Record Not Found	Database Error	10 Minutes
5.	Duplicate entry error	It occurred because another record exist with same id	1 Hour
6.	Null pointer exception	A method of query class getting null as input	45 Minutes
7.	No such method error	It occurred because, a non-existent method was called from.	30 minutes

Code improvement: -

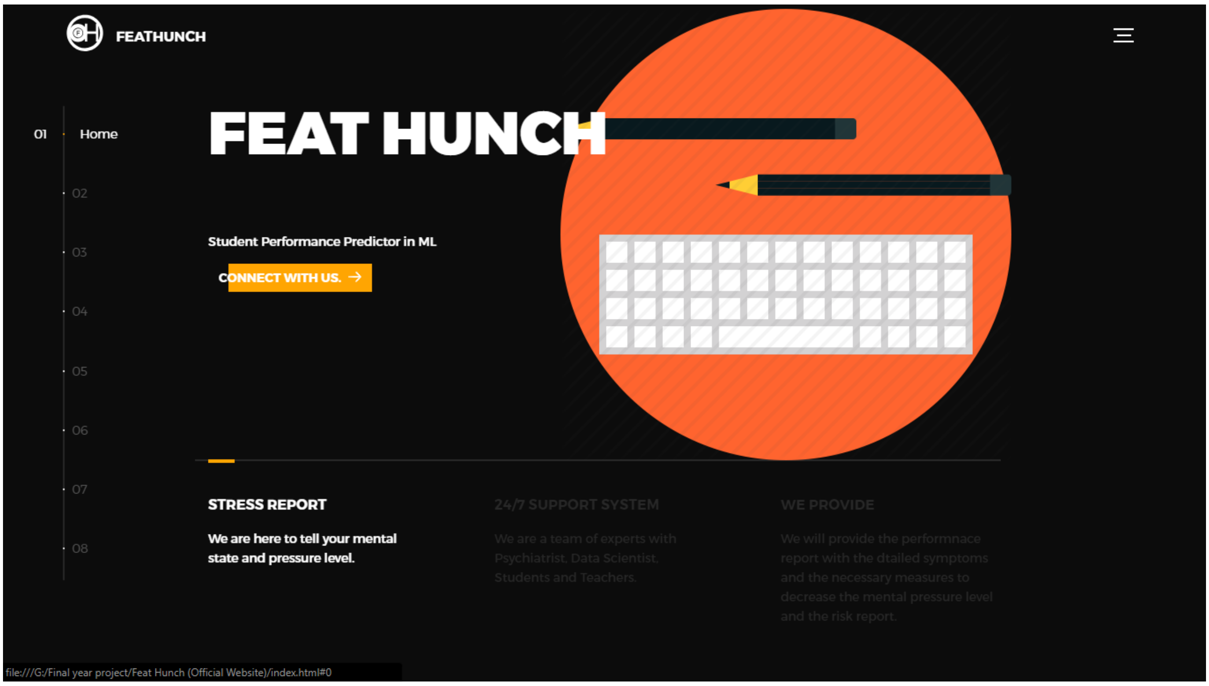
After debugging, it has been that there is some problem in coding and hence to rectify the error occurred during the testing the improvement in the code is necessary. Following improvement has been done in coding after testing: -

- Error prone code has been removed
- All the missing classes has been properly imported
- Automatic table creation code has been added
- More accurate SQL query has been written

CHAPTER-8

SCREENSORT OF THE PROJECT

8.1 HOME PAGE 1- This is page is the opening page of the project.



FEAT HUNCH

Student Performance Predictor in ML

STRESS REPORT

We are here to tell your mental state and pressure level.

24/7 SUPPORT SYSTEM

We are a team of experts with Psychiatrist, Data Scientist, Students and Teachers.

WE PROVIDE

We will provide the performnace report with the dtailed symptoms and the necessary measures to decrease the mental pressure level and the risk report.

file:///G:/Final year project/Feat Hunch (Official Website)/index.html#0

Figure 8.1: Home Page 1

8.2 HOME PAGE 2

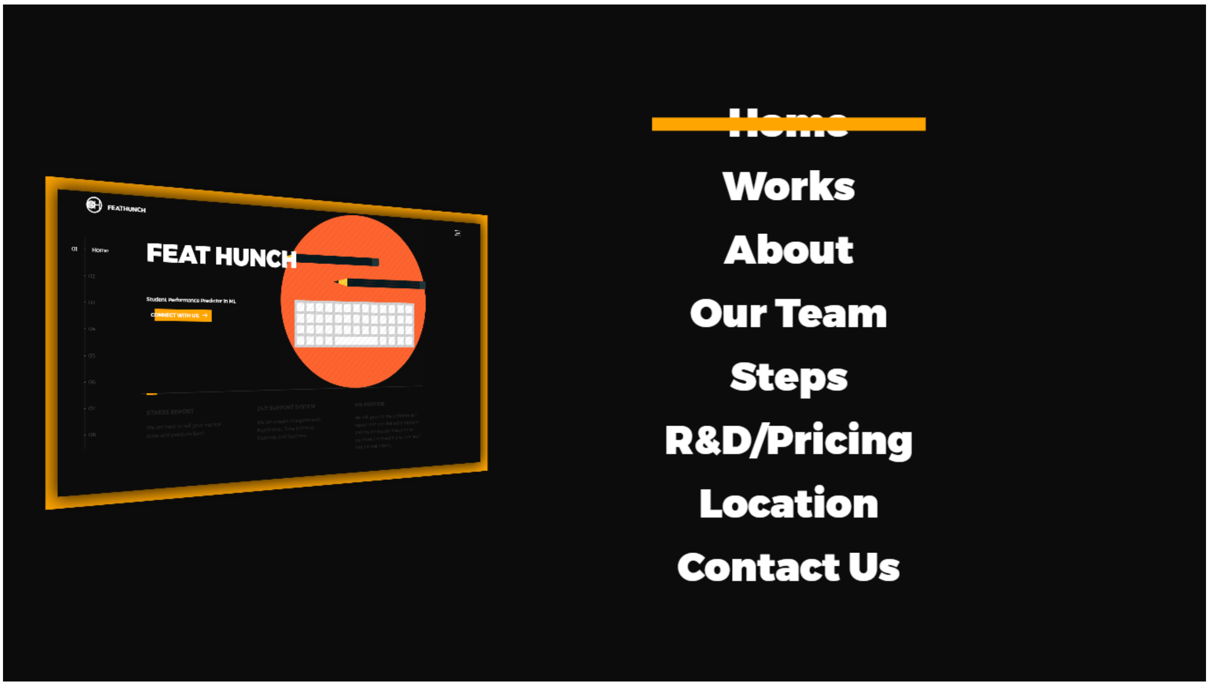
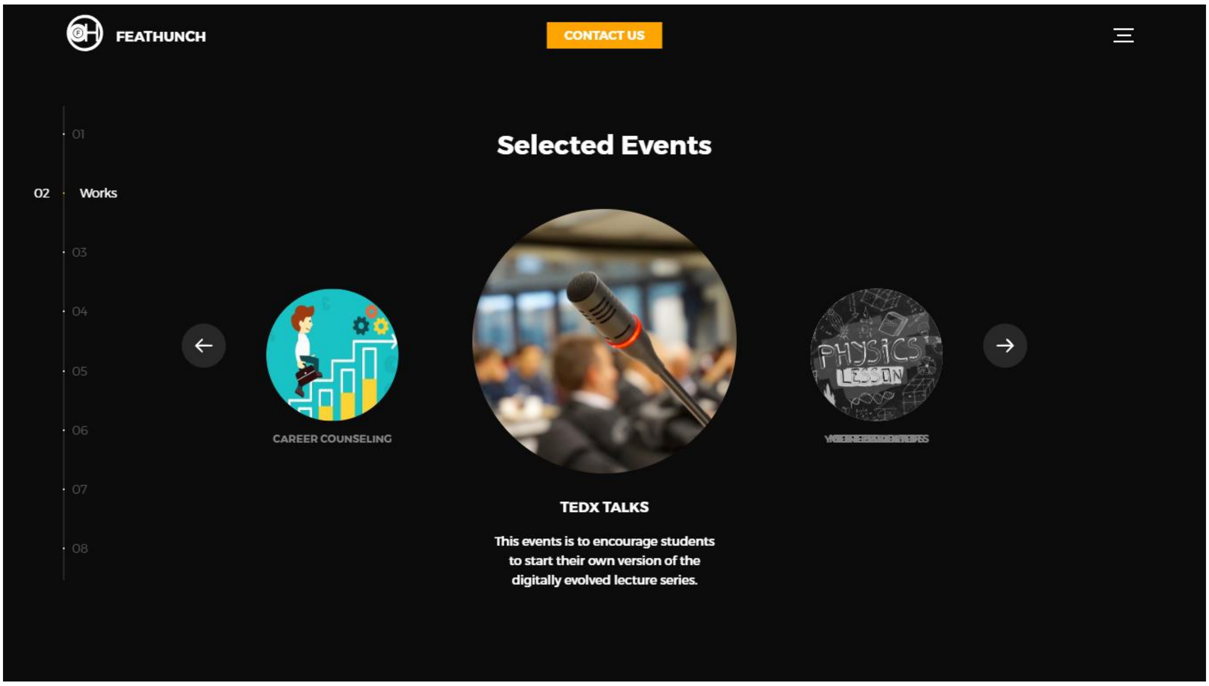


Figure 8.2: Home Page 2

8.3 EVENT PAGE-



This events is to encourage students to start their own version of the digitally evolved lecture series.

Figure 8.3: Event Page

8.4 ABOUT PAGE-

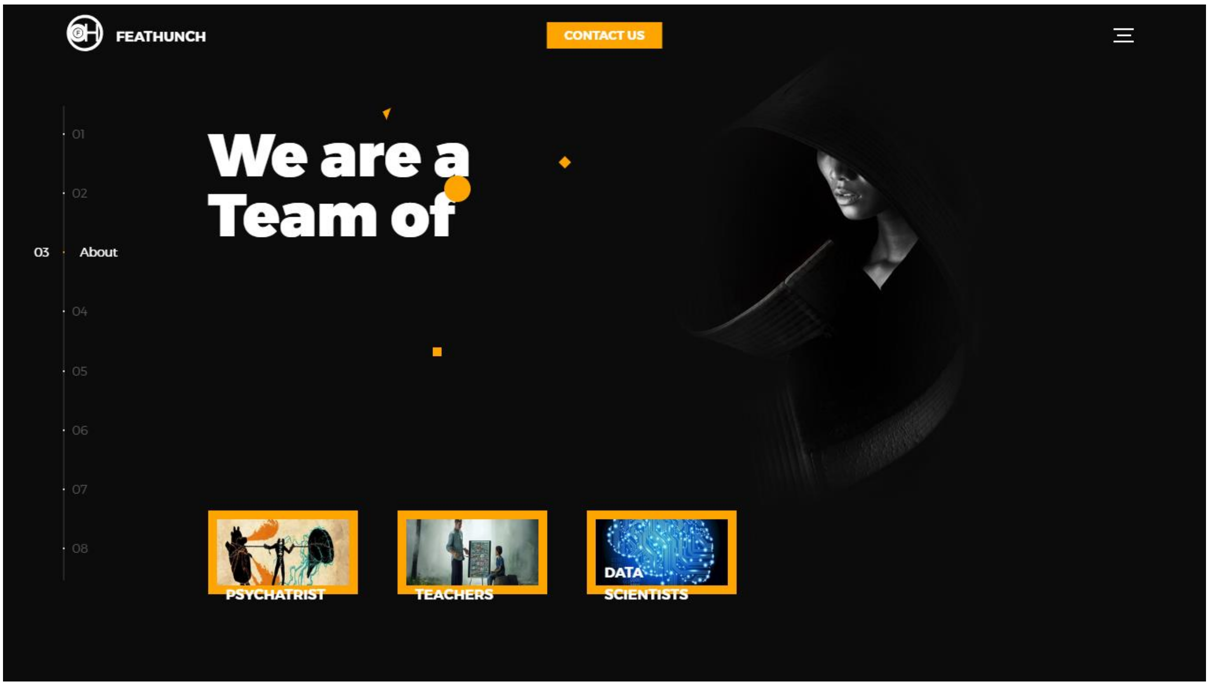


Figure 8.4: About Page

8.5 STEPS-

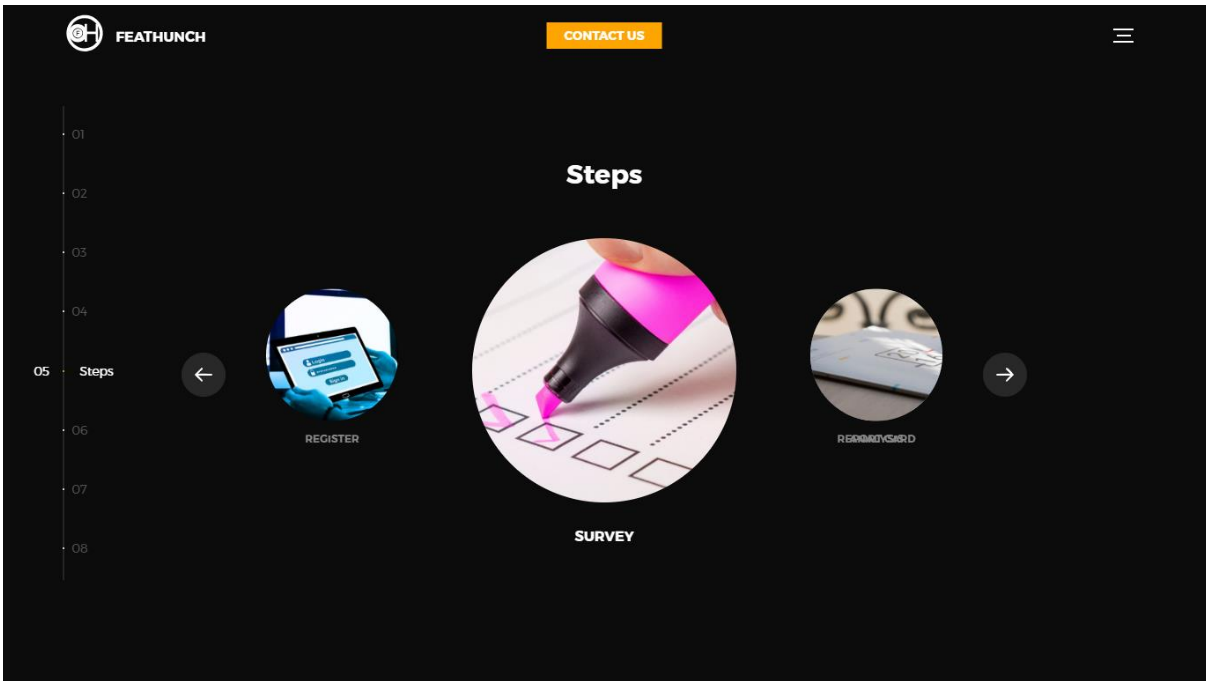


Figure 8.5: Steps

8.6 OUR TEAM-

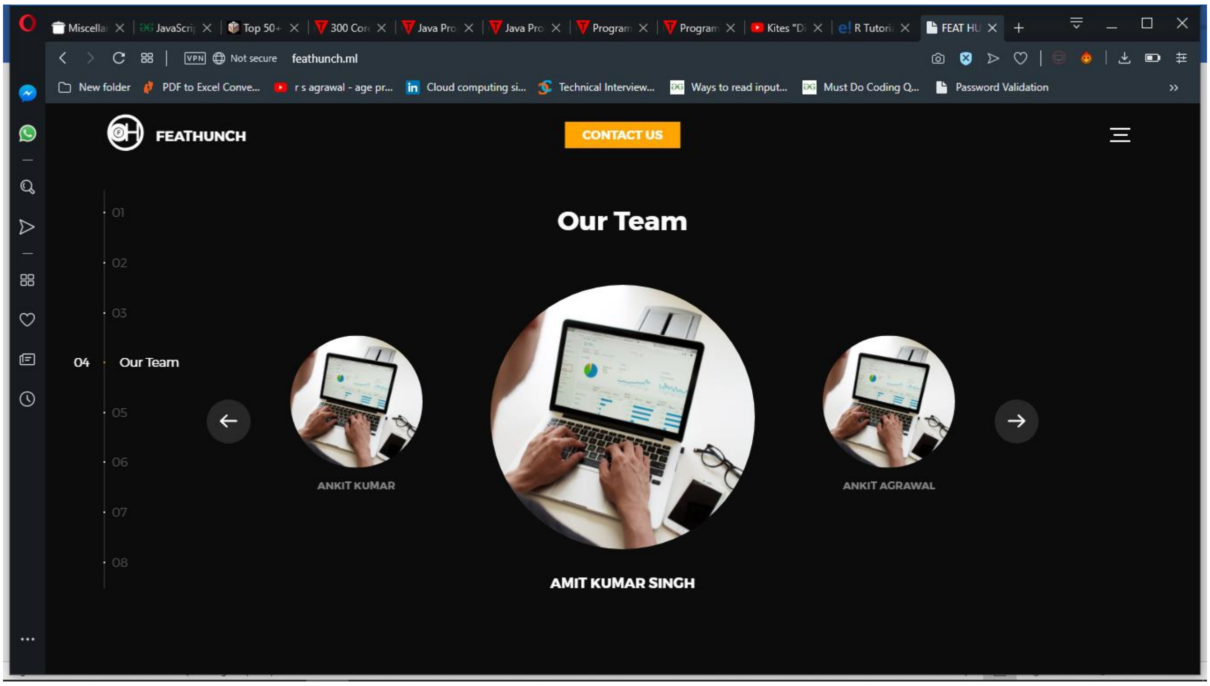


Figure 8.6: Our Team

8.7 FUTURE PLANS-

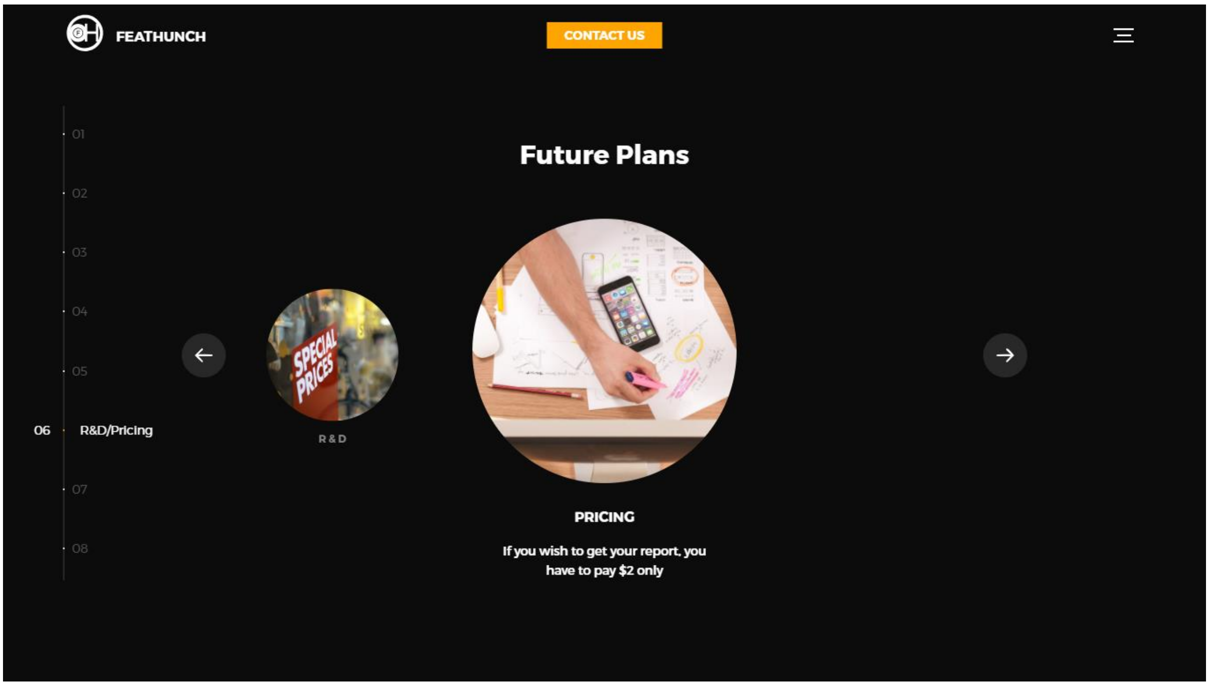


Figure 8.7: Future Plans

8.8 CONTACT US-

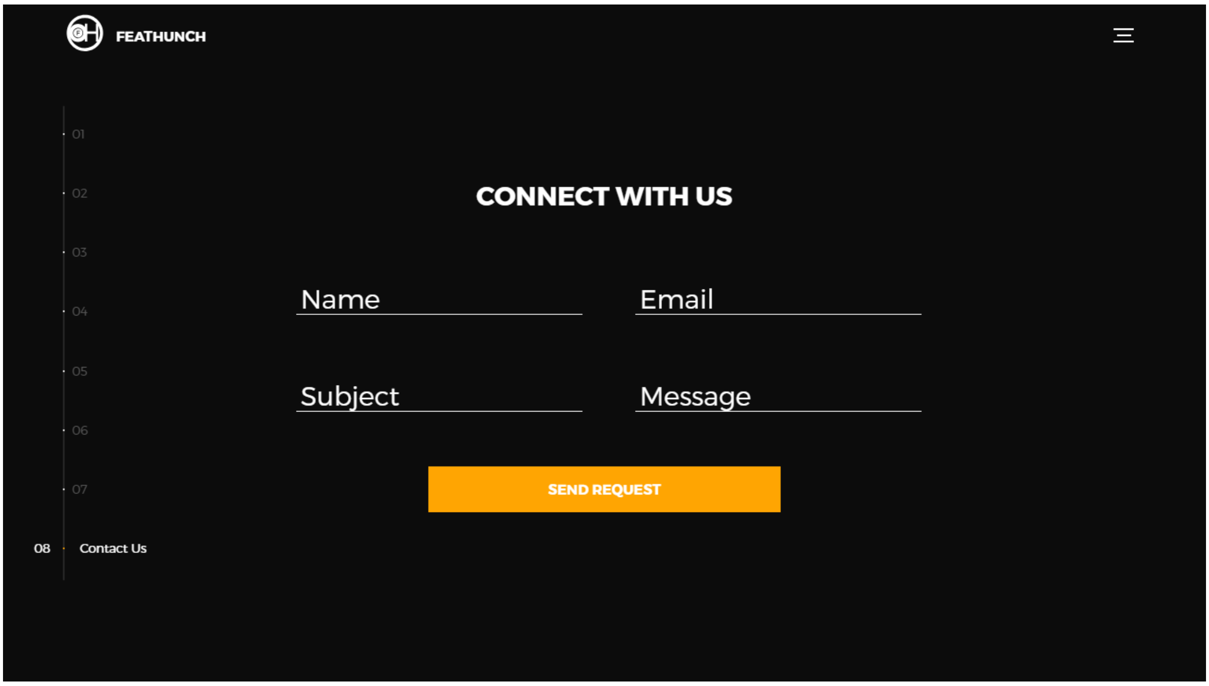
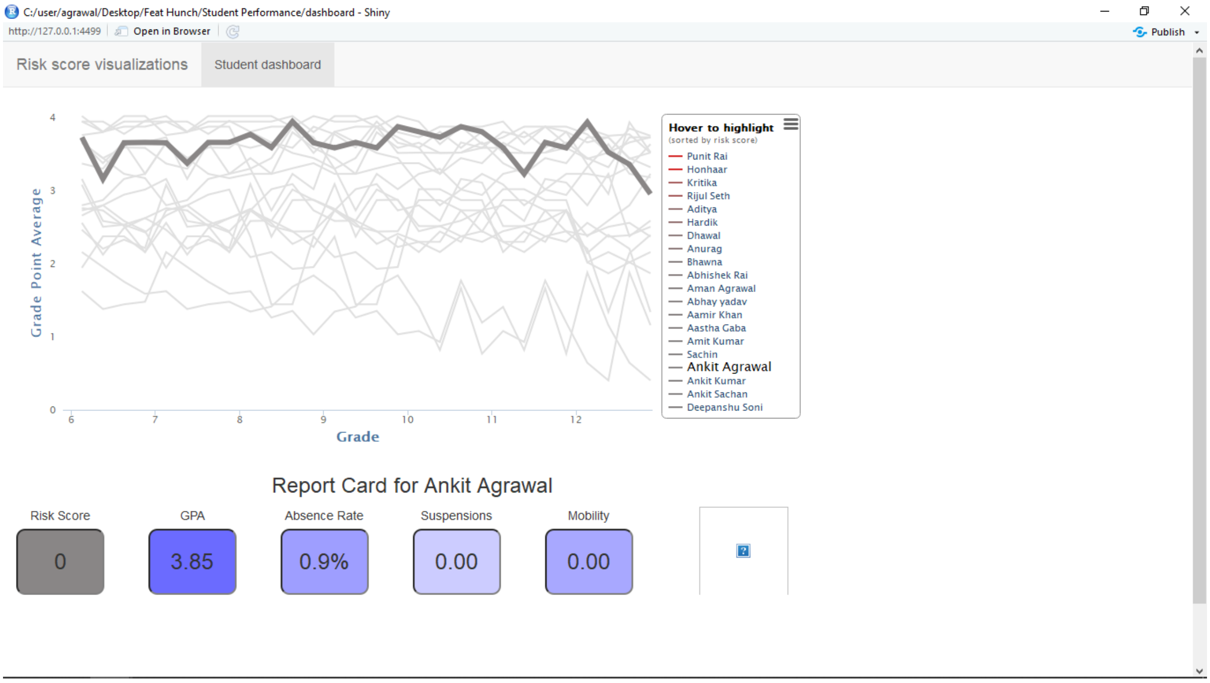


Figure 8.8: Contact Us

8.9 DASHBOARD 1-

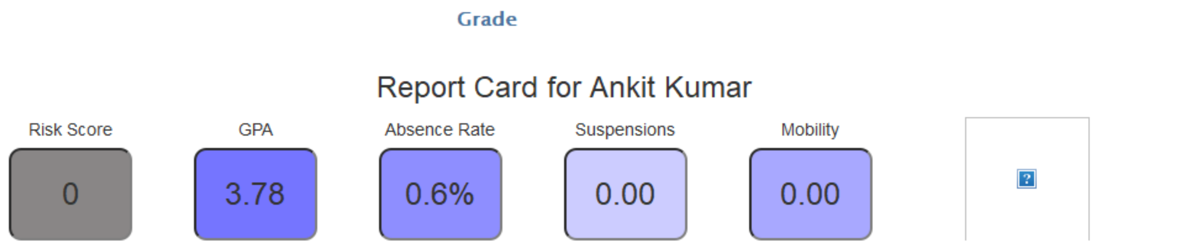


Report Card for Ankit Agrawal

Risk Score	GPA	Absence Rate	Suspensions	Mobility
0	3.85	0.9%	0.00	0.00

Figure 8.9: Dashboard 1

8.10 DASHBOARD 2-

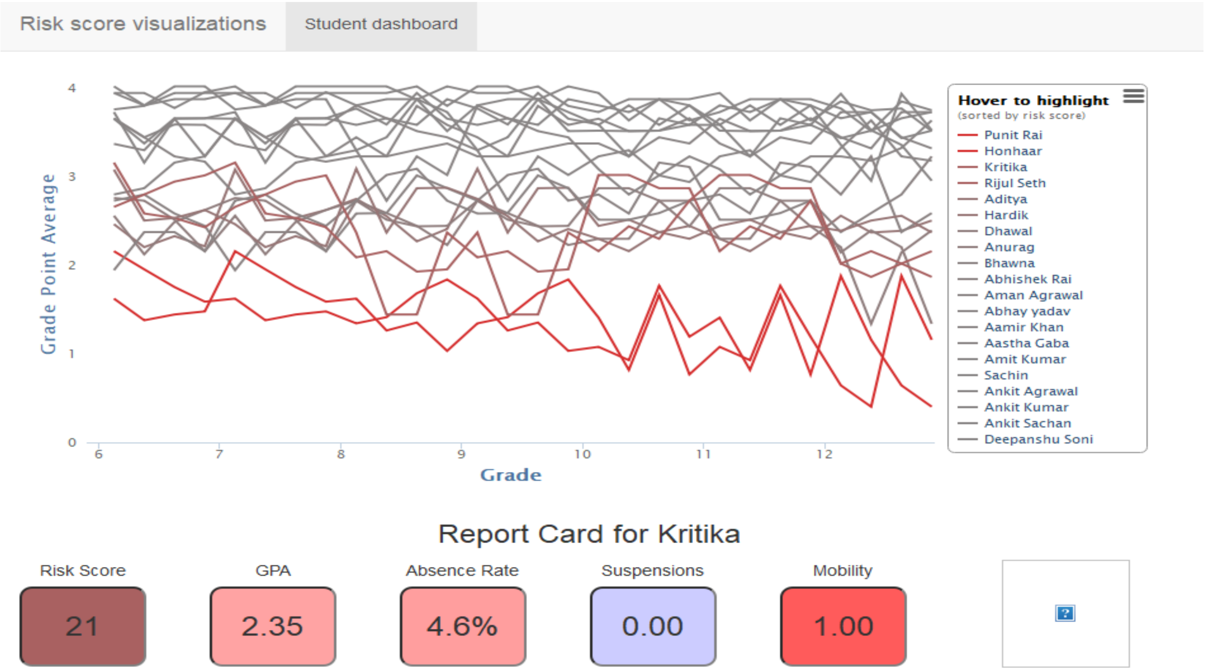


Report Card for Ankit Kumar

Risk Score	GPA	Absence Rate	Suspensions	Mobility
0	3.78	0.6%	0.00	0.00

Figure 8.10: Dashboard 2

8.11 DASHBOARD 3-

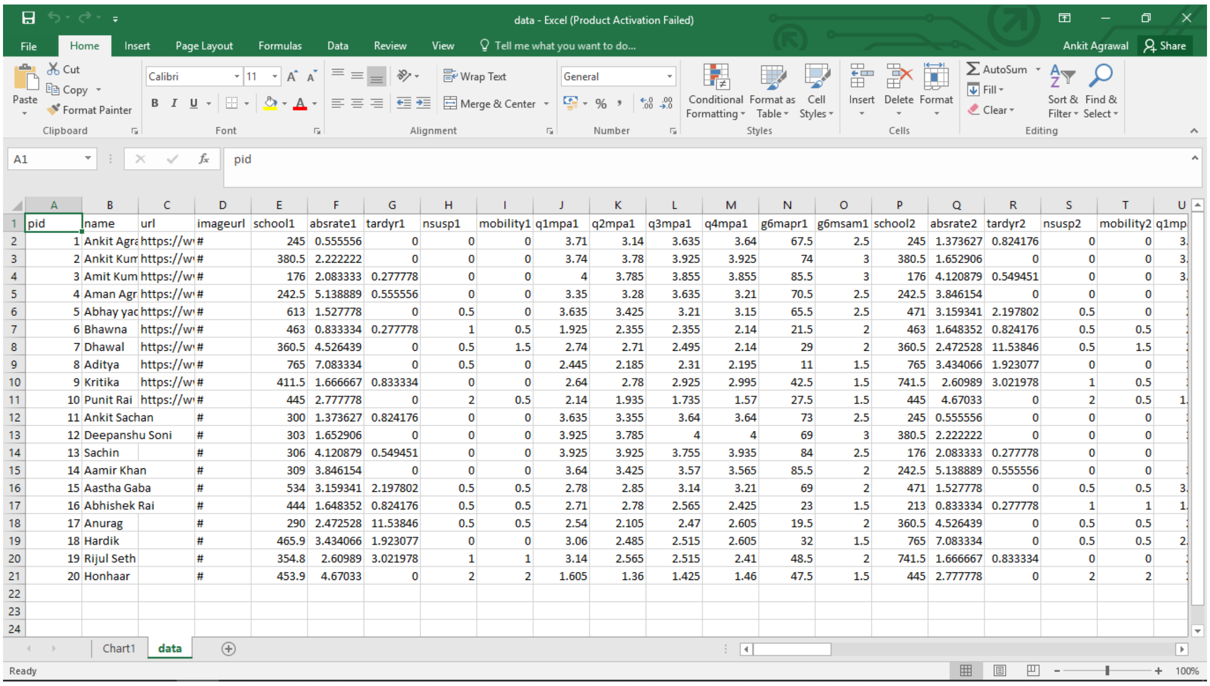


Report Card for Kritika

Risk Score	GPA	Absence Rate	Suspensions	Mobility
21	2.35	4.6%	0.00	1.00

Figure 8.11: Dashboard 3

8.12 DATABASE 1 -

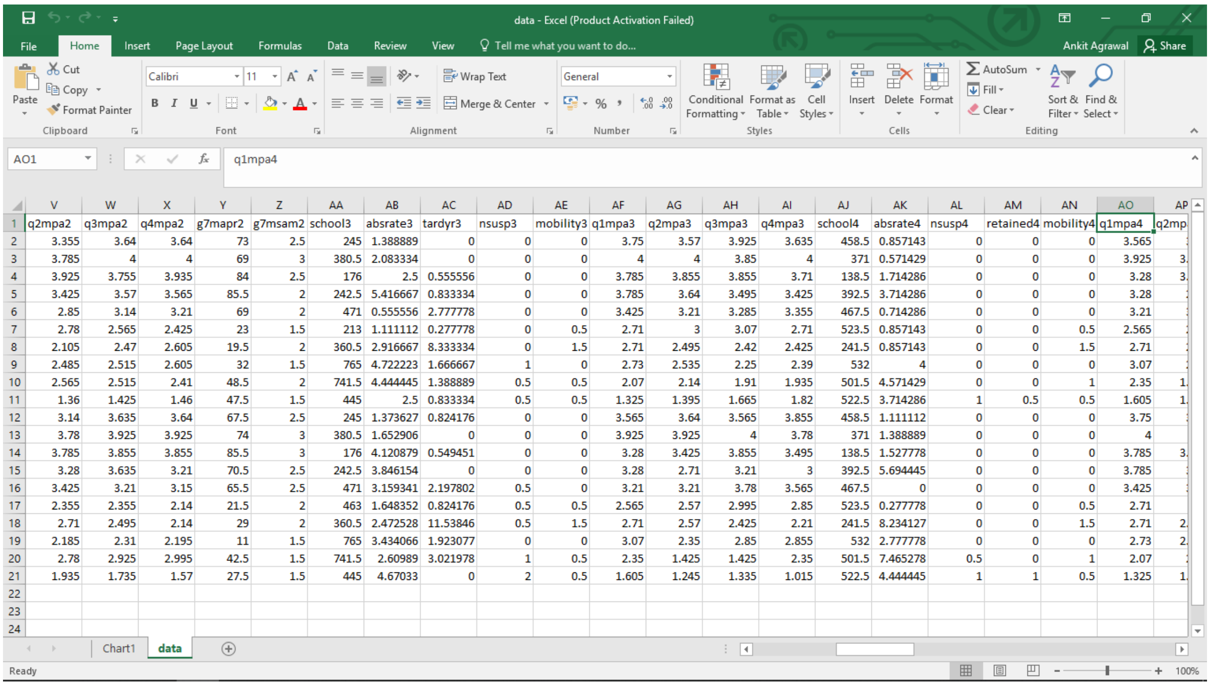


A1 pid

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	pid	name	url	imageurl	school1	absrate1	tardyr1	nsusp1	mobility1	q1mpa1	q2mpa1	q3mpa1	q4mpa1	g6mapr1	g6msam1	school2	absrate2	tardyr2	nsusp2	mobility2	q1mp
2	1	Ankit Agra	https://w	#	245	0.555556	0	0	0	3.71	3.14	3.635	3.64	67.5	2.5	245	1.373627	0.824176	0	0	3.
3	2	Ankit Kum	https://w	#	380.5	2.222222	0	0	0	3.74	3.78	3.925	3.925	74	3	380.5	1.652906	0	0	0	3.
4	3	Amit Kum	https://w	#	176	2.083333	0.277778	0	0	4	3.785	3.855	3.855	85.5	3	176	4.120879	0.549451	0	0	3.
5	4	Aman Agr	https://w	#	242.5	5.138889	0.555556	0	0	3.35	3.28	3.635	3.21	70.5	2.5	242.5	3.846154	0	0	0	
6	5	Abhay yac	https://w	#	613	1.527778	0	0.5	0	3.635	3.425	3.21	3.15	65.5	2.5	471	3.159341	2.197802	0.5	0	
7	6	Bhawna	https://w	#	463	0.833334	0.277778	1	0.5	1.925	2.355	2.355	2.14	21.5	2	463	1.648352	0.824176	0.5	0.5	
8	7	Dhawal	https://w	#	360.5	4.526439	0	0.5	1.5	2.74	2.71	2.495	2.14	29	2	360.5	2.472528	11.53846	0.5	1.5	
9	8	Aditya	https://w	#	765	7.083334	0	0.5	0	2.445	2.185	2.31	2.195	11	1.5	765	3.434066	1.923077	0	0	
10	9	Kritika	https://w	#	411.5	1.666667	0.833334	0	0	2.64	2.78	2.925	2.995	42.5	1.5	741.5	2.60989	3.021978	1	0.5	
11	10	Punit Rai	https://w	#	445	2.777778	0	2	0.5	2.14	1.935	1.735	1.57	27.5	1.5	445	4.67033	0	2	0.5	1.
12	11	Ankit Sachan		#	300	1.373627	0.824176	0	0	3.635	3.355	3.64	3.64	73	2.5	245	0.555556	0	0	0	
13	12	Deepanshu Soni		#	303	1.652906	0	0	0	3.925	3.785	4	4	69	3	380.5	2.222222	0	0	0	
14	13	Sachin		#	306	4.120879	0.549451	0	0	3.925	3.925	3.755	3.935	84	2.5	176	2.083333	0.277778	0	0	
15	14	Aamir Khan		#	309	3.846154	0	0	0	3.64	3.425	3.57	3.565	85.5	2	242.5	5.138889	0.555556	0	0	
16	15	Aastha Gaba		#	534	3.159341	2.197802	0.5	0.5	2.78	2.85	3.14	3.21	69	2	471	1.527778	0	0.5	0.5	3.
17	16	Abhishek Rai		#	444	1.648352	0.824176	0.5	0.5	2.71	2.78	2.565	2.425	23	1.5	213	0.833334	0.277778	1	1	1.
18	17	Anurag		#	290	2.472528	11.53846	0.5	0.5	2.54	2.105	2.47	2.605	19.5	2	360.5	4.526439	0	0.5	0.5	
19	18	Hardik		#	465.9	3.434066	1.923077	0	0	3.06	2.485	2.515	2.605	32	1.5	765	7.083334	0	0.5	0.5	2.
20	19	Rijul Seth		#	354.8	2.60989	3.021978	1	1	3.14	2.565	2.515	2.41	48.5	2	741.5	1.666667	0.833334	0	0	
21	20	Honhaar		#	453.9	4.67033	0	2	2	1.605	1.36	1.425	1.46	47.5	1.5	445	2.777778	0	2	2	
22																					
23																					
24																					

Figure 8.12: Database 1

8.13 DATABASE 2-

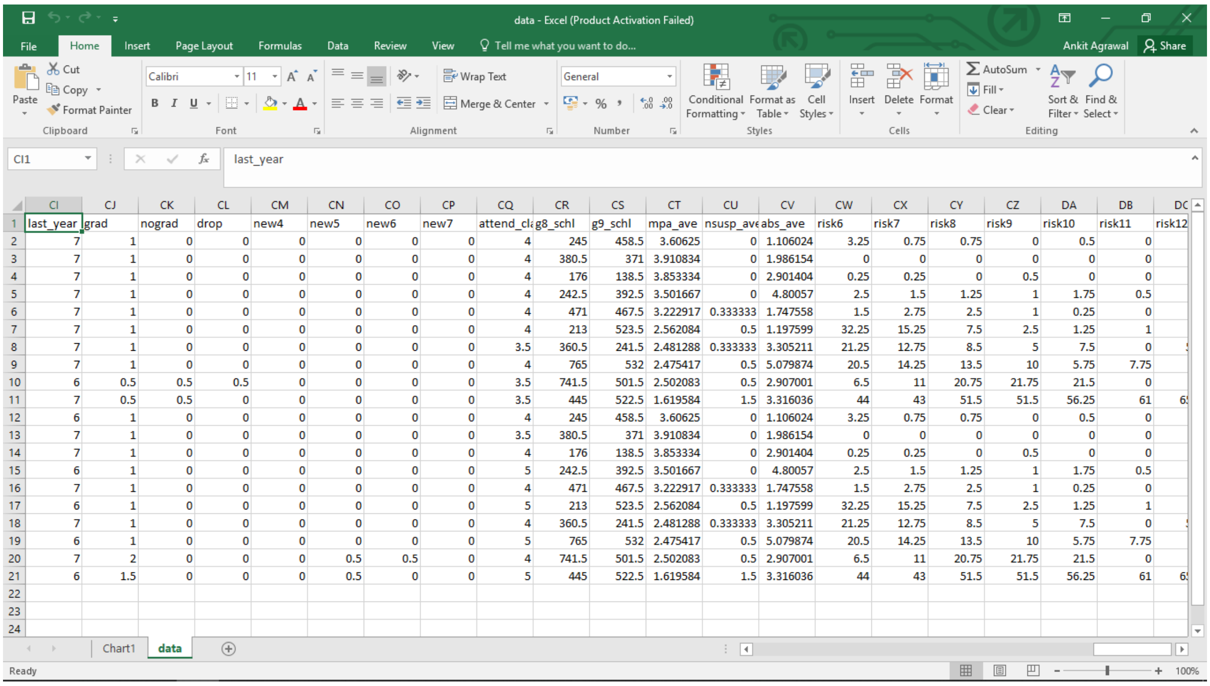


AO1 q1mpa4

	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
1	q2mpa2	q3mpa2	q4mpa2	g7mapr2	g7msam2	school3	absrate3	tardyr3	nsusp3	mobility3	q1mpa3	q2mpa3	q3mpa3	q4mpa3	school4	absrate4	nsusp4	retained4	mobility4	q1mpa4	q2mp
2	3.355	3.64	3.64	73	2.5	245	1.388889	0	0	0	3.75	3.57	3.925	3.635	458.5	0.857143	0	0	0	3.565	
3	3.785	4	4	69	3	380.5	2.083334	0	0	0	4	4	3.85	4	371	0.571429	0	0	0	3.925	3.
4	3.925	3.755	3.935	84	2.5	176	2.5	0.555556	0	0	3.785	3.855	3.855	3.71	138.5	1.714286	0	0	0	3.28	3.
5	3.425	3.57	3.565	85.5	2	242.5	5.416667	0.833334	0	0	3.785	3.64	3.495	3.425	392.5	3.714286	0	0	0	3.28	
6	2.85	3.14	3.21	69	2	471	0.555556	2.777778	0	0	3.425	3.21	3.285	3.355	467.5	0.714286	0	0	0	3.21	
7	2.78	2.565	2.425	23	1.5	213	1.111112	0.277778	0	0.5	2.71	3	3.07	2.71	523.5	0.857143	0	0	0.5	2.565	
8	2.105	2.47	2.605	19.5	2	360.5	2.916667	8.333334	0	1.5	2.71	2.495	2.42	2.425	241.5	0.857143	0	0	1.5	2.71	
9	2.485	2.515	2.605	32	1.5	765	4.722223	1.666667	1	0	2.73	2.535	2.25	2.39	532	4	0	0	0	3.07	
10	2.565	2.515	2.41	48.5	2	741.5	4.444445	1.388889	0.5	0.5	2.07	2.14	1.91	1.935	501.5	4.571429	0	0	1	2.35	1.
11	1.36	1.425	1.46	47.5	1.5	445	2.5	0.833334	0.5	0.5	1.325	1.395	1.665	1.82	522.5	3.714286	1	0.5	0.5	1.605	1.
12	3.14	3.635	3.64	67.5	2.5	245	1.373627	0.824176	0	0	3.565	3.64	3.565	3.855	458.5	1.111112	0	0	0	3.75	
13	3.78	3.925	3.925	74	3	380.5	1.652906	0	0	0	3.925	3.925	4	3.78	371	1.388889	0	0	0	4	
14	3.785	3.855	3.855	85.5	3	176	4.120879	0.549451	0	0	3.28	3.425	3.855	3.495	138.5	1.527778	0	0	0	3.785	3.
15	3.28	3.635	3.21	70.5	2.5	242.5	3.846154	0	0	0	3.28	2.71	3.21	3	392.5	5.694445	0	0	0	3.785	
16	3.425	3.21	3.15	65.5	2.5	471	3.159341	2.197802	0.5	0	3.21	3.21	3.78	3.565	467.5	0	0	0	0	3.425	
17	2.355	2.355	2.14	21.5	2	463	1.648352	0.824176	0.5	0.5	2.565	2.57	2.995	2.85	523.5	0.277778	0	0	0.5	2.71	
18	2.71	2.495	2.14	29	2	360.5	2.472528	11.53846	0.5	1.5	2.71	2.57	2.425	2.21	241.5	8.234127	0	0	1.5	2.71	2.
19	2.185	2.31	2.195	11	1.5	765	3.434066	1.923077	0	0	3.07	2.35	2.85	2.855	532	2.777778	0	0	0	2.73	2.
20	2.78	2.925	2.995	42.5	1.5	741.5	2.60989	3.021978	1	0.5	2.35	1.425	1.425	2.35	501.5	7.465278	0.5	0	1	2.07	
21	1.935	1.735	1.57	27.5	1.5	445	4.67033	0	2	0.5	1.605	1.245	1.335	1.015	522.5	4.444445	1	1	0.5	1.325	1.
22																					
23																					
24																					

Figure 8.13: Database 2

8.14 DATABASE 3-



CI1 | last_year

	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC
1	last_year	grad	nograd	drop	new4	new5	new6	new7	attend_cl	g8_schl	g9_schl	mpa_ave	nsusp_ave	abs_ave	risk6	risk7	risk8	risk9	risk10	risk11	risk12
2	7	1	0	0	0	0	0	0	4	245	458.5	3.60625	0	1.106024	3.25	0.75	0.75	0	0.5	0	
3	7	1	0	0	0	0	0	0	4	380.5	371	3.910834	0	1.986154	0	0	0	0	0	0	
4	7	1	0	0	0	0	0	0	4	176	138.5	3.853334	0	2.901404	0.25	0.25	0	0.5	0	0	
5	7	1	0	0	0	0	0	0	4	242.5	392.5	3.501667	0	4.80057	2.5	1.5	1.25	1	1.75	0.5	
6	7	1	0	0	0	0	0	0	4	471	467.5	3.222917	0.333333	1.747558	1.5	2.75	2.5	1	0.25	0	
7	7	1	0	0	0	0	0	0	4	213	523.5	2.562084	0.5	1.197599	32.25	15.25	7.5	2.5	1.25	1	
8	7	1	0	0	0	0	0	0	3.5	360.5	241.5	2.481288	0.333333	3.305211	21.25	12.75	8.5	5	7.5	0	5
9	7	1	0	0	0	0	0	0	4	765	532	2.475417	0.5	5.079874	20.5	14.25	13.5	10	5.75	7.75	
10	6	0.5	0.5	0.5	0	0	0	0	3.5	741.5	501.5	2.502083	0.5	2.907001	6.5	11	20.75	21.75	21.5	0	
11	7	0.5	0.5	0	0	0	0	0	3.5	445	522.5	1.619584	1.5	3.316036	44	43	51.5	51.5	56.25	61	6
12	6	1	0	0	0	0	0	0	4	245	458.5	3.60625	0	1.106024	3.25	0.75	0.75	0	0.5	0	
13	7	1	0	0	0	0	0	0	3.5	380.5	371	3.910834	0	1.986154	0	0	0	0	0	0	
14	7	1	0	0	0	0	0	0	4	176	138.5	3.853334	0	2.901404	0.25	0.25	0	0.5	0	0	
15	6	1	0	0	0	0	0	0	5	242.5	392.5	3.501667	0	4.80057	2.5	1.5	1.25	1	1.75	0.5	
16	7	1	0	0	0	0	0	0	4	471	467.5	3.222917	0.333333	1.747558	1.5	2.75	2.5	1	0.25	0	
17	6	1	0	0	0	0	0	0	5	213	523.5	2.562084	0.5	1.197599	32.25	15.25	7.5	2.5	1.25	1	
18	7	1	0	0	0	0	0	0	4	360.5	241.5	2.481288	0.333333	3.305211	21.25	12.75	8.5	5	7.5	0	5
19	6	1	0	0	0	0	0	0	5	765	532	2.475417	0.5	5.079874	20.5	14.25	13.5	10	5.75	7.75	
20	7	2	0	0	0	0.5	0.5	0	4	741.5	501.5	2.502083	0.5	2.907001	6.5	11	20.75	21.75	21.5	0	
21	6	1.5	0	0	0	0.5	0	0	5	445	522.5	1.619584	1.5	3.316036	44	43	51.5	51.5	56.25	61	6
22																					
23																					
24																					

Figure 8.14: Database 3

References

[1.] Shaukat, K., Nawaz, I., Aslam, S., Zaheer, S., & Shaukat, U. (2016, December). Student's performance in the context of data mining. In *2016 19th International Multi-Topic Conference (INMIC)* (pp. 1-8). IEEE.

[2.] Sikder, M. F., Uddin, M. J., & Halder, S. (2016, May). Predicting students yearly performance using neural network: A case study of BSMRSTU. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 524-529). IEEE.

[3.] Fok, W. W., Chen, H., Yi, J., Li, S., Yeung, H. A., Ying, W., & Fang, L. (2014, September). Data mining application of decision trees for student profiling at the Open University of China. In 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (pp. 732-738). IEEE.

[4.] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

[5.] Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. arXiv preprint arXiv:1002.2425.

[6.] Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 10(1), 3-12.

[7.] Mallik, P., Roy, C., Maheshwari, E., Pandey, M., & Rautray, S. (2019). Analyzing Student Performance Using Data Mining. In Ambient Communications and Computer Systems (pp. 307-318). Springer, Singapore.

[8.] Vidyavathi, B. Predicting Academic Performance of Students Using Data Mining Technique.

[9.] Sa, C. L., Hossain, E. D., & bin Hossin, M. (2014, November). Student performance analysis system (SPAS). In The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M) (pp. 1-6). IEEE.

[10.] Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. International journal of computer applications, 63(8), 35-39.

[11.] Suryadarma, D., Suryahadi, A., Sumarto, S., & Rogers, F. H. (2006). Improving student performance in public primary schools in developing countries: Evidence from Indonesia. Education Economics, 14(4), 401-429.

[12.] Chemers, M. M., Hu, L. T., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. Journal of Educational psychology, 93(1), 55.

[13.] Runeson, P. (2003, April). Using students as experiment subjects–an analysis on graduate and freshmen student data. In Proceedings of the 7th International Conference on Empirical Assessment in Software Engineering (pp. 95-102). Keele University UK.

[14.] Roy, C., Pandey, M., & Rautaray, S. S. (2018). A Proposal for Optimization of Horizontal Scaling in Big Data Environment. In Advances in Data and Information Sciences (pp. 223-230). Springer, Singapore.

[15.] Das, N., Das, L., Rautaray, S. S., & Pandey, M. (2018). Big data analytics for medical applications. International Journal of Modern Education and Computer Science, 11(2), 35.

[16.] Roy, C., Rautaray, S. S., & Pandey, M. (2018). Big Data Optimization Techniques: A Survey. International Journal of Information Engineering and Electronic Business, 10(4), 41.

[17.] Nonis, S. A., & Hudson, G. I. (2006). Academic performance of college students: Influence of time spent studying and working. Journal of education for business, 81(3), 151-159.