

# **PREDICTING STUDENT'S PERFORMANCE BASED ON MACHINE LEARNING**

Submitted in partial fulfillment of the requirements  
for the award of  
Bachelor of Engineering degree in Computer Science and Engineering

by

R.Manideep (38110296)  
P.V.Adarsh Kumar (38110011)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
SCHOOL OF COMPUTING**

## **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,  
CHENNAI - 600 119**

**MARCH – 2022**



## **SATHYABAMA**

INSTITUTE OF SCIENCE AND TECHNOLOGY

**(DEEMED TO BE UNIVERSITY)**

Accredited with “A” grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

[www.sathyabama.ac.in](http://www.sathyabama.ac.in)



**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**

### **BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **R.Manideep (38110296)** and **P.V.Adarsh Kumar (38110011)** who carried out the project entitled “**PREDICTING STUDENT'S PERFORMANCE BASED ON MACHINE LEARNING**” under my supervision from November 2021 to March 2022.

**Internal Guide**

MRS D.Deepa

**Head of the Department**

---

Submitted for Viva voce Examination held on \_\_\_\_\_

**Internal Examiner**

**External Examiner**

## **DECLARATION**

I **R.Manideep (Reg No:38110296)** and **P.V.Adarsh Kumar (Reg No: 38110011)** hereby declare that the Project Report entitled “**PREDICTING STUDENT'S PERFORMANCE BASED ON MACHINE LEARNING**” done by us under the guidance of **MRS D.Deepa** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

**DATE:**

**PLACE:**

**SIGNATURE OF THE CANDIDATE**

## ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D., Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D.** , Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **MRS D.Deepa** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

## **ABSTRACT**

In education system evaluation and prediction of student performance is a challenging task. In this paper, a model is proposed to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Naïve Bayes and KNN. Further, the importance of several different attributes, or "features" is considered, in order to determine which of these are correlated with student performance. Finally, the results of an experiment follow, showcasing the power of machine learning in such an application. In perspective of this project we are going to predict the student development and examine the greater result through machine learning algorithm. We foresee the student performance by scanning their previous academic details. To execute this prediction we have created a dataset, by using this we can predict student details.

## TABLE OF CONTENTS

| Chapter No. | TITLE                                 | Page No.  |
|-------------|---------------------------------------|-----------|
|             | <b>ABSTRACT</b>                       | v         |
|             | <b>LIST OF FIGURES</b>                | viii      |
|             | <b>LIST OF ABBREVIATIONS</b>          | ix        |
| <b>1</b>    | <b>INTRODUCTION</b>                   | <b>1</b>  |
|             | 1.1. OVERVIEW                         | 1         |
|             | 1.2. MACHINE LEARNING                 | 1         |
|             | 1.3.OBJECTIVE                         | 2         |
| <b>2</b>    | <b>LITERATURE SURVEY</b>              | <b>2</b>  |
|             | 2.1. RELATED WORK                     | 2         |
| <b>3</b>    | <b>METHODOLOGY</b>                    | <b>8</b>  |
|             | 3.1. EXISTING SYSTEM                  | 8         |
|             | 3.2. PROPOSED SYSTEM                  | 9         |
|             | <b>3.3 ALGORITHMS USED</b>            | <b>10</b> |
|             | 3.3.1. NAIVE BAYES ALGORITHM          |           |
|             | 3.3.2 . K-NEAREST NEIGHBOUR ALGORITHM |           |
|             | 3.4 SYSTEM ARCHITECTURE               | 15        |
|             | 3.5 SYSTEM REQUIREMENTS               | 15        |
|             | 3.6 MODULES                           | 16        |
|             | 3.7 UML DIAGRAMS                      | 17        |
|             | 3.8 LANGUAGES USED                    | 21        |
|             | 3.9 REQUIREMENT ANALYSIS              | <b>25</b> |
| <b>4</b>    | <b>RESULTS AND DISCUSSION</b>         | <b>29</b> |
|             | 4.1. WORKING                          | 29        |
| <b>5</b>    | <b>CONCLUSION</b>                     | <b>34</b> |
|             | 5.1. CONCLUSION                       | 34        |
|             | <b>REFERENCES</b>                     | <b>34</b> |
|             | <b>APPENDICES</b>                     | <b>36</b> |
|             | A. SOURCE CODE                        | 36        |

|                      |    |
|----------------------|----|
| B. SCREENSHOTS       | 41 |
| C. PLAGIARISM REPORT | 43 |
| D. JOURNAL PAPER     | 45 |

## LIST OF FIGURES

| Figure No. | Figure Name                   | Page No. |
|------------|-------------------------------|----------|
| 3.3.1      | Posterior probability formula | 11       |
| 3.3.2      | Likelihood table              | 12       |
| 3.4        | System Architecture           | 15       |
| 3.7.1      | Usecase diagram               | 18       |
| 3.7.2      | Sequence diagram              | 19       |
| 3.7.3      | Activity diagram              | 20       |
| 3.8.1      | platform independence method  | 23       |
| 3.8.2      | Collection framework          | 24       |
| B.1        | Input page                    | 41       |
| B.2        | Output page                   | 42       |
| B.3        | Accuracy Graph                | 42       |

## LIST OF ABBREVIATIONS

| ABBREVIATIONS | EXPANSION             |
|---------------|-----------------------|
| ML            | Machine Learning      |
| NB            | Naive Bayes           |
| KNN           | Kth nearest neighbour |

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 OVERVIEW**

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. Students are the major strength for numerous universities. Universities and students play a significant part in producing graduates of superior calibers with its academic performance accomplishment. However, academic performance achievement changes as various sort of students may have diverse degree of performance achievement. Machine learning is the ability of a system to consequently gain from past experience and improve performance. Nowadays machine learning for education gains more attention. Machine learning is used for analyzing information based on past experience and predicting future performance.

### **1.2 MACHINE LEARNING**

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning

algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

### **1.3 OBJECTIVE**

The main objective this paper introduces a ML model that classify and predict student performance by utilizing supervised ML algorithms like Naïve Bayes and K-Nearest Neighbour. Thus, the proposed approach offers a solution to predict performance efficiently and accurately by comparing several ML model.

## **CHAPTER 2**

### **LITERATURE SURVEY**

#### **2.1 RELATED WORK**

Abdallah Namoun et.al[1] proposed accomplishing learning results is estimated principally by the exhibition of the grades (f i.e., level) and the accomplishment scores (i.e., grades). Audit and investigation of AI machines were regularly used to exhibit understudy execution. At long last, online understudy exercises, evaluation scores, and understudy feelings were the main indicators of accomplishment. Distinguish the attributes and strategies that decide the exhibition of understudies utilizing the PICO strategy. This evaluation has had many difficulties, as it is by and large wide, not zeroed in on the utilization of understudy input as a benchmark for understudy execution, experienced quality issues, and has not been distributed in the most as often as possible posed inquiries.

Durgesh Ugale et.al[2] discussed the handling step will be applied to the crude informational index to appropriately utilize the extraction calculation. The "execution" of guiding an understudy can assist with working on their presentation.

The answer for the issue is very distinct. He didn't remark on the understudies' reactions. You didn't talk about the cutoff. You didn't analyze the idea of the prescience.

Ali Salah Hashim et.al[3] In this review, we analyzed the exhibition of machine-controlled calculations. As per the consequences of the overview, the coordination's grouping was the most right in picking the last score of the understudy (68.7% of champs and 88.8% of washouts). Depict the strategies utilized in insightful exploration to anticipate understudy execution. You have not appraised the nature of the preparation. You have not thought about the models.

Alnassar, Fatema et.al[4] talks about the connection between understudy association law calculation, K-key calculation, and choice tree. This review looks at understudy execution dependent on various attributes. The design incorporates questions and replies to illustrations, little and last experimental outcomes, schoolwork, and lab work.

We talk about dynamic trees, data mining techniques, and a blend of strategies that empower understudies to anticipate understudy execution, and educators can make significant strides in creating understudy information. execution. The presentation of different tree calculations can be investigated dependent on its exactness and timing of tree conveyance. Speculations eliminated from the framework assisted the educator with distinguishing understudies with incapacities and further develop execution. Knowing the quantity of gatherings ahead of time. It is hard to know the quantity of gatherings when there is a slight change in the information.

Saifuzzaman et.al[5] discussed outcomes show that calculations and different techniques are utilized to acquire data from instruction. A significant number of these calculations are significant for characterizing and controlling data. In this review, the calculations C5.0, C4.5, and K-esteem were distinguished. At 48% of the locales, these three uncovering calculations are broadly utilized for information examination, particularly in instruction. He wrote about the general accomplishment of the understudies. The gauge study is featured with practically no huge clarification.

Leena H. Alamri et.al[6] proposed consequences of the SVM and RF calculations utilized in the two segments show that the honesty of the two parts is up to 93%, while the least RMSE is 1.13 for the RF. The request in which the examination report is set up is simple and quick to foresee. It functions admirably in numerous classes of speculations. Since SVMs are known as terrible number crunchers, yield as expectations ought not be viewed as significant.

C. Verma et.al[7] discussed the outcomes show that understudy interest and GPA in the principal semester drove the whole choice interaction, and the Bayesian organization is better than the choice tree because of its high precision. It works better as far as pay level than the quantity of factors. In the event that you attempt a little example, the information might be excessively high or excessively high.

Ferda Ünal[8] proposed this paper exhibits the utilization of information recovery innovation to distinguish eventual outcomes dependent on understudy history. In the exploration, three techniques for information mining (choice tree, non-repeating memory, and naive Bayes) were utilized in two lines of information science and Portuguese. The outcomes show that mining strategies are valuable in choosing understudy execution. Build up clear guidelines for anticipating preparing data. With the progression of innovation, e-learning as a web-based learning webpage and progressed mixed media innovation, preparing costs have been decreased and time and difficulties have been eliminated.

### **Performance prediction of students using distributed data mining**

The performance of students in higher education in India is a turning point in the academics for all students for their brightest career. In today's generation the amount of data stored in educational database increasing at a great rate. These databases contain secret information for improvement of students' performance; these data can be located at different nodes in distributed system. Classification and prediction are among the major techniques in Data mining and widely used in various fields. In this paper classification techniques are used for prediction of student performance in distributed environment. Data mining methods are often implemented at many advance universities today for analyzing available data and extracting information

and knowledge to support decisionmaking. While it is important to have models at local level, their results makes it difficult to extract knowledge that can be useful at the global level. Therefore, to support decision making at this area, it is important to generalize the information contained in those models, specific classifier method can be used to generalize these rules for global model.

### **Predicting Student Academic Performance**

Engineering schools worldwide have a relatively high attrition rate. Typically, about 35% of the first-year students in various engineering programs do not make it to the second year. Of the remaining students, quite often they drop out or fail in their second or third year of studies. The purpose of this investigation is to identify the factors that serve as good indicators of whether a student will drop out or fail the program. In order to establish early warning indicators, principal component analysis is used to analyze, in the first instance, first-year engineering student academic records. These performance predictors, if identified, can then be used effectively to formulate corrective action plans to improve the attrition rate.

### **Data Mining Approach For Predicting Student Performance**

This work proposes a novel approach - personalized forecasting - to take into account the sequential effect in predicting student performance (PSP). Instead of using all historical data as other methods in PSP, the proposed methods only use the information of the individual students for forecasting his/her own performance. Moreover, these methods also encode the "student effect" (e.g. how good/clever a student is, in performing the tasks) and "task effect" (e.g. how difficult/easy the task is) into the models. Experimental results show that the proposed methods perform nicely and much faster than the other state-of-the-art methods in PSP.

### **A novel approach for upgrading Indian education by using data mining techniques**

Education is the backbone of all developing countries. Upgrading of the education system, upgrades the country to the world top ranking level. One of the major problems that the education system facing is predicting the behaviour of students from large database. This paper focus on upgrading Indian education system by using one of the techniques in Data mining namely clustering. Cluster analysis solves the given data into some meaningful groups. Normally the performances of the students can be classified into different patterns as normal, average and below average. In this paper we attempt to analyze student's data in different angle beyond the above indicated patterns through newly proposed UCAM (Unique clustering with Affinity Measures) clustering algorithm.

### **A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration**

Educational Data Mining (EDM) is an interdisciplinary ingenuous research area that handles the development of methods to explore data arising in a scholastic fields. Computational approaches used by EDM is to examine scholastic data in order to study educational questions. As a result, it provides intrinsic knowledge of teaching and learning process for effective education planning. This paper conducts a comprehensive study on the recent and relevant studies put through in this field to date. The study focuses on methods of analysing educational data to develop models for improving academic performances and improving institutional effectiveness. This paper accumulates and relegates literature, identifies consequential work and mediates it to computing educators and professional bodies. We identify research that gives well-fortified advise to amend edifying and invigorate the more impuissant segment students in the institution. The results of these studies give insight into techniques for ameliorating pedagogical process, presaging student performance, compare the precision of data mining algorithms, and demonstrate the maturity of open source implements.

### **Data Mining: A prediction of performer or underperformer using classification**

Now a day's students have a large set of data having precious information hidden. Data mining technique can help to find this hidden information. In this paper, data mining techniques name Byes classification method is used on these data to help an institution. Institutions can find those students who are consistently perform well. This study will help to institution reduce the drop put ratio to a significant level and improve the performance level of the institution.

### **Mining students data to analyze e-Learning behavior: A Case Study**

Educational data mining concerns with developing methods for discovering knowledge from data that come from educational environment. In this paper we used educational data mining to analyze learning behavior. In our case study, we collected students' data from DataBase course. After preprocessing the data, we applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, we extracted knowledge that describes students' behavior.

### **Scholastic achievement of higher secondary students in science stream**

The present study was conducted on 400 students (200 boys and 200 Girls) selected from senior secondary school of A.M.U., Aligarh-India, to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The scores obtained on different variables were factor-analyzed to get a smaller number of meaningful variables or factors to establish the predictive validity of these predictors. Factors responsible for success in science stream were identified. The prognostic value of the predictors was compared for high achievers and low achievers in order to identify the factors which differentiate them.

### **Mining student data using decision trees**

Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. This

paper is an attempt to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. For this purpose, the CRISP framework for data mining is used for mining student related academic data. The classification rule generation process is based on the decision tree as a classification method where the generated rules are studied and evaluated. A system that facilitates the use of the generated rules is built which allows students to predict the final grade in a course under study.

### **An Empirical Study of the Applications of Data Mining Techniques in Higher Education**

Few years ago, the information flow in education field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that educational institutions face is the explosive growth of educational data and to use this data to improve the quality of managerial decisions. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the applications of data mining in educational institution to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples.

## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 EXISTING SYSTEM**

In this research on existing methods of prediction is still insufficient to determine the most appropriate methods for predicting student performance in institutions. Second, is the absence of inquiry of the specific courses. There are various types of

classification are implemented in existing framework. The datasets are uploaded in WEKA tool with any WINDOWS OS configuration. K-means clustering algorithms provides reduce number of error rate values.

- **EXISTING ALGORITHM**
- **K-MEANS CLUSTERING ALGORITHM**

What is K-means?

1. Partitional clustering approach
2. Each cluster is associated with a centroid (center point)
3. Each point is assigned to the cluster with the closest centroid
- 4 Number of clusters K must be specified

#### **Details of K-means**

1. Initial centroids are often chosen randomly. - Clusters produced vary from one run to another
2. The centroid is (typically) the mean of the points in the cluster.
3. 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
4. K-means will converge for common similarity measures mentioned above.
  - 5. Most of the convergence happens in the first few iterations. - Often the stopping condition is changed to 'Until relatively few points change clusters'

### **3.2 PROPOSED SYSTEM**

In the proposed system Machine Learning algorithm is used for the classification. An automated evaluation system has been proposed to evaluate student performance and to analyze. A prediction system has been proposed by using their marks, staff opinion, attendance and ragging. The study is evaluated using machine learning classifier. Analysis of the student behaviour has been proposed by using intellectual

parameters of the student which affect their study. Various mining techniques are used to determine the educational data covering some factors. In this paper, a novel approach based on KNN significant academic attributes for performance predictions. The experiment displays good performance of the proposed algorithm and was compared to similar approaches over the same dataset. By analyzing the experimental results, it is observed that the Naïve Bayes and KNN algorithm turned out to be best classifier for student performance prediction because it contains more accuracy and least error rate.

### **ADVANTAGES OF PROPOSED SYSTEM**

- We did same steps to build cumulative predicting model and there may be some change of syntax due to different technology as in paper author used JAVA.
- We have used same dataset given by you to implement algorithms given in paper and in paper author has compare cumulative predictor with Naïve Bayes and KNN.

### **3.3 ALGORITHMS USED**

#### **3.3.1 NAIVE BAYES ALGORITHM**

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Look at the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$

**Fig:3.3.1 : Posterior probability formula**

Above,

- $P(c|x)$  is the posterior probability of class (*c*, target) given predictor (*x*, attributes).
- $P(c)$  is the prior probability of class.
- $P(x|c)$  is the likelihood which is the probability of predictor given class.
- $P(x)$  is the prior probability of predictor.

### How Naive Bayes algorithm works?

Let's understand it using an example. Below I have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition. Let's follow the below steps to perform it.

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

| Weather  | Play |
|----------|------|
| Sunny    | No   |
| Overcast | Yes  |
| Rainy    | Yes  |
| Sunny    | Yes  |
| Sunny    | Yes  |
| Overcast | Yes  |
| Rainy    | No   |
| Rainy    | No   |
| Sunny    | Yes  |
| Rainy    | Yes  |
| Sunny    | No   |
| Overcast | Yes  |
| Overcast | Yes  |
| Rainy    | No   |

| Frequency Table |    |     |
|-----------------|----|-----|
| Weather         | No | Yes |
| Overcast        |    | 4   |
| Rainy           | 3  | 2   |
| Sunny           | 2  | 3   |
| Grand Total     | 5  | 9   |

| Likelihood table |       |       |       |      |
|------------------|-------|-------|-------|------|
| Weather          | No    | Yes   |       |      |
| Overcast         |       | 4     | =4/14 | 0.29 |
| Rainy            | 3     | 2     | =5/14 | 0.36 |
| Sunny            | 2     | 3     | =5/14 | 0.36 |
| All              | 5     | 9     |       |      |
|                  | =5/14 | =9/14 |       |      |
|                  | 0.36  | 0.64  |       |      |

**Fig:3.3.2 :Likelihood table**

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

**Problem:** Players will play if weather is sunny. Is this statement is correct?

We can solve it using above discussed method of posterior probability.

$$P(\text{Yes} \mid \text{Sunny}) = P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$$

Here we have  $P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$ ,  $P(\text{Sunny}) = 5/14 = 0.36$ ,  $P(\text{Yes}) = 9/14 = 0.64$

Now,  $P(\text{Yes} \mid \text{Sunny}) = 0.33 * 0.64 / 0.36 = 0.60$ , which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

**Pros:**

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

### **Cons:**

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict\_proba are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

### **Applications of Naive Bayes Algorithms**

- **Real time Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam

e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

- **Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

### 3.3.2 K-NEAREST NEIGHBOR(KNN) CLASSIFICATION METHOD

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. The neighbors are taken from a set of objects for which the class (for  $k$ -NN classification) or the object property value (for  $k$ -NN regression) is known.

STEP 1: BEGIN

STEP 2: Input:  $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$

STEP 3:  $x = (x_1, \dots, x_n)$  new instance to be classified

STEP 4: FOR each labelled instance  $(x_i, c_i)$  calculate  $d(x_i, x)$

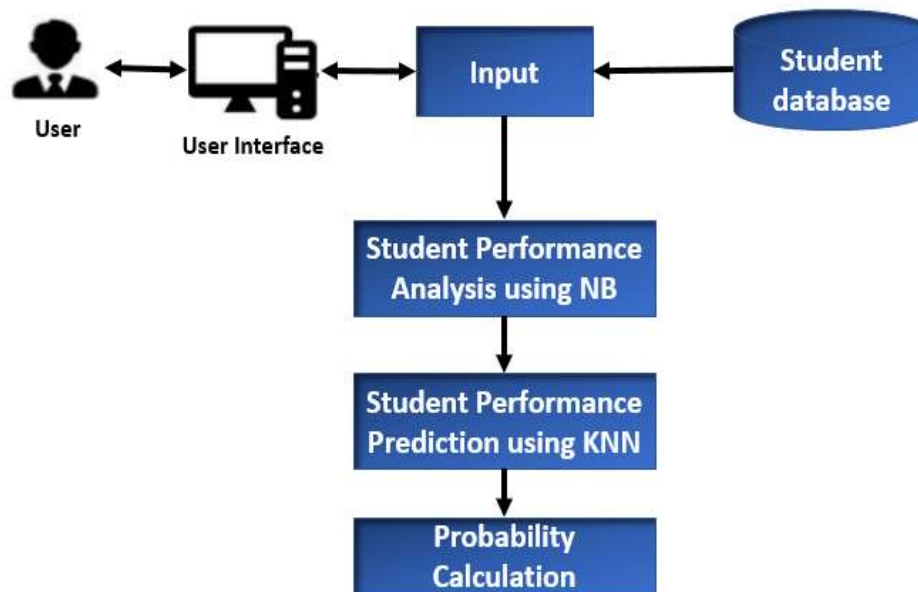
STEP 5: Order  $d(x_i, x)$  from lowest to highest,  $(i = 1 \dots N)$

STEP 6: Select the  $K$  nearest instances to  $x$ :  $D_x^k$

STEP 7: Assign to  $x$  the most frequent class in  $D_x^k$

STEP 8: END

### 3.4 SYSTEM ARCHITECTURE



*Fig:3.4 :System Architecture*

### 3.5 SYSTEM REQUIREMENTS

#### HARDWARE REQUIREMENTS:

- System - Pentium-IV

- Speed - 2.4GHZ
- Hard disk - 40GB
- Monitor - 15VGA color
- RAM - 512MB

## **SOFTWARE REQUIREMENTS:**

- Operating System - Windows XP
- Coding language - Java
- IDE - Net beans
- Database -MYSQL

## **3.6 MODULES**

- Data Collection
- Preprocessing
- Classification module
- Prediction

## **DATA COLLECTION**

In this module, Student data's will be collected from the college. Student's data like mark, attendance, staff opinion, extracurricular activities, Ragging and stress.

## **PREPROCESSING**

Data pre-processing is done to remove the incomplete noisy and inconsistent data. Data must be pre-processed before using in feature selection task.

## **CLASSIFICATION MODULE**

The data mining techniques used for identifying the performance of the student using Naïve Bayes and KNN algorithms. These two algorithm's identifies and analyses the performance of the student.

## **PREDICTION**

In this module, to predict the student performance based upon student mark, attendance, staff opinion, extracurricular activities, Ragging and stress.

### **3.7 UML DIAGRAMS**

UML is simply another graphical representation of a common semantic model. UML provides a comprehensive notation for the full lifecycle of object-oriented development.

#### **ADVANTAGES**

- To represent complete systems (instead of only the software portion) using object oriented concepts
- To establish an explicit coupling between concepts and executable code
- To take into account the scaling factors that are inherent to complex and critical systems
- To creating a modeling language usable by both humans and machines

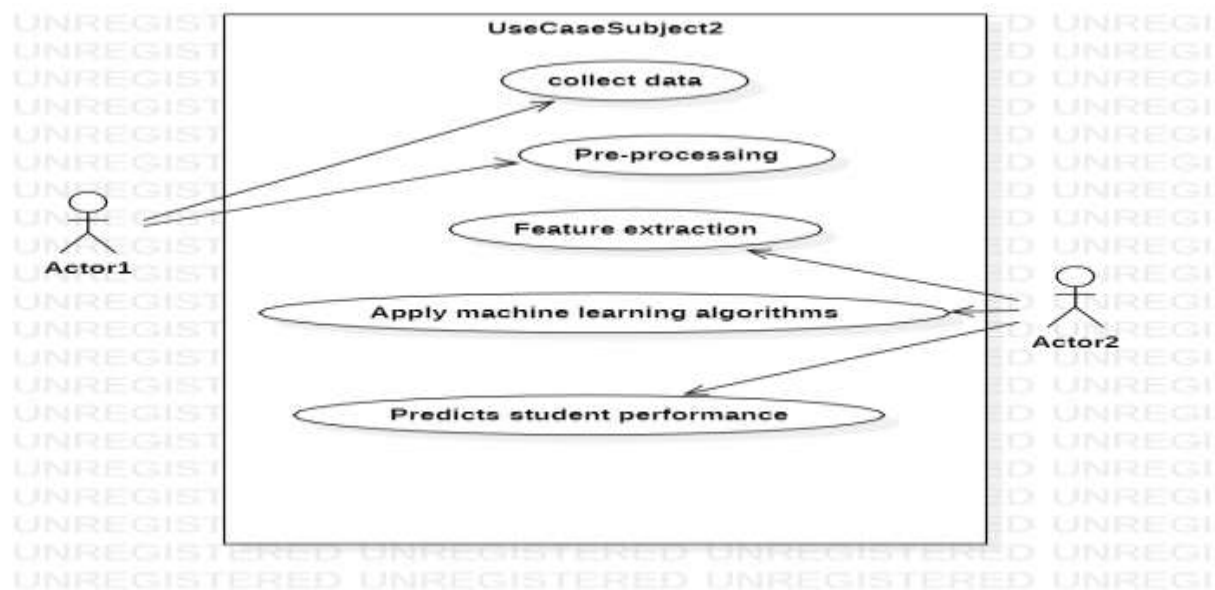
UML defines several models for representing systems

- The class model captures the static structure
- The state model expresses the dynamic behavior of objects
- The use case model describes the requirements of the user
- The interaction model represents the scenarios and messages flows
- The implementation model shows the work units

- The deployment model provides details that pertain to process allocation

- **USECASE DIAGRAM**

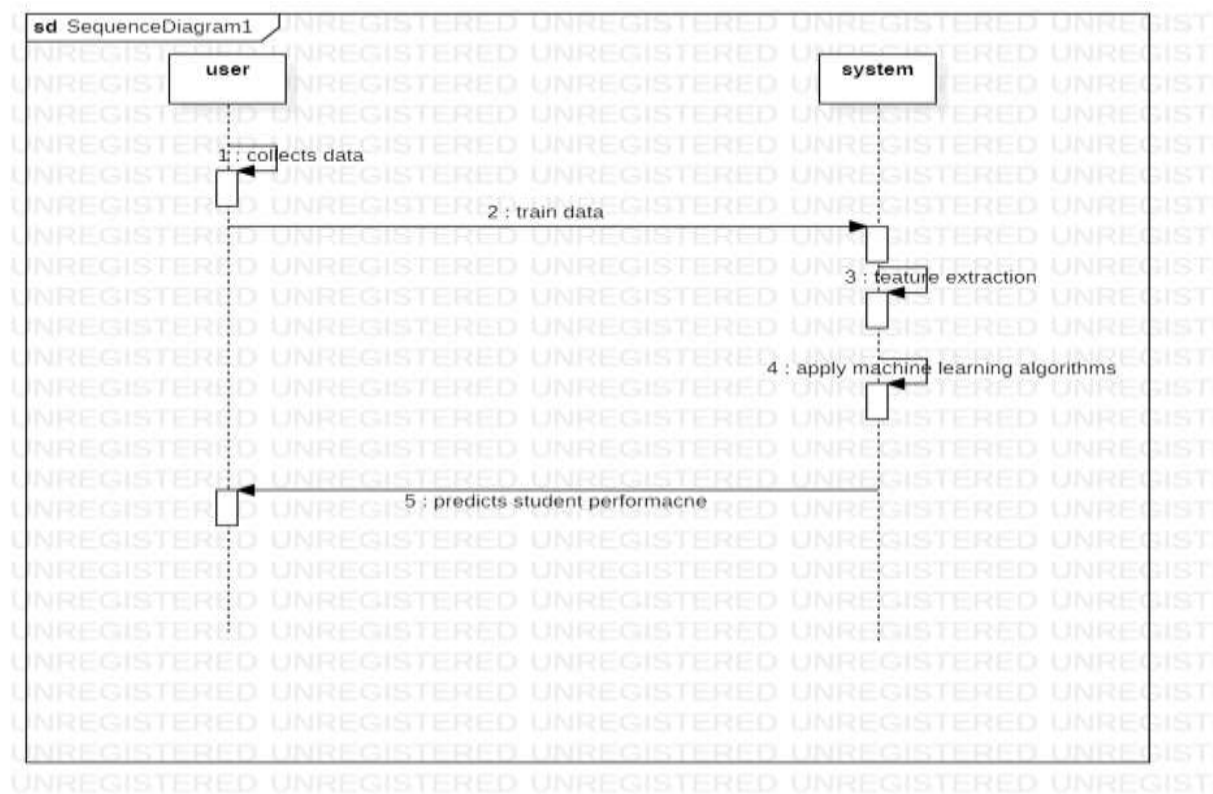
Use case diagrams overview the usage requirement for system. They are useful for presentations to management and/or project stakeholders, but for actual development you will find that use cases provide significantly more value because they describe “the meant” of the actual requirements. A use case describes a sequence of action that provides something of measurable value to an action and is drawn as a horizontal ellipse.



**Fig:3.7.1 :Usecase diagram**

- **SEQUENCE DIAGRAM**

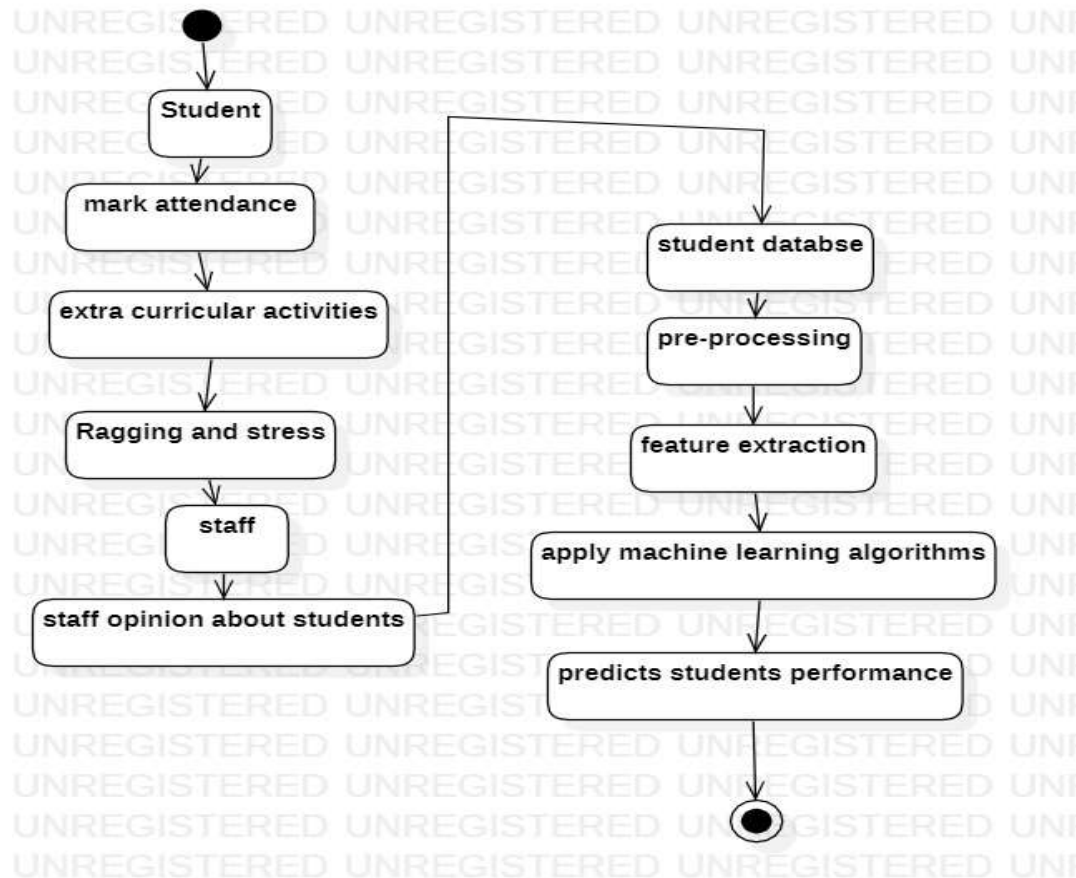
Sequence diagram model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and commonly used for both analysis and design purpose. Sequence diagram are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system.



***Fig:3.7.2 : Sequence diagram***

- **ACTIVITY DIAGRAM**

Activity diagram are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. The activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. Activity diagram consist of Initial node, activity final node and activities in between.



**Fig:3.7.3 : Activity diagram**

### 3.8 LANGUAGES USED

- JAVA

Java is one of the world's most important and widely used computer languages, and it has held this distinction for many years. Unlike some other computer languages whose influence has wearied with passage of time, while Java's has grown.

## **APPLICATION OF JAVA**

Java is widely used in every corner of world and of human life. Java is not only used in softwares but is also widely used in designing hardware controlling software components. There are more than 930 million JRE downloads each year and 3 billion mobile phones run java.

Following are some other usage of Java:

1. Developing Desktop Applications
2. Web Applications like Linkedin.com, Snapdeal.com etc
3. Mobile Operating System like Android
4. Embedded Systems
5. Robotics and games etc.

## **FEATURES OF JAVA**

The prime reason behind creation of Java was to bring portability and security feature into a computer language. Beside these two major features, there were many other features that played an important role in moulding out the final form of this outstanding language. Those features are;

### **1) Simple**

Java is easy to learn and its syntax is quite simple, clean and easy to understand. The confusing and ambiguous concepts of C++ are either left out in Java or they have been re-implemented in a cleaner way.

*Eg:* Pointers and Operator Overloading are not there in java but were an important part of C++.

## **2) Object Oriented**

In java everything is Object which has some data and behaviour. Java can be easily extended as it is based on Object Model.

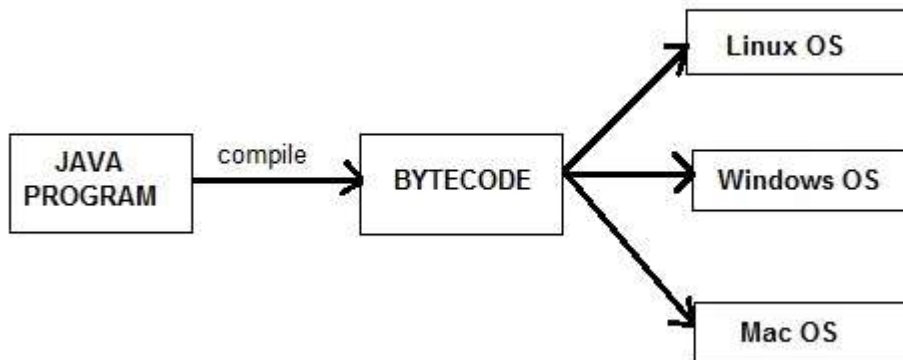
## **3) Robust**

Java makes an effort to eliminate error prone codes by emphasizing mainly on compile time error checking and runtime checking. But the main areas which Java improved were Memory Management and mishandled Exceptions by introducing automatic Garbage Collector and Exception Handling.

## **4) Platform Independent**

Unlike other programming languages such as C, C++ etc. which are compiled into platform specific machines. Java is guaranteed to be write-once, run-anywhere language.

On compilation Java program is compiled into byte code. This byte code is platform independent and can be run on any machine, plus this byte code format also provide security. Any machine with Java Runtime Environment can run Java Programs.



***Fig:3.8.1 : platform independence method***

### **5) Secure**

When it comes to security, Java is always the first choice. With java secure features it enable us to develop virus free, temper free system. Java program always runs in Java runtime environment with almost null interaction with system OS, hence it is more secure.

### **6) Multi-Threading**

Java multithreading feature makes it possible to write program that can do many tasks simultaneously. Benefit of multithreading is that it utilizes same memory and other resources to execute multiple threads at the same time, like While typing, grammatical errors are checked along.

### **7) Architectural Neutral**

Compiler generates byte codes, which have nothing to do with a particular computer architecture, hence a Java program is easy to interpret on any machine.

### **8) Portable**

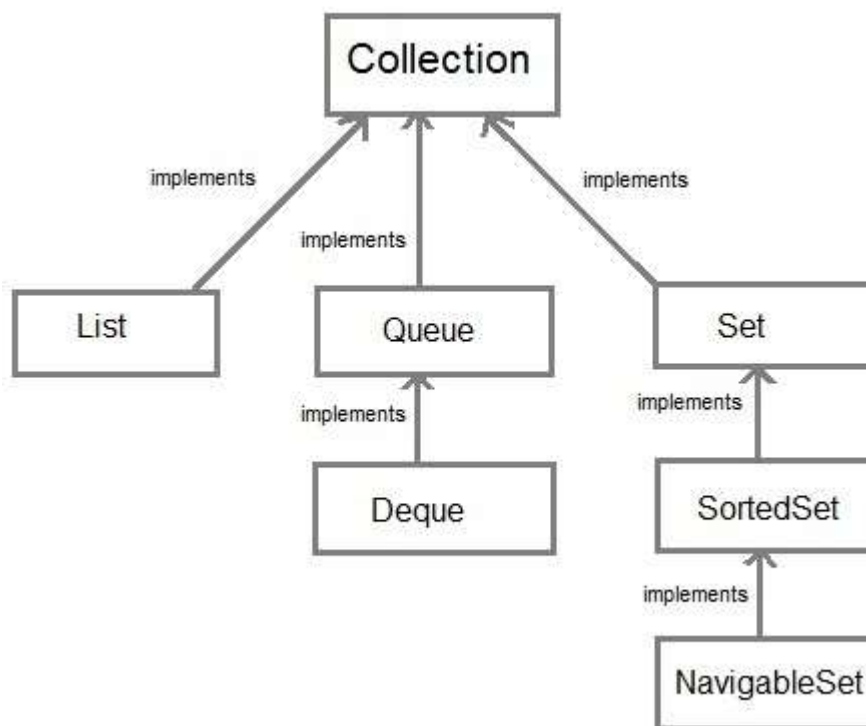
Java Byte code can be carried to any platform. No implementation dependent features. Everything related to storage is predefined, example: size of primitive data types

## 10) High Performance

Java is an interpreted language, so it will never be as fast as a compiled language like C or C++. But, Java enables high performance with the use of just-in-time compiler.

## COLLECTION FRAMEWORK

Collection framework was not part of original Java release. Collections was added to J2SE 1.2. Prior to Java 2, Java provided adhoc classes such as Dictionary, Vector, Stack and Properties to store and manipulate groups of objects. Collection framework provides many important classes and interfaces to collect and organize group of alike objects.



**Fig:3.8.2 : Collection framework**

- **MYSQL**

MySQL, officially, but also called "My Sequel" is the world's most widely used open-source relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases, though SQLite probably has more total embedded deployments. The SQL phrase stands for Structured Query Language.

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation. MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks).

LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Freesoftware-open source projects that require a full-featured database management system often use MySQL.

For commercial use, several paid editions are available, and offer additional functionality. Applications which use MySQL databases include: TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, Drupal and other software. MySQL is also used in many high-profile, large-scale websites, including Wikipedia, Google (though not for searches), Facebook, Twitter, Flickr and YouTube.

### **3.9 REQUIREMENT ANALYSIS**

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks

that determine the need for analysing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

## **FUNCTIONAL REQUIREMENTS**

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

### **Usability**

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

### **Robustness**

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

### **Security**

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

### **Reliability**

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

## **Compatibility**

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

## **Flexibility**

The flexibility of the project is provided in such a way that it has the ability to run on different environments being executed by different users.

## **Safety**

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

## **NON- FUNCTIONAL REQUIREMENTS**

### **Portability**

It is the usability of the same software in different environments. The project can be run in any operating system.

### **Performance**

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

### **Accuracy**

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

### **Maintainability**

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the

application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

## **FEASIBILITY STUDY**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components.

## **ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system is well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

## **TECHNICAL FEASIBILITY**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

## **SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity.

# **CHAPTER 4**

## **RESULTS AND DISCUSSION**

### **4.1 WORKING**

#### **INPUT DESIGN**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing

can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

## **OUTPUT DESIGN**

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

## **Test plan**

Software testing is the process of evaluation a software item to detect differences between given input and expected output. Also to assess the feature of a software

item. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

### **Verification**

Verification is the process to make sure the product satisfies the conditions imposed at the start of the development phase. In other words, to make sure the product behaves the way we want it to.

### **Validation**

Validation is the process to make sure the product satisfies the specified requirements at the end of the development phase. In other words, to make sure the product is built as per customer requirements.

### **Basics of software testing**

There are two basics of software testing: black box testing and white box testing.

#### **Black box Testing**

Black box testing is a testing technique that ignores the internal mechanism of the system and focuses on the output generated against any input and execution of the system. It is also called functional testing.

#### **White box Testing**

White box testing is a testing technique that takes into account the internal mechanism of a system. It is also called structural testing and glass box testing. Black box testing is often used for validation and white box testing is often used for verification.

## **Types of testing**

There are many types of testing like

- Unit Testing
- Integration Testing
- Functional Testing
- System Testing
- Stress Testing
- Performance Testing
- Usability Testing
- Acceptance Testing
- Regression Testing
- Beta Testing

### **Unit Testing**

Unit testing is the testing of an individual unit or group of related units. It falls under the class of white box testing. It is often done by the programmer to test that the unit he/she has implemented is producing expected output against given input.

### **Integration Testing**

Integration testing is testing in which a group of components are combined to produce output. Also, the interaction between software and hardware is tested in integration testing if software and hardware components have any relation. It may fall under both white box testing and black box testing.

### **Functional Testing**

Functional testing is the testing to ensure that the specified functionality required in the system requirements works. It falls under the class of black box testing.

### **System Testing**

System testing is the testing to ensure that by putting the software in different environments (e.g., Operating Systems) it still works. System testing is done with full system implementation and environment. It falls under the class of black box testing.

### **Stress Testing**

Stress testing is the testing to evaluate how system behaves under unfavorable conditions. Testing is conducted at beyond limits of the specifications. It falls under the class of black box testing.

### **Performance Testing**

Performance testing is the testing to assess the speed and effectiveness of the system and to make sure it is generating results within a specified time as in performance requirements. It falls under the class of black box testing.

### **Usability Testing**

Usability testing is performed to the perspective of the client, to evaluate how the GUI is user-friendly? How easily can the client learn? After learning how to use, how proficiently can the client perform? How pleasing is it to use its design? This falls under the class of black box testing.

### **Acceptance Testing**

Acceptance testing is often done by the customer to ensure that the delivered product meets the requirements and works as the customer expected. It falls under the class of black box testing.

### **Regression Testing**

Regression testing is the testing after modification of a system, component, or a group of related units to ensure that the modification is working correctly and is not damaging or imposing other modules to produce unexpected results. It falls under the class of black box testing.

## **CHAPTER 5**

### **CONCLUSION**

#### **5.1 CONCLUSION**

The paper focuses on the student academic growth analysis using machine learning techniques. For analysis Naïve Bayes and KNN classifier are used. This process can help the instructor to decide easily about performance of the students and schedule better method for improving their academics. In future additional features are added to our dataset to acquire better accuracy.

#### **REFERENCES**

- [1] Namoun, Abdallah, and Abdullah Alshanqiti. "Predicting student performance using data mining and learning analytics techniques: a systematic literature review." *Applied Sciences* 11, no. 1 (2021): 237.
- [2] Ugale, Durgesh, Jeet Pawar, Sachin Yadav, and Chandrashekhar Raut. "STUDENT PERFORMANCE PREDICTION USING DATA MINING TECHNIQUES." (2020).

[3] Hashim, Ali Salah, Wid Akeel Awadh, and Alaa Khalaf Hamoud. "Student performance prediction model based on supervised machine learning algorithms." In *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032019. IOP Publishing, 2020.

[4] Alnassar, Fatema, Tim Blackwell, Elaheh Homayounvala, and Matthew Yee-king. "How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment." In *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 1-6. IEEE, 2021.

[5] Saifuzzaman, M., Parvin, M., Jahan, I., Moon, N.N., Nur, F.N. and Shetu, S.F., 2021, June. Machine Learning Approach to Predict SGPA and CGPA. In *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)* (pp. 211-216). IEEE.

[6] H. Alamri, Leena, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan, and Nida Aslam. "Predicting Student Academic Performance using Support Vector Machine and Random Forest." In *2020 3rd International Conference on Education Technology Management*, pp. 100-107. 2020.

[7] Verma, C., Illés, Z. and Stoffová, V., 2019, February. Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-7). IEEE.

[8] Ünal, Ferda. "Data mining for student performance prediction in education." *Data Mining-Methods, Applications and Systems* (2020).

[9] Q. A. Al-Radaideh, E. W. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[10] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. IJACSA - International Journal of Advanced Computer Science and Applications, 2(3), 80-84. Retrieved from <http://ijacsa.thesai.org>.

## **APPENDICES**

### **A. SOURCE CODE**

#### **NaiveBayes.java**

```
import java.util.ArrayList;
```

```
public class NaiveBayes {
```

```
    //holds the mean values for each attribute in the yes class
```

```
    //holds the standard deviation values for each attribute in the yes class
```

```
    private double attributes[];
```

```
    private double yesMean[];
```

```

private double yesStdDev[];

private double noMean[];

private double noStdDev[];


private double pYes;

private double pNo;


public NaiveBayes(ArrayList<Person> training) {

    int size = training.get(0).getAttributes().length;

    yesMean = new double[size];

    noMean = new double[size];

    yesStdDev = new double[size];

    noStdDev = new double[size];

    double y = 0; //Number of yes examples in training set

    double n = 0; //Number of no examples in training set

    for(Person p : training) {

        attributes = p.getAttributes();

        if(p.getOutcome() == Outcome.yes) {

            y++;

            for(int i = 0; i < attributes.length; i++) {

                yesMean[i] += attributes[i];

            }

        }

    }

}

```

```

        else {
            n++;
            for(int i = 0; i < attributes.length; i++) {
                noMean[i] += attributes[i];
            }
        }
    }

    for(int i = 0; i < attributes.length; i++) {
        yesMean[i] = yesMean[i] / y;
        noMean[i] = noMean[i] / n;
    }

    for(Person p : training) {
        attributes = p.getAttributes();
        if(p.getOutcome() == Outcome.yes) {
            for(int i = 0; i < attributes.length; i++) {
                yesStdDev[i] += Math.pow(attributes[i] -
yesMean[i], 2);
            }
        }
        else {
            for(int i = 0; i < attributes.length; i++) {
                noStdDev[i] += Math.pow(attributes[i] - noMean[i],
2);

```

```

        }
    }
}

for(int i = 0; i < attributes.length; i++) {
    yesStdDev[i] = Math.sqrt((yesStdDev[i]/(y-1)));
    noStdDev[i] = Math.sqrt((noStdDev[i]/(n-1)));
}

pYes = y/(y+n); //probability of any given example being yes
pNo = n/(n+y); //probability of any given example being no
}

```

```

public Outcome Test(Person test) {
    double pYesGivenE = 1.0;
    double pNoGivenE = 1.0;
    double attributes[] = test.getAttributes();
    for(int i = 0; i < attributes.length; i++) {
        pYesGivenE *= getProbability(attributes[i], yesMean[i],
yesStdDev[i]);
        pNoGivenE *= getProbability(attributes[i], noMean[i],
noStdDev[i]);
    }
    pYesGivenE *= pYes;
    pNoGivenE *= pNo;
}

```

```

        if(pNoGivenE > pYesGivenE) {
            return Outcome.no;
        }
        else {
            return Outcome.yes;
        }
    }
}

```

```

    private double getProbability(double x, double mean, double stdDev) { //using
an assumed normal distribution

        double denominator = stdDev * Math.sqrt(2*Math.PI);

        double exponent = -1 * ( (Math.pow(x - mean, 2)) /
(2*Math.pow(stdDev, 2)) );

        double probability = (1/denominator) * Math.pow(Math.E, exponent);

        return probability;
    }
}

```

## B. SCREENSHOTS

Marks(Grade)  Attendance(Percentage)  Staff Remarks(Rating)

Ragging  Socialmedia  Extra Curricular Activities

Stress

Process

Result jLabel10

probability jLabel11

***Fig:B.1 : Input page***

Marks(Grade)  Attendance(Percentage)  Staff Remarks(Rating)

Ragging  Socialmedia  Extra Curricular Activities

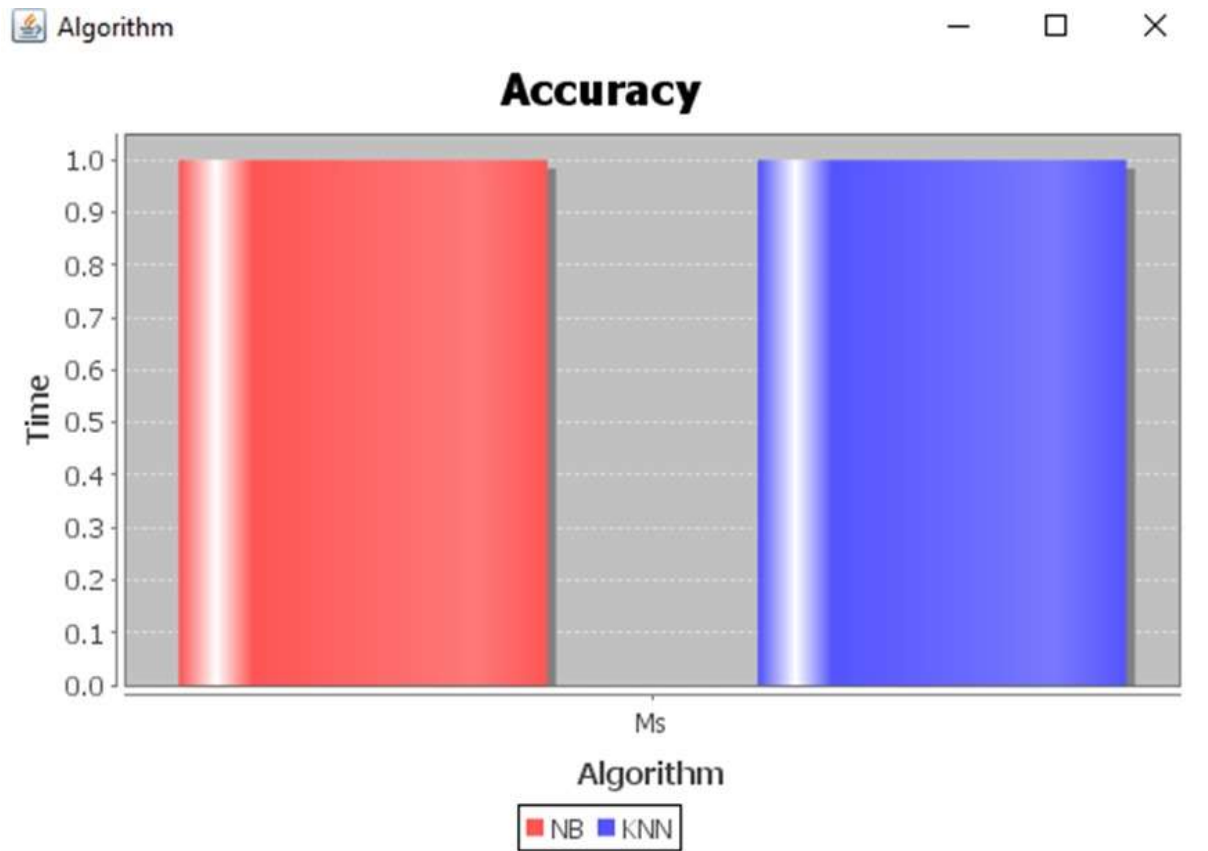
Stress

Process

Result Student Behaviour is: Good

probability 81 %

***Fig:B.2 : Output page***



**Fig:B.3 : Accuracy Graph**

## C. PLAGIARISM REPORT



# A MACHINE LEARNING APPROACH TO CLASSIFY STUDENT'S PERFORMANCE

R.Manideep<sup>1</sup>  
Student

[manideep.spartain@gmail.com](mailto:manideep.spartain@gmail.com)

P.V.Adarsh Kumar<sup>2</sup>  
Student

[adarshkumarpanthula@gmail.com](mailto:adarshkumarpanthula@gmail.com)

Mrs.D Deepa<sup>3</sup>  
Associate Professor

[deepa.cse@sathyabama.ac.in](mailto:deepa.cse@sathyabama.ac.in)

<sup>123</sup>Sathyabama Institute of Science and Technology, India

**Abstract:** In numerous associations, information recovery strategies are utilized to investigate data that can be utilized to decide. In the field of education, data mining is utilized in an assortment of exercises, including understudy evaluation, investment, staff input, unprecedented exercises, and stress. Information mining strategies are utilized to decide understudy execution utilizing Naïve Bayes and KNN calculations. In many organizations, data mining techniques are used for analyzing large amount of available data's, information's for decision making process. In educational sector, Data mining is used for wide variety of applications such as performance of the students like mark, attendance, staff opinion, extracurricular activities, Ragging and stress. The data mining techniques used for identifying the performance of the student using Naïve Bayes and KNN algorithms. These two algorithm's identifies and analyses the performance of the student. The two calculations decide and dissect understudy execution. The main aim of this project is to improvise the student performance in studies based on some important factors. Education is an important element for the improvement and development of a country. It enables the people of a country civilized and well mannered. Now-a-days developing new methods to discover knowledge from educational database in order to analyse student's trends and behaviours towards education. To examine the data from various

angles, categorize it and to summarize the relationships. It motivated us to work on student dataset analysation. The collection of data, assortment and classification is being performed manually.

**Keywords:** Supervised machine learning techniques, KNN, K- Means, Naive bayes, Student's performance.

## I. Introduction

There is a ton of examination interest in the utilization of information mining in schooling. This advanced field, called Information Science Mining, centers around the most effective ways to build information in the field of schooling. Media extraction is an order technique used to extricate pictures from huge chronicles.

These thoughts and procedures can be utilized in an assortment of enterprises, including publicizing, medical services, land, client relationship the executives, development, and web mining. Data can be gathered from old information and functional information put away in the instructive establishment. Understudy data can be close to home data or

scholarly execution. It is additionally open from the e-learning chronicle framework, which is utilized by many organizations.

It utilizes various strategies to successfully carry out information mining, for example, K-implies Clustering, K-Near neighbour calculations. Utilizing these abilities, it is feasible to acquire various kinds of information utilizing hierarchical guidelines, rankings, and coordination. Utilizing this, we infer information that clarifies the presentation of the understudies who took the test and every one of their subtleties.

The first of these numerous information assignments is to order essential data, investigate groups, and arrange them coherently. A gathering is a gathering of unmistakable and immaterial articles, separated into many gatherings at a similar level among them, and doing likewise in bunches that are firmly identified with various gatherings of various species.

There are many tremendous improvement research interests in using data mining in educational sector. This modern emerging sector, called educational Mining data, affected by improved methods that extract knowledge of data come from the educational sector. Data mining is a technique of sorting which is actually used to extract hidden patterns from huge databases. This concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining, etc.

Educational data mining is a new emerging or advanced technique of data exploration that can be applied to data related to the field of education. The data can be collected from past used data and operational data reside in the databases of educational institutes. The data of students can be personal information or academic performance. Exploration that can be applied to data related to the field of education systems which have a huge amount of data and information used by most institutes. It uses many techniques for proper implementation of data mining concepts such as K-means Clustering and K-Nearest neighbour. Using these techniques

different kinds of knowledge can be discovered using association rules, classification and grouping.

By using this we extract knowledge that describes students' performance in examination and all their detail information. From This huge amounts of data, the first task is to sort them out, cluster analysis is to classify the raw data in a reasonable way. Clustering is a bunch of physical or abstract objects, as per the degree of similarity between them, divided into several groups, and makes the same data objects within a groups of high similarity and different groups of data objects which are not similar.

## II. RELATED WORK

Abdallah Namoun et.al[1] proposed accomplishing learning results is estimated principally by the exhibition of the grades (f i.e., level) and the accomplishment scores (i.e., grades). Audit and investigation of AI machines were regularly used to exhibit understudy execution. At long last, online understudy exercises, evaluation scores, and understudy feelings were the main indicators of accomplishment. Distinguish the attributes and strategies that decide the exhibition of understudies utilizing the PICO strategy. This evaluation has had many difficulties, as it is by and large wide, not zeroed in on the utilization of understudy input as a benchmark for understudy execution, experienced quality issues, and has not been distributed in the most as often as possible posed inquiries.

Durgesh Ugale et.al[2] discussed the handling step will be applied to the crude informational index to appropriately utilize the extraction calculation. The "execution" of guiding an understudy can assist with working on their presentation. The answer for the issue is very distinct. He didn't remark on the understudies' reactions. You didn't talk about the cutoff. You didn't analyze the idea of the prescience.

Ali Salah Hashim et.al[3] In this review, we analyzed the exhibition of machine-controlled calculations.

As per the consequences of the overview, the coordination's grouping was the most right in picking the last score of the understudy (68.7% of champs and 88.8% of washouts). Depict the strategies utilized in insightful exploration to anticipate understudy execution. You have not appraised the nature of the preparation. You have not thought about the models.

Alnassar, Fatema et.al[4] talks about the connection between understudy association law calculation, K-key calculation, and choice tree. This review looks at understudy execution dependent on various attributes. The design incorporates questions and replies to illustrations, little and last experimental outcomes, schoolwork, and lab work.

We talk about dynamic trees, data mining techniques, and a blend of strategies that empower understudies to anticipate understudy execution, and educators can make significant strides in creating understudy information. execution. The presentation of different tree calculations can be investigated dependent on its exactness and timing of tree conveyance. Speculations eliminated from the framework assisted the educator with distinguishing understudies with incapacities and further develop execution. Knowing the quantity of gatherings ahead of time. It is hard to know the quantity of gatherings when there is a slight change in the information.

Saifuzzaman et.al[5] discussed outcomes show that calculations and different techniques are utilized to acquire data from instruction. A significant number of these calculations are significant for characterizing and controlling data. In this review, the calculations C5.0, C4.5, and K-esteem were distinguished. At 48% of the locales, these three uncovering calculations are broadly utilized for information examination, particularly in instruction. He wrote about the general accomplishment of the understudies. The gauge study is featured with practically no huge clarification.

Leena H. Alamri et.al[6] proposed consequences of the SVM and RF calculations utilized in the two segments show that the honesty of the two parts is up to 93%, while the least RMSE is 1.13 for the RF.

The request in which the examination report is set up is simple and quick to foresee. It functions admirably in numerous classes of speculations. Since SVMs are known as terrible number crunchers, yield as expectations ought not be viewed as significant.

C. Verma et.al[7] discussed the outcomes show that understudy interest and GPA in the principal semester drove the whole choice interaction, and the Bayesian organization is better than the choice tree because of its high precision. It works better as far as pay level than the quantity of factors. In the event that you attempt a little example, the information might be excessively high or excessively high.

Ferda Ünal[8] proposed this paper exhibits the utilization of information recovery innovation to distinguish eventual outcomes dependent on understudy history. In the exploration, three techniques for information mining (choice tree, non-repeating memory, and naive Bayes) were utilized in two lines of information science and Portuguese. The outcomes show that mining strategies are valuable in choosing understudy execution. Build up clear guidelines for anticipating preparing data. With the progression of innovation, e-learning as a web-based learning webpage and progressed mixed media innovation, preparing costs have been decreased and time and difficulties have been eliminated.

### III. PROPOSED WORK

As of now, the current framework just considers the deficiencies of the framework, which assists with surveying understudy execution. We don't have a framework to assist us with organizing our exercises, and we would rather not be slow-witted. But now they only take into consideration the constitution that the performance of the system is not sufficient to have a system that can help us assess the performance of the student. We won't do it for the reason of having no future for the performance of the, for our help, and they find that the unwanted fit into its own works.

The venture intends to foster a solid model utilizing data mining innovation and the necessary

extraction, with the goal that this instruction can be considered as an essential administration tool.

The proposed framework utilizes mining methods to assess execution and distinguish undesirable practices.

In the field of schooling, information mining is utilized in an assortment of ways, including understudy cooperation, worker input, uncommon exercises, and stress. Information mining procedures are utilized to decide understudy execution utilizing the KNN and NAÏVE BAYES calculations.

## IV. SYSTEM ARCHITECTURE

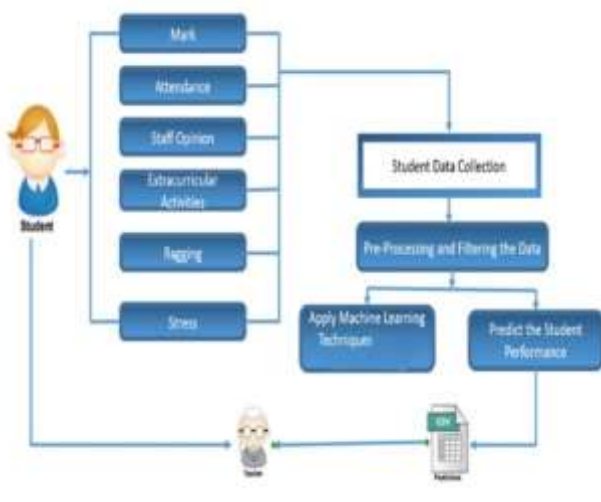


Fig. 1. System Architecture

## V. METHODOLOGY

### A. K-NEAREST NEIGHBOUR(KNN) CLASSIFICATION METHOD

K-NN is a model sort dependent on learning, or lethargic realizing, where execution is slow and slow until the estimation is completely carried out. The k-NN calculation is the least difficult of all AI calculations. Neighbors were taken out from things known by class (for k-NN classifications) or property things (for k-NN withdraws).

- STEP 1: BEGIN

- STEP 2: Input:  $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$
- STEP 3:  $x = (x_1, \dots, x_n)$  new instance to be classified
- STEP 4: FOR each labelled instance  $(x_i, c_i)$  calculate  $d(x_i, x)$
- STEP 5: Order  $d(x_i, x)$  from lowest to highest,  $(i = 1 \dots N)$
- STEP 6: Select the K nearest instances to x:  $D_{kx}$
- STEP 7: Assign to x the most frequent class in  $D_{kx}$
- STEP 8: END

### B. NAIVE BAYES ALGORITHM

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Look at the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability  
Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Above,

- $P(c|x)$  is the posterior probability of class ( $c$ , target) given predictor ( $x$ , attributes).
- $P(c)$  is the prior probability of class.
- $P(x|c)$  is the likelihood which is the probability of predictor given class.
- $P(x)$  is the prior probability of predictor.

We should pay attention to a model. Underneath I have data about the climate planning and the motivation behind the "Play" change (showing the play prospects). Presently we really want to list whether or not the players will play, contingent upon the climate. To do this, follow the subsequent stage.

Stage 1: Change the data put on the table.

Stage 2: Create a likelihood table to search for elements like cloud likelihood = 0.29 and playability 0.64.

Stage 3: Now utilize the Naive Bayesian condition to ascertain the greatest conceivable size for every class. The first and most significant advance is the result of the prediction.

## VI. MODULES

### A. Data Collection

This module assists you with gathering understudy data. Understudy scores, participation, staff criticism, unprecedented exercises, stress, and significantly more. In this module student's data will be gathered from the respective educational institution. Student's data like mark, attendance,

staff opinion, extracurricular activities, ragging etc. will be collected.

### B. Preprocessing

Adapting to data kills commotion and struggle. The determination of highlights should begin working before use. Data pre-processing is done to remove the noisy and inconsistent data. Data should be pre-processed before utilizing it in feature selection job.

### C. Classification Module

The strategy is utilized to decide understudy execution utilizing K-techniques and KNN calculations. Two calculations that investigate understudy execution that is the performance of the respective student.

### D. Prediction

This module features understudy execution dependent on understudy scores, participation, staff input, remarkable exercises, and stress. This module will predict the performance of the student by the above criteria.

## VII. RESULTS

### A. Dataset

|     |    |   |   |   |   |       |
|-----|----|---|---|---|---|-------|
| 9.5 | 90 | 5 | 1 | 0 | 1 | 1 Yes |
| 3.2 | 40 | 3 | 1 | 0 | 1 | 0 Yes |
| 3.7 | 45 | 2 | 0 | 1 | 0 | 0 No  |
| 5.8 | 69 | 1 | 0 | 1 | 0 | 1 No  |
| 5   | 50 | 3 | 1 | 0 | 1 | 0 Yes |
| 7.8 | 80 | 4 | 1 | 0 | 1 | 1 Yes |
| 7.5 | 75 | 3 | 1 | 1 | 0 | 0 No  |
| 8   | 80 | 4 | 1 | 1 | 1 | 1 Yes |
| 3.5 | 46 | 2 | 0 | 1 | 0 | 1 No  |
| 1.5 | 15 | 1 | 0 | 1 | 0 | 0 No  |
| 6.5 | 74 | 4 | 1 | 0 | 1 | 1 Yes |
| 5.8 | 62 | 5 | 1 | 0 | 1 | 1 Yes |
| 1   | 10 | 1 | 1 | 1 | 0 | 0 No  |
| 1.5 | 15 | 2 | 1 | 1 | 1 | 0 No  |
| 1.8 | 20 | 2 | 0 | 1 | 1 | 0 No  |
| 1.7 | 60 | 3 | 0 | 1 | 1 | 0 Yes |

Fig. 2. Dataset.

This data approach student achievement in secondary education of two schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance, the two datasets were modeled under binary/five-level classification and regression tasks.

### B. Input

Input form fields:

- Marks(Grade): 50
- Attendance(Percentage): 74
- Staff Remarks(Rating): 5
- Ragging: 1
- Socialmedia: 1
- Extra Curricular Activities: 0
- Stress: 0

Result: Student Behaviour is: Bad  
probability: 69%

Fig. 3. Input.

In the field of schooling, information mining is utilized in an assortment of ways, including understudy cooperation, worker input, uncommon exercises, and stress.

### C. Prediction

Prediction form fields:

- Marks(Grade): 50
- Attendance(Percentage): 74
- Staff Remarks(Rating): 5
- Ragging: 1
- Socialmedia: 1
- Extra Curricular Activities: 0
- Stress: 0

Result: Student Behaviour is: Bad  
probability: 69%

Fig. 4. Prediction.

### D. Graph

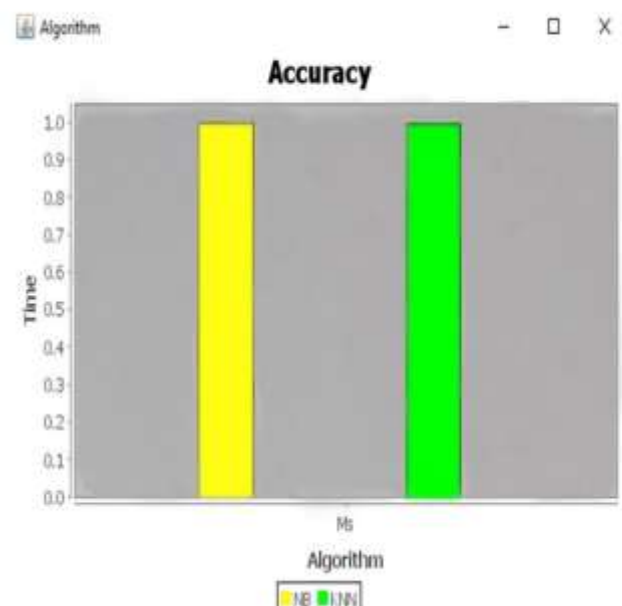


Fig. 5. Graph.

## VIII. CONCLUSION

In the news story, the arranging capacity is utilized to illuminate the understudy division utilizing the fundamental chronicles displayed in the understudy files.

There are such countless sorts it's difficult to say.

In view of key regions, like participation, school tests, studios, and results gave, the understudy has finished the grades. Essential exploration helps understudies and educators further develop understudy privileges.

Administration preparing additionally centers around distinguishing understudies who need to zero in on the best way to assist them with finishing the following test. This assists the understudies with working on their learning and perform well in their tests.

## IX. REFERENCES

- [1] Namoun, Abdallah, and Abdullah Alshanqiti. "Predicting student performance using data mining and learning analytics techniques: a systematic literature review." *Applied Sciences* 11, no. 1 (2021): 237.
- [2] Ugale, Durgesh, Jeet Pawar, Sachin Yadav, and Chandrashekhar Raut. "STUDENT PERFORMANCE PREDICTION USING DATA MINING TECHNIQUES." (2020).
- [3] Hashim, Ali Salah, Wid Akeel Awadh, and Alaa Khalaf Hamoud. "Student performance prediction model based on supervised machine learning algorithms." In *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032019. IOP Publishing, 2020.
- [4] Alnassar, Fatema, Tim Blackwell, Elaheh Homayounvala, and Matthew Yee-king. "How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment." In *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 1-6. IEEE, 2021.
- [5] Saifuzzaman, M., Parvin, M., Jahan, I., Moon, N.N., Nur, F.N. and Shetu, S.F., 2021, June. Machine Learning Approach to Predict SGPA and CGPA. In *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)* (pp. 211-216). IEEE.
- [6] H. Alamri, Leena, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan, and Nida Aslam. "Predicting Student Academic Performance using Support Vector Machine and Random Forest." In *2020 3rd International Conference on Education Technology Management*, pp. 100-107. 2020.
- [7] Verma, C., Illés, Z. and Stoffová, V., 2019, February. Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-7). IEEE.
- [8] Ünal, Ferda. "Data mining for student performance prediction in education." *Data Mining-Methods, Applications and Systems* (2020).
- [9] Aggarwal, Deepti, Sonu Mittal, and Vikram Bali. "Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques." *International Journal of Recent Technology and Engineering* 8 (2019): 497-503.
- [10] Chen, J., Zhang, Y., Wei, Y. and Hu, J., 2021. Discrimination of the contextual features of top

performers in scientific literacy using a machine learning approach. *Research in Science Education*, 51(1), pp.129-158.

[11] Sekeroglu, B., Dimililer, K. and Tuncal, K., 2019, March. Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11).

[12] Kučak, D., Juričić, V. and Đambić, G., 2018. MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS. *Annals of DAAAM & Proceedings*, 29.

[13] Alam, Mirza Mohtashim, Karishma Mohiuddin, Amit Kishor Das, Md Kabirul Islam, Md Shamsul Kaonain, and Md Haider Ali. "A reduced feature based neural network approach to classify the category of students." In *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, pp. 28-32. 2018

