TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

**ADVANCED COLLEGE OF ENGINEERING AND MANAGEMENT**

DEPARTMENT OF COMPUTER AND ELECTRONICS



MINOR PROJECT PROPOSAL ON:

**STUDENT PERFORMANCE PREDICTION**

|  |  |
| --- | --- |
| **SUBMITTED BY**: | AJAY PAKHRIN LAMA (073-BCT-504)  AYUB SHRESTHA (073-BCT-515)  JEEWAN SHARMA (073-BCT-531)  MANJIL MUNANKARMI (073-BCT-537) |

LALITPUR, NEPAL

MAY 26, 2019.

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

**ADVANCED COLLEGE OF ENGINEERING AND MANAGEMENT**

DEPARTMENT OF COMPUTER AND ELECTRONICS



MINOR PROJECT PROPOSAL ON:

**STUDENT PERFORMANCE PREDICTION**

|  |  |
| --- | --- |
| **SUBMITTED BY**: | AJAY PAKHRIN LAMA (073-BCT-504)  AYUB SHRESTHA (073-BCT-515)  JEEWAN SHARMA (073-BCT-531)  MANJIL MUNANKARMI (073-BCT-537) |

**SUBMITTED TO:**

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING (ACEM)

KUPONDOLE, KATHMANDU

LALITPUR, NEPAL

MAY 26, 2019.

# Acknowledgement

In all humility and gratitude, we are over-helmed to acknowledge our profundity to all those who helped us put our ideas and assigned work well above the level of simplicity and into something concrete.

To our HOD, Mr. Yuvraj Siwakoti, we would like to express our sincere gratitude for the continuous support for our Minor Project, for his patience, motivation, enthusiasm and immense knowledge.

We would also like to show our gratitude to the Mr. Pradip Khanal for sharing his pearls of wisdom with us during the course of this project. We are very thankful to our mentors for their valuable suggestions, guidance and encouragement.

Finally, we would also like to express our gratitude to all our friends and family for their moral support, cooperation and encouragement without which the completion of the proposal would be incomplete.

Ajay Pakhrin Lama

Ayub Shrestha

Jeewan Sharma

Manjil Munankarmi

# Abstract

The student performance prediction is a very important step in improving the educational process of a student. The early prediction of the student performance will help for further development in the educational process of a student. However, predicting student performance is becoming challenging day by day due to many factors which might affect in the performance of a student. So, by applying data mining techniques and using various algorithms on the datasets of the student, the performance of a student can be predicted more accurately.

In this proposed project, we attempt to implement Educational Data Mining techniques to predict the performance by extracting useful information and patterns from database of the students. In this proposed project, we would be using Naïve Bayesian classification algorithm on educational dataset of the students. The datasets were collected from online source. Two datasets are provided regarding the performance in two distinct subjects. The characteristics of the datasets is multivariate and consists of 33 attributes with 649 instances. The datasets don’t have any missing values [1].

# List of Figures

[Figure 1:System Flow Diagram 6](#_Toc9756233)

[Figure 2: Use Case Diagram 7](#_Toc9756234)

[Figure 3: DFD Level 0 8](#_Toc9756235)

[Figure 4: DFD level 1 8](#_Toc9756236)

[Figure 5: Waterfall Model 9](#_Toc9756237)

# List of Table

[**Table 1: Gantt Chart** 14](#_Toc9756359)

**Table of Contents**

[Acknowledgement iii](#_Toc9760788)

[Abstract iv](#_Toc9760789)

[List of Figures v](#_Toc9760790)

[List of Table vi](#_Toc9760791)

[1. Introduction 1](#_Toc9760792)

[2. Objectives 2](#_Toc9760793)

[3. Literature Review 3](#_Toc9760794)

[4. Requirement Analysis 4](#_Toc9760795)

[4.1 Software Requirements: 4](#_Toc9760796)

[4.2 Functional Requirements 4](#_Toc9760797)

[4.3 Non-Functional Requirements 4](#_Toc9760798)

[4.4 Feasibility Study 4](#_Toc9760799)

[4.4.1 Economic Feasibility: 5](#_Toc9760800)

[4.4.2 Technical feasibility: 5](#_Toc9760801)

[4.4.3 Operation Feasibility: 5](#_Toc9760802)

[5. System Design and Architecture 6](#_Toc9760803)

[5.1 System Flow Diagram 6](#_Toc9760804)

[5.2 Use Case Diagram 7](#_Toc9760805)

[5.3 Data Flow Diagram 8](#_Toc9760806)

[6. Methodology 9](#_Toc9760812)

[6.1 Software Development Approach 9](#_Toc9760813)

[6.2 Algorithm 9](#_Toc9760814)

[7. Time Schedule 14](#_Toc9760815)

[7.1 Gantt Chart: 14](#_Toc9760816)

[8. Expected Output 15](#_Toc9760817)

[9. References 16](#_Toc9760818)

# Introduction

Education is one of the fundamental infrastructures of development and lifestyle. In the present context. There are many Schools, Colleges and Universities facilitating students with various programs and degrees. Students enroll in such programs with the ambition of achieving a degree and pursuing a career. Students can deviate from the process of learning easily. Many factors are responsible for causing them to stray away from their studies. This is a major problem for Schools, Colleges and Universities. Now, due to major technological advancement Educational Data mining has been introduced.

Data mining refers to the process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data. Educational Data Mining is simply data mining implemented in the field of education. Educational Data Mining is an evolving discipline concerned with developing ways for exploring the unique types of data that are derived from educational environment and using these ways to understand students properly and the environment in which they learn on. It facilitates in ways such as generating patterns of student results, factors affecting the student performance, predicting the performance of students based on past records and so on. Many Universities and colleges perform adopt various measures to evaluate on the student’s data and reach to a certain conclusion on the ongoing performance and status of an individual student.

In this proposed project, Prediction of student performance based on their previous records which can be done with the help of certain Data mining techniques. In Predictive Data Mining, Classification and Regression can be used. We have opted to implement Naïve Bayes Classification Algorithm. With the available past data records of students with different attributes, the performance grade of a particular student can be predicted with this algorithm.

# Objectives

Student Performance Prediction based on Bayes Classification allows the evaluation of the past student data records available and predict the future performance or grade of the student. Some of the major objectives of our proposed work are:

* To predict the grade/performance of the student based on the past records.
* To provide the system with past data to get output of future data

# Literature Review

There are many research and papers that have been published that encompass the Student Performance Prediction Systems and Algorithms. A paper was published on Student Performance Prediction using KNN and Naïve Bayesian by Amra and Maghari for International Conference on Information Technology (ICIT) in the year 2017. The comparison between two classification algorithms KNN and Naïve Bayes for performance prediction were conducted to figure out the one with the highest accuracy. [2] The paper concluded that Naïve Bayes had the highest accuracy among the two algorithms proving to have strong relation with the features that affect the performance of the Students. A paper was published on Prediction of Students Performance using Educational Data Mining by authors Devasia T., Vinushree T P, & Hedge V. in the year 2016 for International Conference on Data Mining and Advanced Computing. The paper included a proposed system which consisted of a web-based application which implemented Naïve Bayesian mining technique. The system used data of 700 students with 19 attributes. The student registration number, marks of 10th grade, 12th grade, degree marks in each semester, assignment, gender, parent’s education, income, etc. were taken as attributes [4]. The system predicted the grades of the students with these inputs using Naïve Bayes Mining Technique and the output was represented in a diagram. Data Mining Techniques for Predicting Student Performance by authors Shaleena K.P. & Paul S. was published on International Conference on Engineering and Technology [8]. Prediction System for Student Performance using Data Mining Classification by Patil, R., Salunke, S., Kalbhor, M., & Lomte, R. was published on Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) in the year 2018. These papers focused on the different data mining techniques, algorithms and their feasibility Study [6]. Naive Bayes Using to Predict Students Academic Performance at Faculty of Literature by Pujianto, U., Azizah, E. N., & Damayanti, A. S. for 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) in the year 2017 was also done [7]. These papers would be an immense help in the completion of our proposed project.

# Requirement Analysis

## Software Requirements:

In the project we plan on developing the system with the use of following Programming language software:

Python: It provides a vast library for data mining and predictions.

MySQL: For the database that would be used in the system.

## Functional Requirements

* The system must provide the predicted performance grade of the student that has been selected by the user.
* The system must have an easy to use interface for using the system for all the users.
* System should provide a graphical interface for the prediction generated and for the User Interface.
* The Admin must be able to update/modify the Database of the System.
* The Dataset of the students must be available for the system.

## Non-Functional Requirements

* Scalability

The proposed system should be scalable and accommodate a number of users based on the implementation without any changes.

* Security

The proposed system should be safe from unauthorized user and access. The database of the system must also be safe.

* Portability

For the portability of the system, it should run properly in different hardware.

* Maintainability

The Proposed system should be easy to maintain.

## Feasibility Study

The proposed work has to be operational and be feasible to every running environment to achieved its goal. Following points describes feasibility of the project:

### Economic Feasibility:

Expenditure for the proposed project required is only computational power. It doesn’t require anything complex rather easily available. The required Dataset can be found from an online source and computational need from laptops.

### Technical feasibility:

The proposed project works with the database of students having different attributes which will be available with the institutions that would implement the system. So, there would be no problem in the availability of data resource for the system. The completion and development would be difficult but its operation would be quite feasible. The dataset can be trained with the available computational power of i5 or latest i7 processors. And the software would be easily operable.

### Operation Feasibility:

For a project to be feasible it has to be implemented in real world. It proposed project analyzes the available data to predict the future outcome which has gained a lot of importance. The analysis for the performance prediction can be implemented in different educational institutions for specific student progress reason. So, the proposed project would be feasible operationally.

# System Design and Architecture

## System Flow Diagram

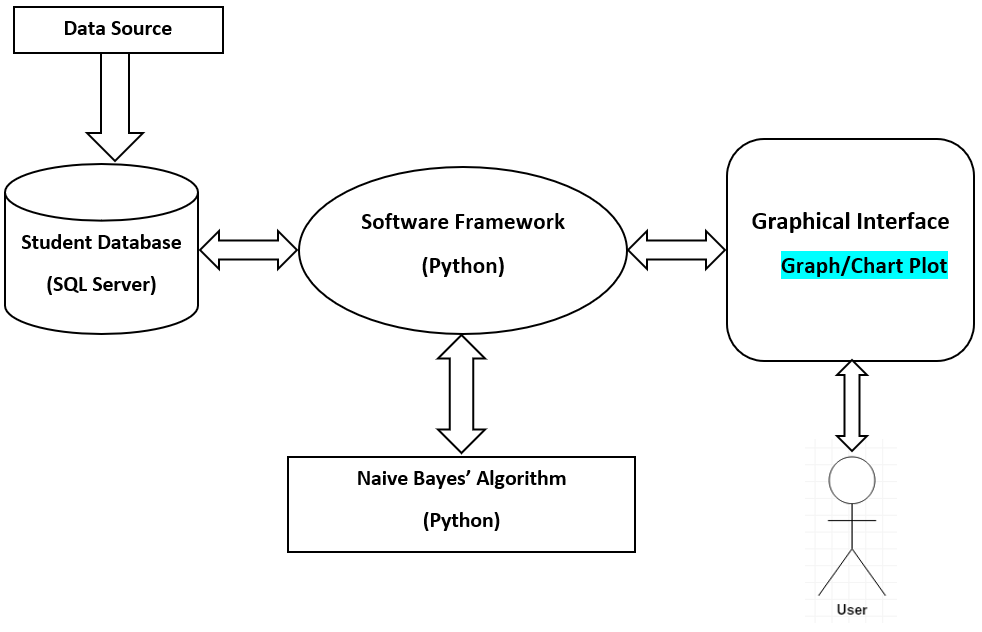


Figure 1: System Flow Diagram

## Use Case Diagram

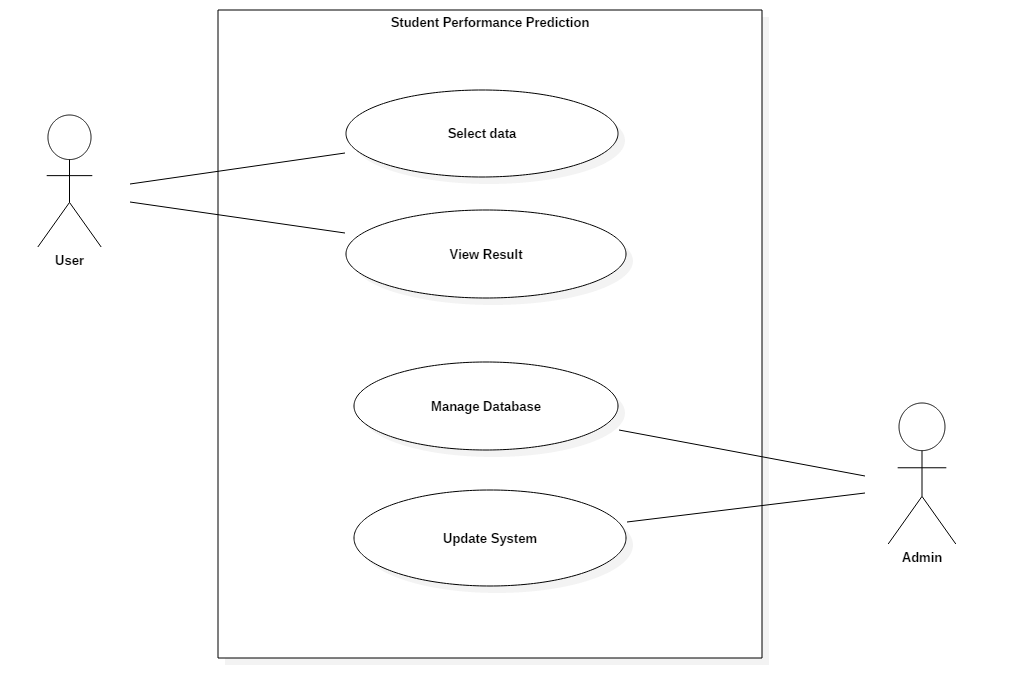


Figure 2: Use Case Diagram

## Data Flow Diagram

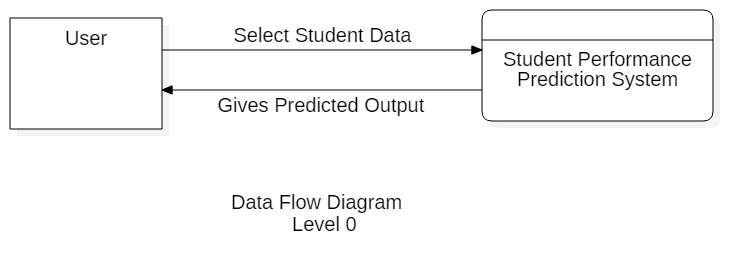


Figure 3: DFD Level 0

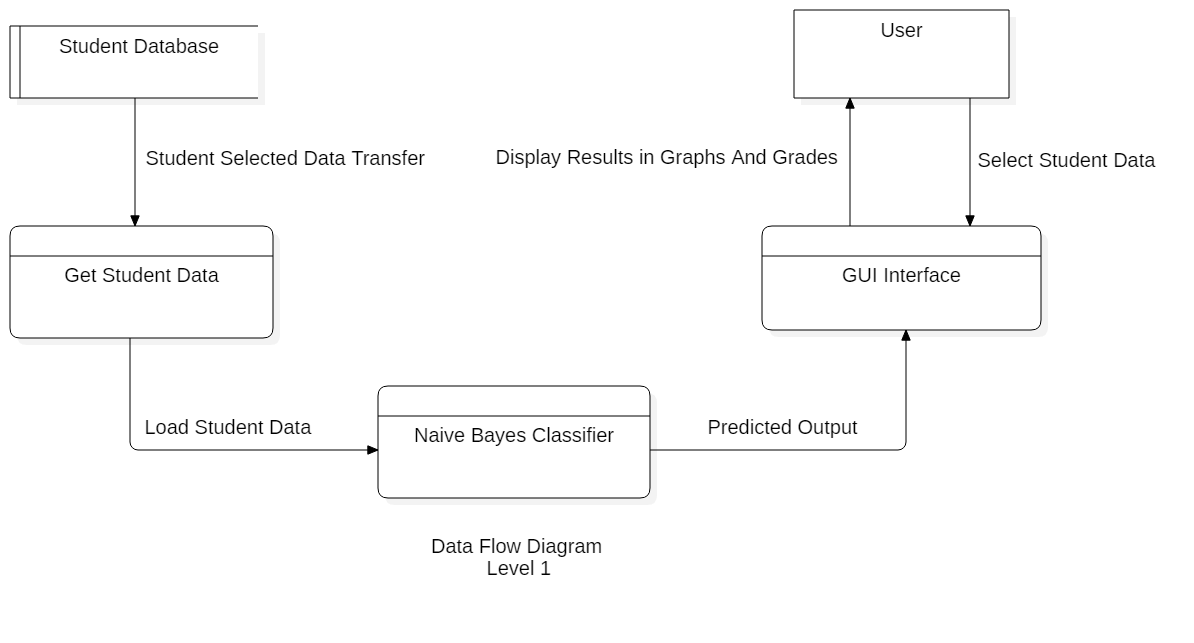


Figure 4: DFD level 1



# Methodology

## Software Development Approach

The waterfall model is a linear, sequential approach to the software development life cycle (SDLC) which emphasizes a logical progression of steps. Similar to the direction water flows over the edge of a cliff, distinct endpoints or goals are set for each phase of development and cannot be revisited after completion [9]. The goal is to provide a system with overall functionality.

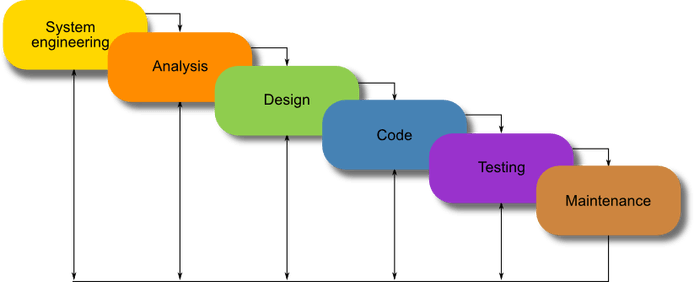


Figure 5: Waterfall Model (source: Retrieved from Study.com)

As in our proposed system, we need to complete a phase before beginning the next phase, and each phase needs to be documented properly after its completion, for this reason we choose this model which enabled us to force structured and disciplined project. This model will also help us to allow for early design and specification changes can be made easily.

## Algorithm

Naïve Bayes Algorithm is a probabilistic machine learning algorithm which can be widely used in various classification tasks which is based on Bayes Theorem. The term naïve is given because it assumes the data that is given to the model are independent of each other, that is they have independent distribution. So, if we change the value of one feature than it doesn’t affect the value of other features used in the algorithm.

There are many applications of Naïve Bayes Algorithm like real time prediction, multi class prediction, text classification, spam filtering, recommendation system etc.

However, the algorithm is getting its popularity because of its robustness ability to noise and outliers as well as to irrelevant attributes. The missing values are easily handled. The predictions are made real-quick because of which, it is easily scalable and is traditionally the choice for real-world application. Before going to Naïve Bayes theorem, we have to understand conditional probability and Bayes theorem.

Conditional Probability is a measure of the probability of an event given that the another even has occurred.  If the event of interest is ‘A’ and the event ‘B’ is known or assumed to have occurred, "the conditional probability of ‘A’ given ‘B’", or "the probability of ‘A’under the condition ‘B’", is usually written as P(A|B) [3].

For example, if we pick a card from the deck, we can guess the probability of getting a king given the card is a heart. We already have a condition that the card is a heart. So, the denominator is 13 and not 52. And since there is only one king in spades, the probability it is a king given the card is a heart is 1/13 =0.077

Mathematically, conditional probability of A given B can be computed as:

P(A|B) = P(AՈB)/P(B)….(1)

P(B|A) = P(AՈB)/P(A)….(2)

And the Bayes theorem used for Naïve Bayes is derived from above given equations.

Bayes theorem describes the [probability](https://en.wikipedia.org/wiki/Probability) of an [event](https://en.wikipedia.org/wiki/Event_(probability_theory)), based on prior knowledge of conditions that might be related to the event. The Bayes Rule is a way of going from P(A|B), known from the training dataset, to find P(B|A).

We have,

P(A|B) = P(AՈB)/P(B)

which is known from training data.

P(B|A) = P(AՈB)/P(B)

which is to be predicted from test data.

Here, A is evidence and B is outcome.

And the Bayes Theorem is now given as:

P(B|A) = [P(A|B) \* P(B)]/P(A)

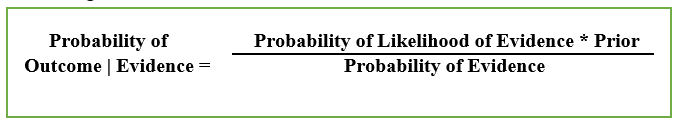
Hence, Bayes theorem gives us the formula for probability of B given A. However, in real world scenario, we will encounter with multiple values of A. So, when the features are independent, we can extend the Bayes theorem to Naïve Bayes theorem. Due to the naïve assumptions that the multiple variables of A are independent of each other, it is called Naïve Bayes.

Assuming, A’s are independent, we have:

where, K is class of B

Becomes Naïve Bayes, i.e.

The above given formula can be understood as:



Here, Probability of Evidence is same for all classes of B.

In the technical jargon, the LHS of the equation is simply the posterior probability. In the numerator of RHS, we have two terms. Here, Probability of likelihood of evidence is conditional probability of each A’s given B is of particular class ‘c’.

Since all the A’s are assumed to be independent of each other, we can simply multiply the likelihoods of all the A’s and call it the ‘Probability of Likelihood of Evidence’ which is known from the training datasets by filtering records where B= ‘c’.

The second term is called ‘Prior’ which is the overall probability of B=c, where c is class of B. Simply, Prior = count(B=c)/n\_Records.

Sometimes, when we have a model with many features, the entire probability will be zero because one of the features’ value is zero. To avoid this situation, we increase the count of the zero to one in numerator so that we will avoid the entire probability to be zero, this condition is called Laplace Correction [5].

Algorithm for Naïve Bayes:

Step 1: Scan the student data set

Step 2: Calculate the probability of each attribute value. [n, n\_c, m, p]

Step 3: Apply the formulae

where,

n = the number of training data item for which v = vj

nc = number of examples for which v = vj and a = ai

p = a priori estimate for P(ai,vj)

m = the parallel size of the sample

Step 4: Multiply the probabilities by p

Step 5: Compare the values and classify the attribute values to one of the predefined set of class [2].

# Time Schedule

## Gantt Chart:

**Table 1: Gantt Chart**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.N. | Work | Time (Month) | | | | | |
| Jestha  (10-12) | Jestha  (12-32) | Ashad  (1-15) | Ashad  (15-30) | Shrawan  (1-15) | Shrawan  (15-30) |
| 1 | Proposal |  |  |  |  |  |  |
| 2 | System Design |  |  |  |  |  |  |
| 3 | Software Development |  |  |  |  |  |  |
| 4 | Testing |  |  |  |  |  |  |
| 5 | Documentation |  |  |  |  |  |  |

# Expected Output

The system shall give a result of predicted performance grade of a student.

And the predicted result shall be in graphical manner.

# References

1. (2019, May 25). Retrieved from UCI: https://archive.ics.uci.edu/ml/datasets/Student+Performance
2. Amra, I. A. (2017). *Students performance prediction using KNN and Naïve Bayesian.* 8th International Conference on Information Technology (ICIT).
3. *Conditional Probability*. (n.d.). Retrieved from Wikipedia.com: https://en.wikipedia.org/wiki/Conditional\_probability
4. Devasia, T. V. (2016). *Prediction of students performance using Educational Data Mining.* International Conference on Data Mining and Advanced Computing (SAPIENCE).
5. *How Naive Bayes algorithm works?* (n.d.). Retrieved from Machine Learning Plus: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/?fbclid=IwAR0-aj4YE37OCb5eboT\_3K60tglNC5qibJQp5UPTGick-PC53CVqVNlXuRk
6. Patil, R. S. (2018). *Prediction System for Student Performance Using Data Mining Classification.* Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
7. Pujianto, U. A. (2017). *Naive Bayes Using to Predict Students Academic Performance at Faculty of Literature .* 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) .
8. Shaleena, K. P. (2015). *Data mining techniques for predicting student performance.* IEEE International Conference on Engineering and Technology (ICETECH).
9. *Waterfall Model*. (n.d.). Retrieved from Study.com: https://study.com/academy/lesson/waterfall-model-methodology-phases.html