PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE USING MULTIPLE LINEAR REGRESSION

CHAN MYAE MYINT ZU

M.C.Sc. **JUNE 2022**

PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE USING MULTIPLE LINEAR REGRESSION

By

CHAN MYAE MYINT ZU

B.C.Sc.(Honous)

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science (M.C.Sc.)

University of Computer Studies, Yangon

JUNE 2022

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin**, Rector of the University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw**, Professors, Faculty of Computer Science, the University of Computer Studies, Yangon, for their superior suggestion, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr. Kyi Lai Lai Khine,** Lecturer, Faculty of Computer Science , the University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards the completion of this thesis.

I also wish to express my deepest gratitude to **Daw Win Lai Lai Bo**, Assistant Lecturer, English Department, the University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

I especially thank to my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

ABSTRACT

Nowadays, the education system has been changed from teacher-centered approach to learner-centered approach. Thus, students' related features need to analyze to predict students' academic performance to provide the active learning in educational system. In this system, Multiple Linear Regression (MLR) is applied to predict the students' academic performance on the UCI Machine Learning Repository's student performance data set. The student's academic performance prediction model is built by using Multiple Linear Regression method. The purpose of this system is to predict the students' final grade based on previous grades and relevant features. Feature selection method is used to reduce the number of input variables that are believed to be the most useful to a regression model in order to predict the target variable. For evaluating the result of prediction model performance applied on students' academic performance model use four different measures: accuracy, precision, recall and f-measure. This system is implemented using C# programming language with Microsoft Visual Studio IDE and Microsoft SQL Server as Database Engine.

Key Words: Multiple Linear Regression (MLR), learner-centered approach, UCI

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Education is a pivotal component in our general public society. It is a critical variable for accomplishing a drawn-out financial advancement. Further developing understudy's scholastic accomplishment began with working on understudy's way of behaving and urges the understudy to take part in the homeroom. The quality and amount of understudies' mental and behavioral reactions close to home and conduct responses to the educational experience as well as to in-class/out-of-class scholarly and social exercises are critical to accomplish fruitful learning results. The separated information will assist schools with upgrading understudy's scholarly achievement and will assist executives with further developing learning frameworks.

Analysts and designers in the training local area started to investigate the possibilities in taking on comparable to procedures for acquiring knowledge into learning the executives and conveyance. Two significant regions are presently being worked on towards the incorporation and investigation of huge information capacities in the instructive climate. They are Educational Data Mining (EDM) and Learning Analytics (LA).

Regression analysis could be helpful in advancing educational system by enabling scholars to better understand variables that may predict specific outcomes such as student achievement or program retention. There has been an increased interest in regression techniques within education to allow researchers to determine variables that may predict a specific outcome.

1.2 Objectives of the Thesis

The main objectives of this thesis are:

- to allow teachers for predicting the student grades easily
- to apply the multiple linear regression model in this study to predict the students' academic performance
- to provide the important and relevant features that may have an impact on student's academic performance
- to show the teaching, learning outcomes that is based on the students' other related features
- to improve the quality of students and enhance school resource management

1.3 MLR for Student's Academic Performance Evaluation

Predicting students' academic performance has long been an important area of research in education. Most existing literature had used of traditional statistical methods that run into the problems of over fitted models, inability to effectively handle large numbers of participants and predictors, and inability to pick out non-linearities that may be present. Regression-based ML methods that can produce highly interpretable yet accurate models for new predictions are able to provide some solutions to the aforementioned problems.

1.4 Motivations

Predicting Students' Academic Performance (SAP) is one of the significant examination regions in Higher Learning Institutions. Scientists showed that Data Mining Techniques are generally utilized in instructive field to view as new and concealed designs from understudy's information. A powerful prescient model requires great info information (boundary), reasonable Data Mining techniques and devices for the information investigation. The primary thought of this framework is to introduce the model by utilizing

the Multiple Linear Regression method which can be utilized for forecast of understudy's exhibition in scholastic by utilizing their grades and related highlights.

Feature selection is a way to reduce the number of features and, hence, reduces the computational complexity of the model. Feature selection is used many times because it is very useful to overcome the over fitting problem. It helps us in determining the smallest set of features that are needed to predict the response variable with high accuracy.

1.5 Organization of the Thesis

The thesis is organized in five chapters. They are as follows:

In Chapter 1, introduction of the system, objectives of the thesis, MLR for students' academic performance evaluation, motivations and organization of the thesis are described. **Chapter 2** presents the background theory of Multiple Linear Regression (MLR) method and feature selection (Chi-Square). **Chapter 3** describes the proposed system of prediction of students' academic performance using multiple linear regression, **Chapter 4** expresses the design and implementation of the proposed system, academic student performance evaluation with feature selection and experimentation of the system. Finally, **Chapter 5** presents the conclusion, benefits, limitations and further extensions of the system.

CHAPTER 2
BACKGROUND THEORY

Firstly, this chapter expresses a detail explanation of Regression analysis and feature selection. Regression is the oldest and most widely used multivariate technique in the social sciences. Regression is an example of dependence analysis in which the variables are not treated symmetrically. In regression analysis, the object is to obtain a prediction of one variable given the values of the others.

2.1 Regression Analysis

Regression analysis is used to address inquiries concerning how one variable relies upon at least one different factor. For example, does less calories associate with cholesterol level, and does this relationship rely upon different variables, like age, smoking status, and level of activity? Relapse models can respond to these inquiries. They depict the connection between a reliant variable which is diet in our model, and a free factor or factors which are cholesterol level, age, smoking status, and level of exercise.

2.1.1 Linear Regression Model

A linear regression model depicts the connection between a reliant variable “y”, and at least one free factors “X”. The reliant variable is likewise called the reaction variable. Autonomous factors are additionally called informative or indicator factors. Nonstop indicator factors are likewise called covariates, and downright indicator factors are additionally called factors.

2.1.1.1 Definitions, Basic Concepts, and Examples

A **regression model** is a numerical condition that portrays the connection between at least two factors; in some cases we call it relapse condition. Also, by straight relapse model, we mean a model that expects a direct connection between at least two factors. Sorts of straight relapse models:

- Straightforward Linear Regression: When we consider the connection between one ward variable and one autonomous variable, we utilize Simple Linear Regression.
- Various Linear Regression: When we consider the connection between one ward variable and more than one autonomous variable, we utilize Multiple Regression.
- Multivariate direct relapse: When we consider the connection between more than one ward variable and one autonomous variable, we utilize Multivariate Regression.
- Multivariate Multiple Linear Regression: When we consider the connection between more than one ward variable and more than one autonomous variable, we utilize Multivariate Regression.
- Connection examination is a factual methodology used to gauge the strength of the relationship among factors. The term connection alludes frequently to the direct relationship between two amounts or factors, or at least, the inclination for one factor to increment or abatement as different increments or diminishes in an orderly fashion pattern or relationship. Connection and relapse investigation are connected as in the two of them manage connections among factors.
- The connection coefficient (additionally called the Pearson direct relationship coefficient) is a mathematical file of the strength of connection between two factors. Upsides of the relationship coefficient are generally between - 1 and +1. A connection coefficient of +1 demonstrates that the two factors are impeccably related in a positive direct sense. A relationship coefficient of - 1 demonstrates that two factors are impeccably related in a negative direct sense. A connection coefficient of 0 demonstrates that there is no straight connection between the two factors.
- The populace relationship coefficient "ρ (rho)" gauges the strength and course of the straight relationship between the factors.
- The example connection coefficient "r" is a gauge of ρ and is utilized to quantify the strength of the straight relationship in the example perceptions. The

nearer r is to +1, the more grounded the positive connection is. The nearer r is to - 1, the more grounded the negative connection is. The two factors are impeccably corresponded if |r| is equal to 1 precisely. A worth of zero for r doesn't intend that there is no relationship.

Relapse models are broadly utilized for forecast that can anticipate the worth of a reaction variable from information on the upsides of at least one logical factors. The following are a few models:

- A meteorologist might conjecture that it will rain tomorrow.
- A leader of an insurance agency might foresee that there will be more street mishaps and losses one year from now.
- An analyst in schooling might guarantee that instructive achievement relies upon insight, financial and social class of an understudy.

2.1.1.2 Regression Equation

The basic relapse condition takes the logarithmic structure for a straight line: y = mx + b, where m is the slant of the line, and b is the y-catch. It is known from variable based math that a line is recognized by its incline (the point of the line depicting the adjustment of y per unit x) and catch (where the line crosses the y pivot). So relapse portrays the connection among x and y with simply such a line. The condition or recipe is utilized for the straight line that limits the complete mistake.

Based on the relapse conditions, the worth of the reliant variable is foreseed by using fixed values of the free variable(s). For the most part, when we have k indicator factors, the recipe of relapse condition takes the structure

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = x_i^T \beta + \varepsilon_i, i = 1,2, \ldots\ldots, n \qquad (2.1)$$

where T denotes the transpose, so that $x_i^T\beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in vector form as:

$$Y = X\beta + \varepsilon, \qquad (2.2)$$

Where

- Y is the segment vector of n perceptions of the reaction variable. The choice concerned with variable in an informational index is displayed as the reliant variable and which are demonstrated as the autonomous factors might be founded on an assumption that the worth of one of the factors is brought about by, or straightforwardly affected by different factors.
- X is known as the plan lattice comprising of section vectors of perceptions on the indicator factors.

Remark:

Typically a consistent is incorporated as one of the indicator factors. For instance we can take $x_{i1} = 1$ for I = 1, ..., n. The comparing component of β is known as the capture. Numerous measurable derivation strategies for straight models require a catch to be available, so it includes many times regardless of whether hypothetical contemplations recommend that its worth ought to be zero. Hence, the relapse condition will take the structure:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = x_i^T \beta + \varepsilon_i$, where β_0 represents the intercept parameter.

In some cases, one of the indicator factors can be a non-straight capacity of another indicator as in polynomial relapse. The model remaining parts straight for however long it is direct in the boundary vector β.

β is the segment vector of coefficients to be assessed. Measurable assessment and induction in straight relapse centers around β.

ε is the blunder term, or commotion. This variable catches any remaining elements which impact the reliant variable yi other than the regressors xi.

The specific populace relapse line (upsides of the relapse coefficients), and the objective of straight relapse techniques is to view as the "best" decisions of the assessed values for the constants $\beta_0, \beta_1, \beta_2, \ldots, \beta k$ to make the relapse equation exactly. The relapse line that we acquire from an example gives a gauge of the populace relapse line. The assessed relapse model:

$$E(y_i) = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_k} x_{ik} \text{ (population)} \qquad (2.3)$$

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_k} x_{ik} \text{ (sample)} \quad (2.4)$$

where

$\hat{y}$: is the estimated y value.

$\hat{\beta}_i$: are the estimated values of the regression coefficients.

$\hat{\beta}_0$: is the estimation of the regression intercept.

The individual random error term ε_i have a mean of zero. A **residual** (or the error term) is the difference between the observed response y and the predicted response $\hat{y}$, $\varepsilon_i = y_i - \hat{y}_i$.

2.1.1.3 Linear Regression Assumptions

The another one is significant suspicions made by standard straight relapse models with standard assessment strategies (for example common least squares):

- Mistake values (ε) are genuinely autonomous. This expects that the mistakes of the reaction factors are uncorrelated with one another. A few strategies are equipped for taking care of related blunders. In spite of the fact that they require essentially more information, some kind of regularization is not utilized to incline the model towards accepting uncorrelated mistakes. Bayesian direct relapse is an overall approach to dealing with this issue.
- The likelihood dispersion of the blunders is ordinary with mean zero.
- The likelihood dispersion of the blunders has steady difference.
- The free factors are estimated with no blunder. The noticed upsides of x are thought to be a bunch of known constants. At the end of the day, the indicator factors are thought to be without blunder.
- Linearity: The fundamental connection between the x variable and the y variable is direct. All in all, the mean of the reaction variable is a direct mix of the boundaries (relapse coefficients) and the indicator factors.
- Factors are treated as fixed values, linearity is truly just a limitation on the boundaries. The indicator factors themselves can be randomly changed. Actually numerous duplicates of a similar fundamental indicator variable can be added and each one changes in an unexpected way.

- The indicators are directly free, that is to say, communicating any indicator as a straight mix of the others is beyond the realm of possibilities.

At that times one of the indicators can be a non-straight capacity of another indicator or of the information as in polynomial relapse and portioned relapse. As long as the model is linear in the parameter vector β, it remains linear. For instance, it consider what is happening and where a little ball is being thrown up in the air. Afterward we measure its levels of climb hey at different minutes in time ti. Physical science shows by disregarding the drag, the relationship can be demonstrated as:

$$h_i = \beta_1 t_i + \beta_2 t_{i2} + \varepsilon_i, \qquad (2.5)$$

where β1 decides the underlying speed of the ball, β2 is relative to the standard gravity, and εi is estimation mistakes. Straight relapse can be utilized to appraise the upsides of β1 and β2 from the deliberate information. This model is non-straight in the time variable, however it is direct in the boundaries β_1 and β_2. Assuming we take regressors xi = (x_{i1}, x_{i2}) = (t_i, t_{i2}), the model takes on the standard structure:

$$h_i = x_i T\beta + \varepsilon_i \qquad (2.6)$$

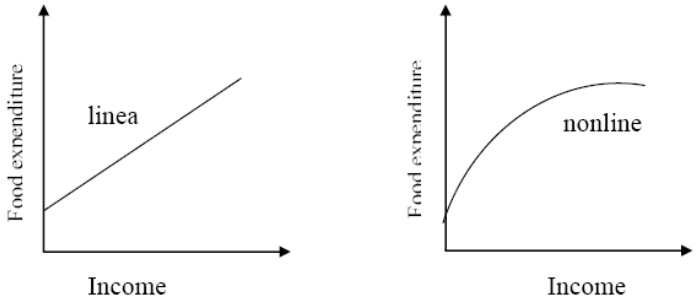
Remark: In statistics, the independence and normality assumptions about the errors are called the Gauss–Markov conditions.

2.1.2 Nonlinear Regression

The supposition of linearity in relapse models expects that the relationship among the factors should be straight. That is a straight-line connection between the factors (the reaction variable is a direct blend of the boundaries and of the indicator factors). In direct relapse, we attempt to find the best straight line fitted to information while the genuine relationship that we need to display is bended. For instance, in the event that something is developing dramatically and implies developing at a consistent rate, the connection among x and y is bend. To fit something like this, we really want non-straight relapse which is a type of relapse examination where observational information are displayed by a capacity which is a nonlinear blend of the model boundaries and relies upon at least one autonomous factors.

Nonlinear relapse is an overall strategy to fit a bend through the information. It fits information to any situation that characterizes y as an element of x and at least one boundaries. It observes the upsides of those boundaries that create the bend which comes nearest to the information (limits the amount of the squares of the upward distances between data of interest and bend).

Both straight and nonlinear relapse tracks down the assessed upsides of the boundaries (slant and block for direct relapse) that make the line (in straight relapse) or the bend (in nonlinear relapse) which come as close as conceivable to the information. The two charts in Figure 2.1 show a direct and a nonlinear connection between the reliant variable food consumption and the free factor pay. A direct connection among pay and food consumption displayed in Figure 2.1 (a) shows that pay expands and the food use generally increments at a steady rate. A nonlinear connection among pay and food consumption, as displayed in Figure 2.1 (b), shows that if amount is increased and the food also increase. Although, after a point, the pace of expansion in food consumption is lower for each resulting expansion in pay.



2.1 (a) Linear relationship 2.1(b) non-Linear relationship

Figure 2.1: Relationship between food expenditure and income

Numerous connections in science and different areas of science don't follow a straight line. To dissect such information, you have two options:

- Utilize nonlinear relapse strategies: Like the normal least squares (OLS) move toward which gives the best fit bend that limits the amount of squared residuals.
- Do numerical changes to compel the information into a straight relationship. Then utilize direct relapse. Although these strategies are usually utilized and

they ought to be kept away from it. They are less exact than nonlinear relapse and are no simpler.

2.1.3 The Method of Least Squares

Least squares straight relapse (otherwise called "standard least squares", "OLS", or frequently "least squares"), is quite possibly the most essential and most normally utilized expectation method known as humanity with applications in fields as assorted as measurements, finance, medication, financial aspects, and brain research.

The numerical idea of least squares is the reason for a few strategies to fit particular kinds of bends and surfaces to information. Issues of fitting bends and surfaces have a set of experiences spreading over a few centuries. The fundamental thought of the strategy for least squares is straightforward. It might appear to be strange when a few group measure similar amount and they typically don't get similar outcomes. As a matter of fact, a similar individual estimates similar amount a few times, the outcomes will change.

Prominence of Least Square: Least squares is such an exceptionally famous strategy that when individuals utilize the expression "direct relapse" frequently, they are as a matter of fact alluding to "least squares relapse". A significant part of the utilization of least squares can be credited to a few elements:

- It is one of the earliest broad expectation techniques known to humanity.
- Its execution on present day PCs is productive, so it tends to be immediately applied issues with many highlights and a huge number of data of interest.
- It is simpler to investigate numerically than numerous other relapse strategies.
- It is straightforward for non-mathematicians to comprehend at an essential level.
- It is the ideal strategy from a specific perspective in specific exceptional cases.

Specifically, assuming that the framework being concentrated really is direct with added substance that free typically dispersed blunders (for example with mean zero and consistent difference). Then the constants addressed for by least squares are as a matter of fact the most probable coefficients to have been utilized and to create the information.

2.1.4 Simple Linear Regression Model

The straightforward direct relapse model which includes just a single ward variable (y) and one autonomous variable(x) states that the genuine mean of the reliant variable changes at a steady rate as the worth of the free factor increments or diminishes. The condition of the line relating y to x is known as the straightforward direct relapse condition.

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.7)$$

where,

y: the dependent (response) variable.

x: the independent (explanatory) variable.

β_0: the y-intercept, the value of y when x = 0.

β_1: the slope, the expected change in y relative to one unit increase in x.

ε: is the random error.

The estimate of the simple linear regression equation is given by substituting the least squares estimates into equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{y}$ is the expected value of y for a given value of x.

The straightforward direct relapse model has two coefficients β0 and β_1, which are to be assessed from the information, that is we need to find $\hat{\beta}_0$ and $\hat{\beta}_1$. Assuming there was no arbitrary blunder in yi, any two data of interest (x_i, y_i) could be utilized to settle expressly for the upsides of the boundaries. The irregular variety in y may causes each sets of noticed information focuses to give various outcomes. (All assessments would be indistinguishable provided that the noticed information fell precisely on the straight line). The technique for least squares will be utilized to consolidate all the data to give one arrangement which is "ideal" by some measure.

The least squares assessment technique utilizes the basis that the best arrangement should give the littlest conceivable amount of squared deviations of the standard noticed yi from the evaluations of their actual means given by the arrangement. The best fit relapse line is the line that limits the amount of mistakes.

Integrity of Fit of the model: Once fit a model, a characteristic inquiry rings a bell, how great is the fit? Are the illustrative factors helpful in expectation? To respond to

these inquiries, we really want to survey the relapse model either utilizing t-test or involving ANOVA for relapse. By and large, the motivation behind examination of fluctuation (ANOVA) is to test for huge contrasts between implies. However, in relapse models, it comprises of estimations that give data about degrees of inconstancy inside a relapse model and structure a reason for trial of importance.

It is not difficult to demonstrate the way that the complete variety of the reaction variable y can be decayed into two sections: the leftover variety of y (blunder amount of squares (SSE)) and the made sense of variety of y (relapse amount of squares (SSR)). Consider the total sum of squares:

$$\sum(y_i - \bar{y}) = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \quad (2.8)$$

$\sum(y_i - \bar{y}) = \sum(y_i - \hat{y}_i) + \sum(\hat{y}_i - \bar{y})^2$, which we usually rewrite as:

$$SST = SSE + SSR$$

- SST represents the "complete amount of squares" this is basically the absolute variety in the informational index. That is the absolute variety of food use.
- SSR means "amount of squares because of relapse" - this is the squared variety around the mean of the assessed food that are used. This is here and there called the complete variety made sense of by the relapse.
- SSE means "amount of squares because of mistake" - this is essentially the amount of the squared residuals, and the variety in the y variable remains parts unexplained in the wake of considering the variable x.

For the simple regression case, these are computed as:

$$SST = Syy = \sum y_i^2 - n\bar{y}^2 \quad (2.9)$$

$$SSR = \widehat{\beta_1} Sxy \quad (2.10)$$

$$SSE = SST - SSR \quad (2.11)$$

Each amount of squares can be isolated by a suitable consistent (levels of opportunity) to get the mean amount of squares because of relapse MSR, and the mean

amount of squares because of blunder MSE. It is frequently valuable to sum up the deterioration of the variety in y concerning an examination of difference (ANOVA). In such a case, the all out made sense of and unexplained varieties in y are changed over into fluctuations by partitioning by the proper levels of opportunity. This fosters a conventional strategy to test the decency of fit by the relapse line.

At first, the system set the invalid theory that the fit isn't great. At the end of the day, the theory is that the general relapse isn't critical and the informative variable can't make sense of the reaction variable in a palatable manner.

Table 2.1 ANOVA table for simple regression

Source of Variation (Source)	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F statistic
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Total	SST	n-1		

From the ANOVA table, we can undoubtedly close the general relapse that is huge at the 5% degree of importance, i.e., the OLS relapse line enough fits the information. Assuming the determined worth of the measurement falls in the basic district, the system reject the invalid speculation and presume that the relapse coefficient is huge. All in all, the system says that the informative variable affects the reaction variable. The basic locale (or the dismissal) not entirely settled by the worth of F-classified, Fα,1,n-2.

If the worth of the measurement falls outside the basic area, The system don't dismiss the invalid speculation and infer that the relapse coefficient isn't huge, i.e., the informative variable affects the reaction variable.

2.1.5 Multiple Linear Regression Model

This part presents another convoluted model, and fosters the typical conditions and answer for the ordinary conditions for a broader straight model including limited number of free factors. This present different relapse examination in framework documentation. In this model, the system think about the connection between one ward variable and more than one autonomous variable. The direct model for relating a reliant variable to k autonomous factors is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (2.12)$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_k} x_k \quad (2.13)$$

The numerous direct relapse can be considered an expansion of basic straight relapse, where there are k informative factors, or straightforward straight relapse can be thought as an exceptional instance of various straight relapse, where k = 1. Additionally, here we utilize the least squares technique to track down the assessed coefficients $\beta\hat{}_0$, $\beta\hat{}_1$, …,$\beta\hat{}_k$ that minimize the sum of squares $\Sigma(y - \hat{y})^2$

The method is to write the following formulas,

$$Sx_1y = \widehat{\beta_1} Sx_1x_1 + \widehat{\beta_2} Sx_1x_2 + \cdots + \widehat{\beta_k} Sx_1x_k$$

$$Sx_2y = \widehat{\beta_1} Sx_1x_2 + \widehat{\beta_2} Sx_2x_2 + \cdots + \widehat{\beta_k} Sx_2x_k$$

.

.

.

$$Sxky = \hat{\beta} S_1 xk + \widehat{\beta_2} Sx_2x_k + \cdots + \widehat{\beta_k} Sx_kx_k$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\overline{x_{11}} - \widehat{\beta_2}\overline{x_2} - \cdots - \widehat{\beta_k} x\overline{y_k} \quad (2.14)$$

These conditions are called ordinary conditions, so we can handle these conditions to track down the assessed coefficients $\beta\hat{}_0$, $\beta\hat{}_1$, …,$\beta\hat{}_k$. It is note that, in light of the fact that the ordinary conditions are straight and in light of the fact that there are however many conditions as obscure relapse coefficients (k+1), there is typically remarkable answer for the coefficients $\beta\hat{}_0$,$\beta\hat{}_1$, …,$\beta\hat{}_k$.

To represent how the technique for least squares work in this model, we will consider an exceptionally straightforward illustration of numerous relapse model on account of only two informative factors, that is the equation of the model: $\hat{y} = \beta\hat{}_0 + \beta\hat{}_1 x_1 + \beta\hat{}_2 x_2$. Through this model, we will perceive how to register the relapse coefficients by doing comparable advances like in basic relapse in a created way. **Goodness of Fit of the model:** As in straightforward relapse, we use ANOVA for relapse to test the decency of the different relapse models. If the general relapse isn't huge, it could be said that the informative factors can't make sense of the reaction variable in an agreeable manner.

$$SST = Syy = \sum y_i^2 - n\overline{y^2} \quad (2.15)$$

$$SSR = \widehat{\beta_1} Sx_1 y + \widehat{\beta_2} Sx_2 y \quad (2.16)$$

$$SSE = SST - SSR \quad (2.17)$$

Table 2.2 ANOVA table for multiple regression

Source of Variation (Source)	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F statistic
Regression	SSR	K	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-k-1	$MSE = \frac{SSE}{n-k-1}$	
Total	SST	n-1		

At 5% level of significance, the rejection region is determined by $F_{0.05,k,nk-1}$.

Multiple Regression Model in Matrix Notation: For the multiple regression model

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$, for all i= 1,2,…,n

where the error terms assume to have the following properties:

- $E(\varepsilon_i) = 0$.
- $Var(\varepsilon_i) = \sigma 2$ (constant).
- $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

The addendum I indicates the observational unit from which the perceptions on y and the k free factors were taken. The subsequent addendum assigns the autonomous variable.

The example size is meant with n, I = 1, . . . , n, and k indicates the quantity of free factors. There are (k + 1) assessed coefficients βˆj, j = 0, . . . , k when the straight model incorporates the assessed capture βˆ0.

Four grids are expected to communicate the direct model in lattice documentation: Y : the n×1 is segment vector of perceptions on the reliant variable y. X: the n × (k+1) is lattice comprising of a segment of ones which is named 1 trailed by the k section vectors of the perceptions on the free factors.

β: the (k+1) × 1 vector of parameters to be estimated.

Ɛ: the n × 1 vector of random errors

With these definitions, the linear model can be written as in equation,

$$Y = X\beta + \varepsilon \qquad (2.18)$$

or :

$$\underset{(n\times 1)}{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{pmatrix}} = \underset{(n\times(k+1))}{\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ . & . & & & . \\ . & & & \dots & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}} \underset{((k+1)\times 1)}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}} + \underset{(n\times 1)}{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{pmatrix}}$$

The assumptions become:

- $E(\varepsilon) = 0$
- $Cov(\varepsilon) = E(\varepsilon\, \varepsilon^T) = \sigma^2 I$

The components of a specific column of X which say line I are the coefficients on the relating boundaries in β. It is notice that β0 has the steady multiplier 1 for all perceptions; subsequently, the section vector 1 is the principal segment of X.

2.1.6 The normal equations and their solution

For the least square estimation, our main objective is to find a vector of parameters which minimizes the error sum of squares

$$SSE = \sum \varepsilon_i^2 = \varepsilon^T \varepsilon \quad (2.19)$$

In matrix notation, the normal equations are written as:

$$X^T X\hat{\beta} = X^T Y \quad (2.20)$$

The normal equations are always consistent and, hence, will always have a solution of the form

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.21)$$

If X^TX has an inverse, then the normal equations have a unique solution given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.22)$$

here $\hat{Y} = X\hat{\beta}$ is the predicted value of the response variable (in matrix form), and $\varepsilon = Y - \hat{Y}$.

2.1.7 Root-Mean-Square Deviation

The **Root-Mean-Square deviation** (**RMSD**) or **Root-Mean-Square Error** (**RMSE**) is an oftentimes utilized proportion of the distinctions between values (test or populace values) anticipated by a model or an assessor and the qualities noticed. The RMSD addresses the square base of the second example snapshot of the distinctions between anticipated esteems and noticed values or the quadratic mean of these distinctions. These deviations are called residuals when the computations are performed over the information test that was utilized for assessment and are called mistakes (or forecast blunders) when figured out-of-test. The RMSD effectively totals the extents of the mistakes in forecasts for different data of interest into a solitary proportion of prescient power.

RMSD is a proportion of precision to think about determining blunders of various models for a specific dataset and not between datasets as it is scale-subordinate.

RMSD is generally non-negative, and a worth of 0 (never accomplished practically speaking) would demonstrate an ideal fit to the information. By and large, a lower RMSD is superior to a higher one. Correlations across various kinds of information would be invalid on the grounds that the action is reliant upon the size of the numbers utilized. RMSD is the square base of the normal of squared blunders. The impact of every blunder on RMSD is relative to the size of the squared mistake; along these lines, bigger mistakes significantly affect RMSD. RMSD is delicate to anomalies.

2.2 Feature Selection

In machine learning and statistics, **feature selection**, also known as factor choice, trait determination or variable subset determination, is the method involved with choosing a subset of applicable elements (factors, indicators) for using in model development. Include choice strategies are utilized in light of multiple factors:

- improvement of models to make them simpler to decipher by specialists/clients,
- more limited preparing times,
- to stay away from the scourge of dimensionality,
- work on information's similarity with a learning model class,
- encode innate balances present in the information space.

The focal reason while utilizing a component choice strategy is that the information contains a few elements that are either excess or superfluous, and can hence be eliminated without bringing about much loss of data. Repetitive and unimportant are two unmistakable ideas, since one important element might be excess within the sight of one more applicable component with which it is firmly connected.

Feature selection procedures ought to be recognized from highlight extraction. Include extraction makes new highlights from elements of the first highlights while include determination returns a subset of the elements. Include determination procedures are much of the time utilized in areas where there are many highlights and nearly couple of tests (or

data of interest). Model cases for the use of element choice incorporate the examination of composed texts and DNA microarray information where there are a huge number of highlights and two or three tens to many examples.

2.3 Feature Selection Algorithms

A feature selection algorithm should be visible as the mix of a quest method for proposing new component subsets alongside an assessment measure which scores the different element subsets. The least complex calculation is to test every conceivable subset of highlights observing the one which limits the mistake rate. This is a comprehensive hunt of the space and is computationally obstinate for everything except the littlest of capabilities. The decision of assessment metric intensely impacts the calculation and it is the assessment measurement which recognizes the three primary classes of element determination calculations: coverings, channels and installed strategies.

- Wrapper methods utilize a prescient model to score highlight subsets. Each new subset is utilized to prepare a model, which is tried on a hold-out set. Counting the quantity of missteps make that hold-out set (the mistake pace of the model) gives the score for that subset. As covering techniques train another model for every subset, they are computationally escalated and ordinarily give the best performing highlight set for that specific sort of model or average issue.
- Filter methods utilize an intermediary measure rather than the blunder rate to score a component subset. This action is decided to be quick to register, while catching the helpfulness of the list of capabilities. Normal measures incorporate the common data, the pointwise shared data, Pearson item second relationship coefficient, Relief-based calculations, and bury/intra class distance or the scores of importance tests for each class/highlight mixes. Channels are generally less computationally serious than coverings, and they produce a list of capabilities which isn't tuned to a particular kind of prescient model. This absence of tuning implies a list of capabilities from a channel which is more broad than the set from a covering and is generally giving lower expectation execution than a covering. Anyway the list of capabilities doesn't contain the presumptions of a forecast model, as is more helpful

for uncovering the connections between the elements. Many channels give an element positioning instead of an unequivocal best component subset and the limit in the positioning is picked by means of cross-approval. Channel techniques have additionally been utilized as a preprocessing venture for covering strategies and permitting a covering to be utilized on bigger issues. Another well-known approach is the Recursive Feature Elimination calculation which utilizes generally with Support Vector Machines to repeatedly construct a model and remove features with low weights.

- Embedded methods are a catch-all gathering of strategies which perform feature choice as a component of the model development process. The model of this approach is the LASSO technique for building a straight model which punishes the relapse coefficients with a L1 punishment and contracts a considerable lot of them to nothing. Any elements which have non-zero relapse coefficients are 'chosen' by the LASSO calculation. Enhancements to the LASSO incorporate Bolasso which bootstraps tests; Elastic net regularization which joins the L1 punishment of LASSO with the L2 punishment of edge relapse; and FeaLect which scores every one of the highlights in light of combinatorial examination of relapse coefficients. AEFS further stretches out LASSO to nonlinear situation with auto encoders. These methodologies will quite often be among channels and coverings concerning computational intricacy.

In customary relapse investigation, the most famous type of element determination is stepwise relapse, which is a covering strategy. An eager calculation adds the best component (or erases the most obviously terrible element) at each round. The primary control issue is choosing when they stop the calculation. In AI, this is commonly finished by cross-approval. In measurements, a few rules are streamlined. This prompts the intrinsic issue of settling. More powerful strategies have been investigated, like branch and bound and piecewise straight organization.

Subset Selection: Subset selection assesses a subset of highlights collectively for appropriateness. Subset determination calculations can be separated into coverings, channels, and inserted strategies. Coverings utilize a pursuit calculation to look through the space of potential highlights and assess every subset by running a model on the subset.

Coverings can be computationally cost and have a gamble of over fitting to the model. Channels are like coverings in the hunt approach, however rather than considering in contrast to a model, a more straightforward channel is assessed. Installed procedures are implanted in and intended for a model.

Numerous famous inquiry approaches utilize covetous slope climbing which iteratively assesses a competitor subset of highlights then alters the subset and assesses on the off chance that the new subset is an improvement over the old. Assessment of the subsets requires a scoring metric that grades a subset of highlights. Through inquiry is by and large unreasonable, so some practitioner (or administrator) characterizes place to pause and the subset of highlights with the most elevated score found up to that point is chosen as the palatable component subset. The halting rule differs by calculation; potential measures include a subset score surpasses an edge and a program's most extreme permitted run time has been outperformed, and so forth.

Elective hunt put together methods are based with respect to designated projection pursuit which finds low-layered projections of the information that score profoundly: the highlights that have the biggest projections in the lower-layered space are then chosen.

Search approaches include:

- Exhaustive
- Best first
- Simulated annealing
- Genetic algorithm
- Greedy forward selection
- Greedy backward elimination
- Particle swarm optimization
- Targeted projection pursuit
- Scatter search
- Variable neighborhood search

Two famous channel measurements for order issues are connection and shared data, albeit nor are valid measurements or 'distance measures' in the numerical sense, since they neglect to submit to the triangle disparity and hence figure no genuine 'distance' - they

ought to rather be viewed as 'scores'. These scores are figured between an up-and-comer component (or set of elements) and the ideal result class. There are notwithstanding genuine measurements that are a basic capacity of the shared data.

Other available filter metrics include:

- Class separability
  - Error probability
  - Inter-class distance
  - Probabilistic distance
  - Entropy
- Consistency-based feature selection
- Correlation-based feature selection

2.4 The Connection of Feature Selection with Chi-Square

Chi-Square is to be utilized when the component is all out, the objective variable is some way can be thought as clear cut. It estimates the level of relationship between two straight out factors. In the event that both are numeric, we can utilize Pearson's item second connection, and assuming the property is mathematical. There are two classes that we can utilize a t-test in the event that beyond what two classes we can utilize ANOVA.

2.5 Chi-Square Test for Feature Selection

The Chi-Square test is helpful in AI and how this test has an effect. Highlight determination is a significant issue in AI where we will have a few elements in line and need to choose the best highlights to fabricate the model. The chi-square test assists you with tackling the issue which include determination by testing the connection between the highlights.

2.5.1 Chi-Square Distribution

A random variable follows chi-square dispersion in the event that it very well may be composed very well as an amount of squared standard ordinary factors.

$$X^2 = \sum Z_i^2 \quad (2.23)$$

Z_1, Z_2.. are standard normal variables.

2.5.2 Degrees of Freedom

Degrees of freedom allude to the greatest number of intelligently autonomous qualities which have the opportunity to differ. In basic words, it tends to be characterized as the absolute number of perceptions less the quantity of autonomous imperatives forced on the perceptions.

$$Df = N - 1 \quad (2.24)$$

Df = degrees of freedom

N= sample size

The levels of opportunity increment Chi-Square dissemination approximates to typical dispersion.

2.5.3 Chi-Square Test

A chi-square test is utilized in insights to test the autonomy of two occasions. Because of the information of two factors, we can get noticed count O and anticipated count E. Chi-Square estimates how expected count E and noticed count O strays one another.

The formula for Chi-Square id

$$X^2 = \sum (O - E)^2 / E \quad (2.25)$$

Where,

O = Observed frequency

E = Expected frequency

$\sum$ = Summation

χ^2 = Chi-Squared value

It considers a situation where we really want to decide the connection between the free classification include (indicator) and ward class feature (response). In highlight choice, we plan to choose the elements which are exceptionally reliant upon the reaction.

At the point, when two highlights are free, the noticed count is near the normal count, hence, we will have more modest Chi-Square worth. So high Chi-Square worth demonstrates that the theory of freedom is inaccurate. In straightforward words, higher the Chi-Square worth the component is more reliant upon the reaction and it very well may be chosen very well for model preparation.

2.5.4 Steps for Chi-Square Test with an example

Consider a data-set where have to determine why customers are leaving the bank. It performs a Chi-Square test for two variables. ***Gender*** of a customer with values as Male/Female as the predictor and ***Exited*** describes whether a customer is leaving the bank with values Yes/No as the response. In this test, we will check *"is there any relationship between Gender and Exited"*.

Steps to perform the Chi-Square Test:

I. Define Hypothesis.
II. Build a Contingency table.
III. Find the expected values.
IV. Calculate the Chi-Square statistic.
V. Accept or Reject the Null Hypothesis.

2.5.4.1 Define Hypothesis

Null Hypothesis (H0): Two variables are independent.

Alternate Hypothesis (H1): Two variables are not independent.

2.5.4.2 Contingency table

The table shows the distribution of one variable in rows and another in columns. It uses to study the relation between two variables.

Exited\ Gender	**Yes**	**No**	**Total**
Male	38	178	216

Female	44	140	184
Total	82	318	400

Table 2.3 Contingency table for observed values

Degrees of freedom for contingency table is given as (r-1) * (c-1) where r,c are rows and columns. Here df = (2–1) * (2–1) = 1.

In the above table we have figured out all observed values. Next steps are to find expected values and, get the Chi-Square value and check for relationship.

2.5.4.3 Find the Expected Value

Based on the null hypothesis, the two variables are independent. If A and B are two independent events

$$P(A \cap B) = P(A) * P(B) \qquad (2.26)$$

Calculation of the expected value for the first cell is those who are Males and are Exited from the bank.

E1 = n * p

p = p(Yes) * p(Male)

p = (82/400) * (216/400)

p = 0.1107

now, E1 = 400 * 0.1107 = 4

In similar, we calculate E2, E3, E4 and get the following results.

Exited\ Gender	**Yes**	**No**
Male	44	172
Female	38	146

Table 2.4 Contingency table for expected values

2.5.4.4 Calculate Chi-Square value

The following table summarizes the observed values and calculates expected values into a table and determines the Chi-Square value.

Gender, Exited	O	E	O-E	Square of O-E	(Square of O-E)/E
Male, Yes	38	44	-6	36	0.818181818
Male, No	178	172	6	36	0.209302326
Female, Yes	44	38	6	36	0.947368421
Female, No	140	146	-6	36	0.246575342
Chi Square Value					2.221427907

Table 2.5 Calculation of Chi Square Value

In the above table,

O – Observed values

E – Expected values

Chi-Square is calculated as 2.22 by using the Chi-Square statistic formula.

2.5.4.5 Accept or Reject the Null Hypothesis

With 95% certainty that is alpha = 0.05, we will check the determined Chi-Square worth which falls in the acknowledgment or dismissal area. Having levels of opportunity =1 is (calculated with possibility table) and alpha =0.05. The Chi-Square worth is 3.84. The Chi-Square qualities are still up in the air with the Chi-Square table. The chi-square dissemination is the right side since the distinction in Observed and Expected is enormous.

Chi-Square ranges from 0 to boundlessness and alpha reaches from 0 to 1 the other way. We will dismiss the Null speculation if Chi-Square worth falls in the blunder locale (alpha from 0 to 0.05). Since the Chi-Square worth is not exactly the basic Chi-Square worth, the null hypothesis is accepted. Because the two factors are free, Gender variable can't be chosen for preparing the model.

CHAPTER 3
PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE USING MULTIPLE LINEAR REGRESSION

Nowadays, the education systems have been changed from teacher center approach to learner centered approach. Thus, students' related features need to analyze by predicting students' academic performance to give the active learning in educational system. In this system, Multiple Linear Regression is applied to predict the students' academic performance on the UCI's student performance data set. The student's academic performance prediction model is built by using Multiple Linear Regression method with feature selection (Chi-Square). The purpose of this system is to predict the students' final grade based on previous grades and relevant features.

3.1. The Proposed System

Education has been offered in classes where students can interact directly with their teachers and making students' related features presence very important. The key role of the teacher in the teaching-learning process, it is need to prepare to identify the diversity of students' nature. The education system tries to highlight the role of teachers in identifying the students' performance to become from passive learning to active learning. In this system, feature selection process is applied to choose the important features of the student.

The aim of the system is to build the prediction model of students' academic performance by using multiple linear regression (MLR). This study examines the role of satisfaction on students' academic performance and investigates the relationship between satisfaction of students and academic performance and explores other factors that contribute academic performance using Multiple Linear Regression Method.

3.2. The System Flow

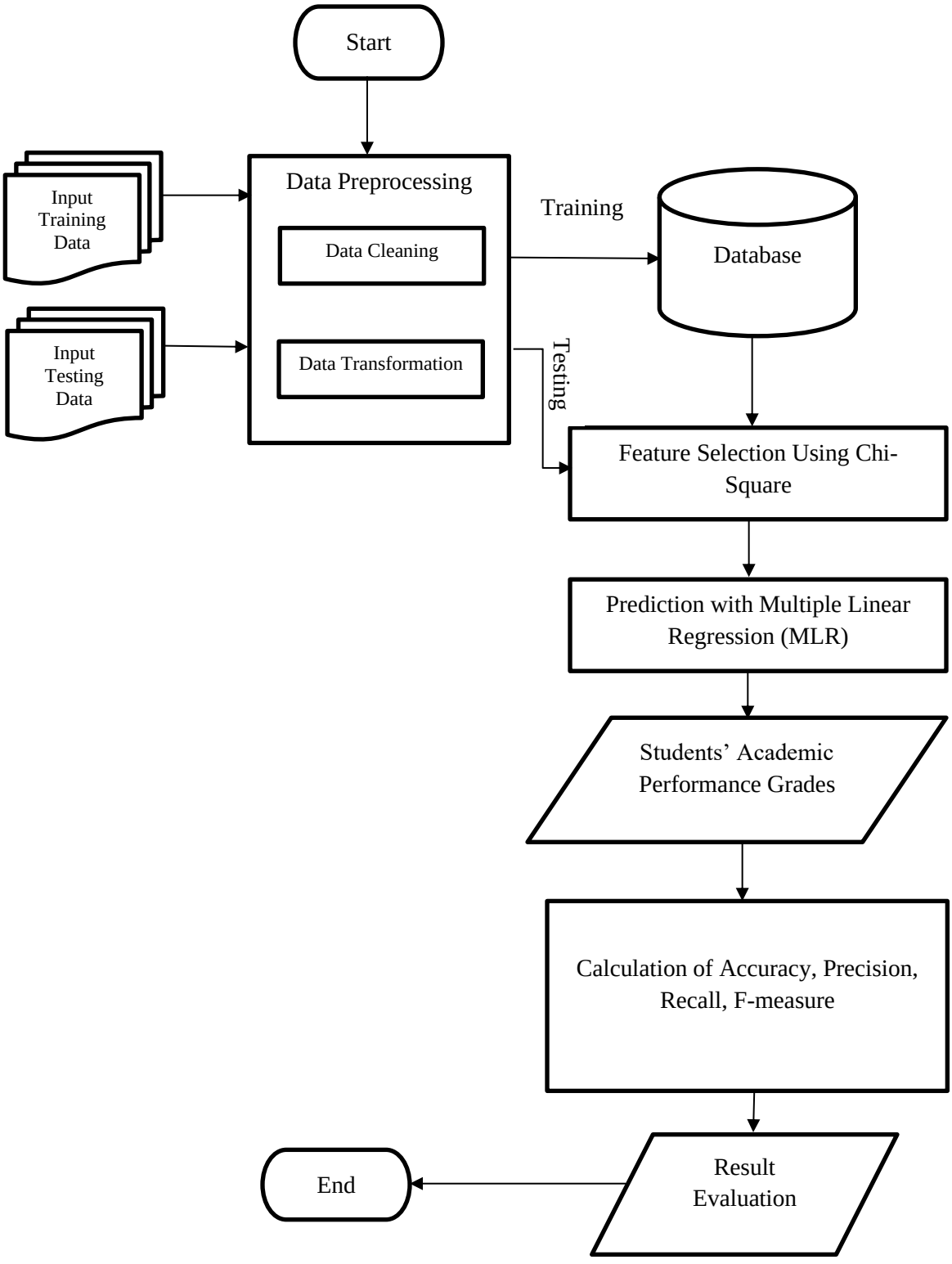


Figure 3.1: The System Flowchart

The dataset collected from UCI consists of achievement of student in the secondary school of education. The dataset consists of 649 student's records with 33 features. The features are students' grades, demographic features, social features. After Data collection process, some preprocessing techniques are applied on dataset. Data cleaning is used to solve irrelevant missing part such as ignore the tuple or fill the missing values, inconsistent data and to deal with incomplete values. Data transformation is used to transform the raw data in a useful and efficient format so that it can be easily accepted and used by multiple linear regression method. After preprocessing, data are stored in students' training database. This data are used MLR method to predict the student academic performance.

And then, results are evaluated by accuracy calculation and then show the result of prediction. The results show the students' grades and which features are related to improve students' academic performance.

3.3. Attributes Values of System Dataset

The following table explains the detailed explanation of attributes values of System Dataset. It contains 33 features. The features are student's grades (G1,G2,G3), demographic features (school, sex, age, address, etc.)and social features (internet, romantic, freetime, gout, etc.)

Table 3.1 Detail explanation of attributes in dataset

ID	Attribute	Explanation
1	school	UCI student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)L
2	sex	student's sex (binary: "F" - female or "M" - male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: "U" - urban or "R" - rural)
5	famsize	family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6	Pstatus	parent's cohabitation status (binary: "T" - living together or "A" - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 “ 5th to 9th grade, 3 “ secondary education or 4 â€“ higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 “ 5th to 9th grade, 3 “ secondary education or 4 “ higher education)
9	Mjob	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10	Fjob	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11	reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12	guardian	student's guardian (nominal: "mother", "father" or "other")
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)

ID	Attribute	Explanation
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)

3.4. Case Study (Calculation on 30 Data Samples)

The following table describes the dataset of the system especially used G1 and G2 as independent variables and G3 as a predictor with 30 students' data sample.

Table 3.2 30 data samples of students using in manual calculation

No	higher	Internet	famrel	health	absences	G1	G2	G3
1	yes	No	4	3	4	0	11	11
2	yes	Yes	5	3	2	9	11	11
3	yes	Yes	4	3	6	12	13	12
4	yes	Yes	3	5	0	14	14	14
5	yes	No	4	5	0	11	13	13
6	yes	Yes	5	5	6	12	12	13
7	yes	Yes	4	3	0	13	12	13
8	yes	No	4	1	2	10	13	13
9	yes	Yes	4	1	0	15	16	17
10	yes	Yes	5	5	0	12	12	13
11	yes	Yes	3	2	2	14	14	14
12	yes	Yes	5	4	0	10	12	13
13	yes	Yes	4	5	0	12	13	12
14	yes	Yes	5	3	0	12	12	13
15	yes	Yes	4	3	0	14	14	15
16	yes	Yes	4	2	6	17	17	17

No	higher	Internet	famrel	health	absences	G1	G2	G3
17	yes	Yes	3	2	10	13	13	14
18	yes	No	5	4	2	13	14	14
19	yes	Yes	5	5	2	8	8	7
20	yes	Yes	3	5	6	12	12	12
21	yes	Yes	4	1	0	12	13	14
22	yes	yes	5	5	0	11	12	12
23	yes	yes	4	5	0	12	13	14
24	yes	yes	5	5	2	10	10	10
25	yes	yes	4	5	2	10	11	10
26	yes	yes	1	5	6	10	11	12
27	yes	yes	4	5	8	11	12	12
28	yes	yes	2	1	0	11	11	11
29	yes	yes	5	5	2	12	12	13
30	yes	yes	4	5	4	12	11	12

Calculation (UCI dataset - G1, G2, G3)

Samples - 30

Variables - 3 (one dependent variable- G3 and 2 independent variables – G1 and G2)

n = 30

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon} \quad (3.1)$$

$$\sum Y = \beta_0 n + \beta_1 \sum X_1 + \beta_2 \sum X_2 \quad \text{---------- Eq 1} \quad (3.2)$$

$$\sum X_1 Y = \beta_0 \sum X_1 + \beta_1 \sum X_1^2 + \beta_2 \sum X_1 X_2 \quad \text{---------- Eq 2} \quad (3.3)$$

$$\sum X_2 Y = \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum X_2^2 \quad \text{---------- Eq 3} \quad (3.4)$$

Table 3.3 Calculation of 2 variables (require for Eq 1, 2 and 3)

No	G3(Y)	G1(X_1)	G2(X_2)	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
1	11	0	11	0	121	0	0	121
2	11	9	11	99	121	99	81	121
3	12	12	13	144	156	156	144	169

No	G3(Y)	G1(X_1)	G2(X_2)	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
4	14	14	14	196	196	196	196	196
5	13	11	13	143	169	143	121	169
6	13	12	12	156	156	144	144	144
7	13	13	12	169	156	156	169	144
8	13	10	13	130	169	130	100	169
9	17	15	16	255	272	240	225	256
10	13	12	12	156	156	144	144	144
11	14	14	14	196	196	196	196	196
12	13	10	12	130	156	120	100	144
13	12	12	13	144	156	156	144	169
14	13	12	12	156	156	144	144	144
15	15	14	14	210	210	196	196	196
16	17	17	17	289	289	289	289	289
17	14	13	13	182	182	169	169	169
18	14	13	14	182	196	182	169	196
19	7	8	8	56	56	64	64	64
20	12	12	12	144	144	144	144	144
21	14	12	13	168	182	156	144	169
22	12	11	12	132	144	132	121	144
23	14	12	13	168	182	156	144	169
24	10	10	10	100	100	100	100	100
25	10	10	11	100	110	110	100	121
26	12	10	11	120	132	110	100	121
27	12	11	12	132	144	132	121	144
28	11	11	11	121	121	121	121	121
29	13	12	12	156	156	144	144	144
30	12	12	11	144	132	132	144	121
Total	**381**	**344**	**372**	**4478**	**4816**	**4361**	**4178**	**4698**

$381 = 30\ \beta_0 + 344\ \beta_1 + 372\ \beta_2$ Eq 1

$4478 = 344\ \beta_0 + 4178\ \beta_1 + 4361\ \beta_2$ Eq 2

$4816 = 372\ \beta_0 + 4361\ \beta_1 + 4698\ \beta_2$ Eq 3

***Coefficients** >>>*

a = - 0.50480, b = 0.052380, c = 1.016465

Multiple Linear Regression Model is:

$Y = - 0.50480 + 0.052380\ \beta_1 + 1.016465\ \beta_2$

3.5 Data Preparation for Processing

Assuming all of the assumptions for a multiple linear regression have been met, this can be done by generalizing to unseen data. Any model whose goal is prediction, this is best determined by splitting the data into two pieces:

- Training data (~75% of the data)
- Testing data (~25% of the data)

The test data should be set aside and does not look at until ready to determine how well proposed regression model is generalizing. The train data is used to fit proposed model.

CHAPTER 4
IMPLEMENTATION OF SYSTEM

The primary purpose of assessment and evaluation is to improve student learning. Information gathered through assessment and evaluations helps teachers to identify students' difficulties as well as to detect weaknesses in programs. Assessment and evaluations are important tools for adapting curriculum and instructional approaches to students' needs and for determining the overall effectiveness of programs and classroom practices. Assessment is the process of gathering information from a variety of sources (including assignments, projects, and a midterm) that accurately reflect how well students are achieving the curriculum expectations. As part of assessment, teachers provide students with descriptive feedback that guides their effort towards improvement.

The breakdown of the final mark of any course will be as follows: 70% of the grade is based on evaluations conducted throughout the course with special consideration given to more recent and more consistent evidence of achievement and 30% of the grade is based on a final evaluation in a form that is suitable to the course content.

Predicting students' academic performance has long been an important area of research in education. Most existing literature have made by using of traditional statistical methods that run into the problems of over fitted models, inability to effectively handle large numbers of participants and predictors and inability to pick out non-linearity that may be present. Regression-based ML methods that can produce highly interpretable yet accurate model for new predictions are able to provide some solutions to the aforementioned problems.

4.1 Implementation of the System

This academic student performance evaluation is based on 32 variables and the out will be predict by multiple linear regression (MLR). To get more related data value, Feature selection is also concerned. This system is developed by C#.Net Language on Microsoft Visual Studio 2015 and Microsoft SQL server 2017 Expression version as database engine.

Figure 4.1 The System Login Page

This system only permits to use the registered user only. So, the user must be login first and authentication process will be made. After the authentication phase is completed, the main page of the system can be seen as shown in figure 4.2.

The main page of the system has seven menus:

“Load Student Data” menu: Load student data menu is supported to load the training data to the system. By this menu, the user can load and submit various size of data set for different prediction.

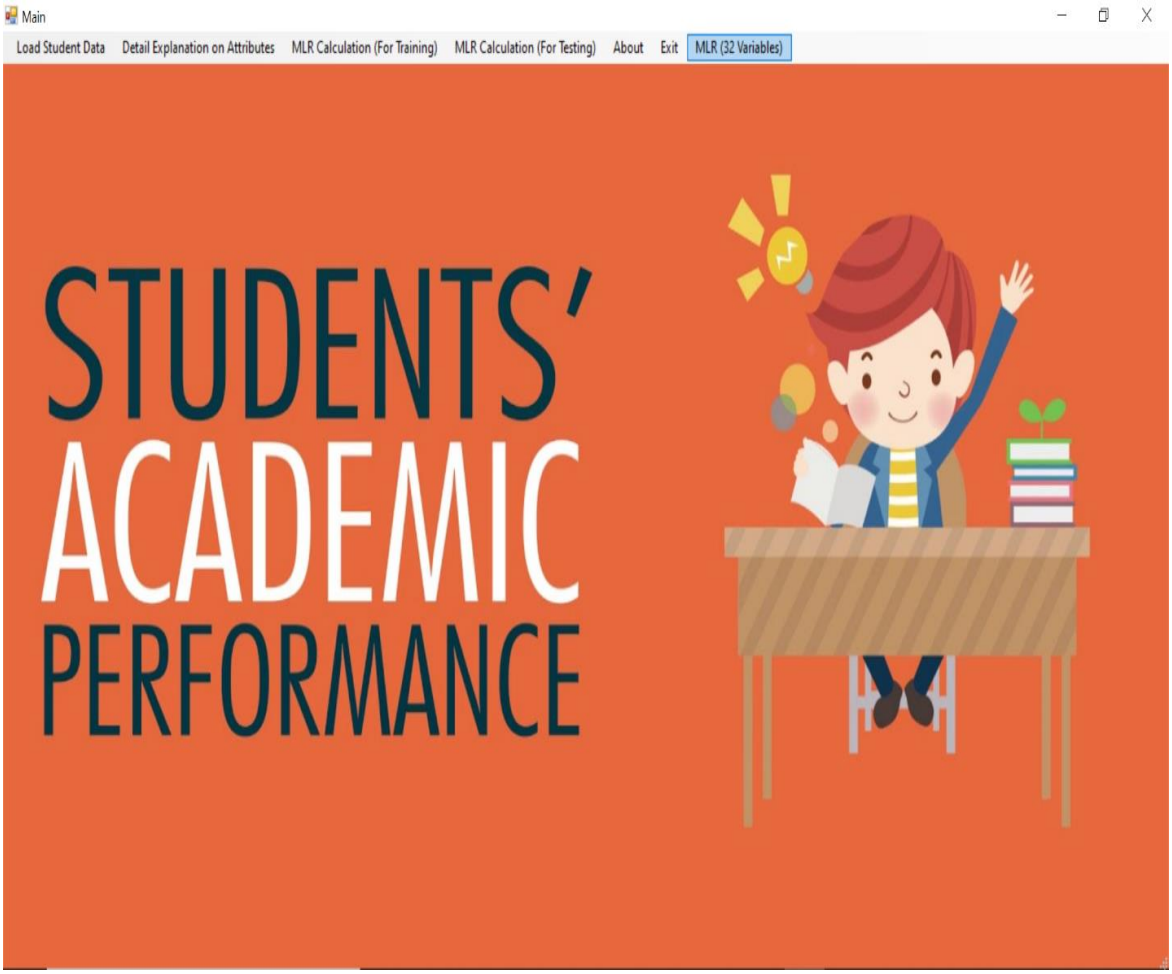


Figure 4.2 Main Page of System

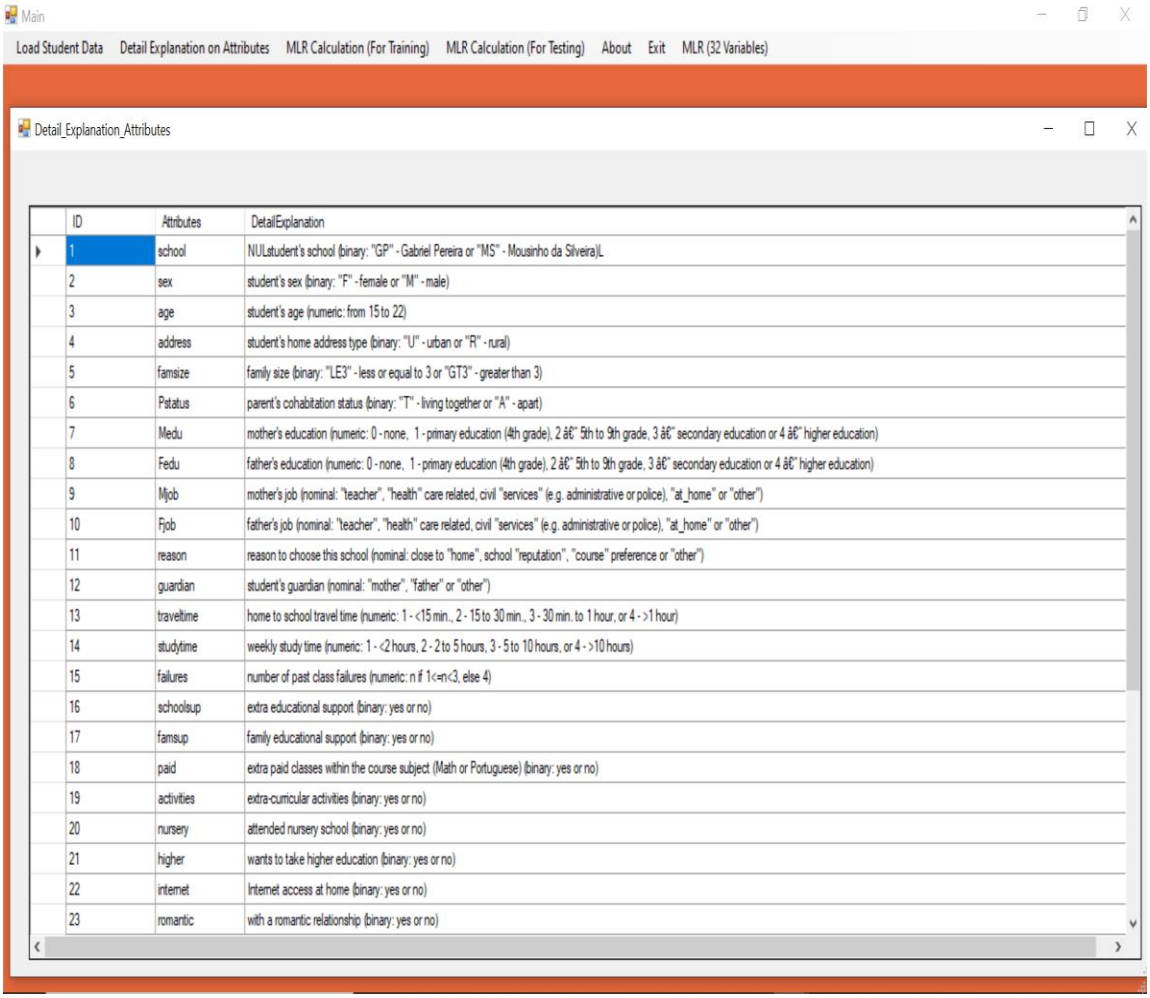
"Data Explanation on Attribute" menu: is shown and explain the 33 variables of the academic student information on each record. Data explanation on each attribute's detail explanations are shown in figure 4.3 (System Design of Attribute Explanation). Detail and complete explanation can also be learned in Table 4.1.

Table 4.1: Detail Explanation of All Attributes

ID	Attributes	Detail Explanation
1	school	NULstudent's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)L
2	sex	student's sex (binary: "F" - female or "M" - male)
3	age	student's age (numeric: from 15 to 22)

ID	Attributes	Detail Explanation
4	address	student's home address type (binary: "U" - urban or "R" - rural)
5	famsize	family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6	Pstatus	parent's cohabitation status (binary: "T" - living together or "A" - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 " 5th to 9th grade, 3 " secondary education or 4 " higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 " 5th to 9th grade, 3 " secondary education or 4 " higher education)
9	Mjob	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10	Fjob	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11	reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12	guardian	student's guardian (nominal: "mother", "father" or "other")
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)

ID	Attributes	Detail Explanation
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20, output target)



ID	Attributes	DetailExplanation
1	school	NULstudent's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)L
2	sex	student's sex (binary: "F" - female or "M" - male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: "U" - urban or "R" - rural)
5	famsize	family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6	Pstatus	parent's cohabitation status (binary: "T" - living together or "A" - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)
9	Mjob	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10	Fjob	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11	reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12	guardian	student's guardian (nominal: "mother", "father" or "other")
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)

Figure 4.3 System Design of Detail Attribute Explanation

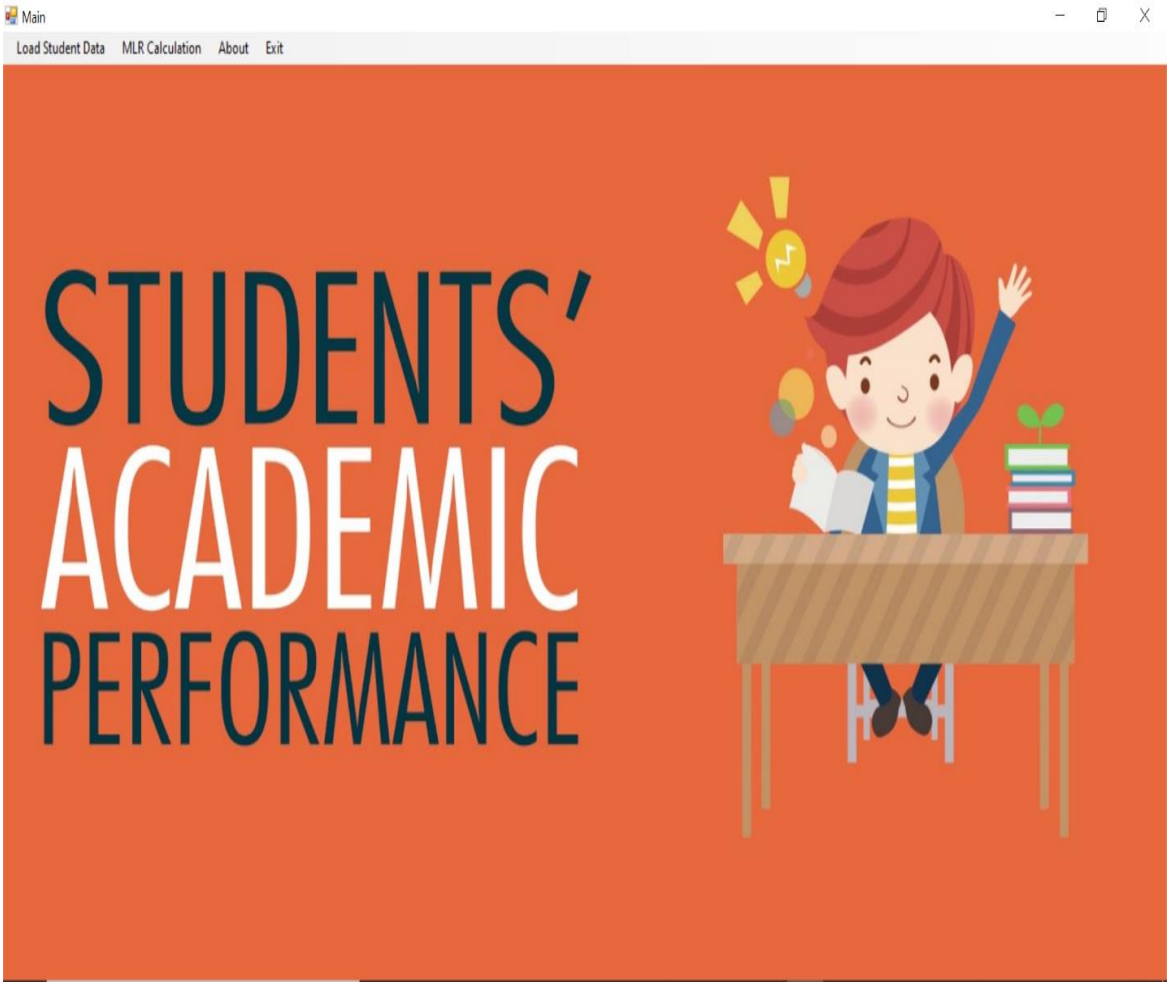


Figure 4.4 "MLR (32 variables)" Menu's Content Page

The Design and implementation of "MLR (32 variables)" menu's content page is as shown in figure 4.4. This page is implemented to calculate the multi linear regression on student information (based on 32 variables) without concerning the feature selection consideration. In this calculation phase, the training dataset is loaded to the system database by clicking "Load Student Data" tool strip menu button and the data loading process is as shown in figure 4.5. So, this system supports the open dialog box to choose the desire data set. This system can be trained various size dataset of student data for better performance and evaluation results. The loaded training data is as shown in figure 4.6.

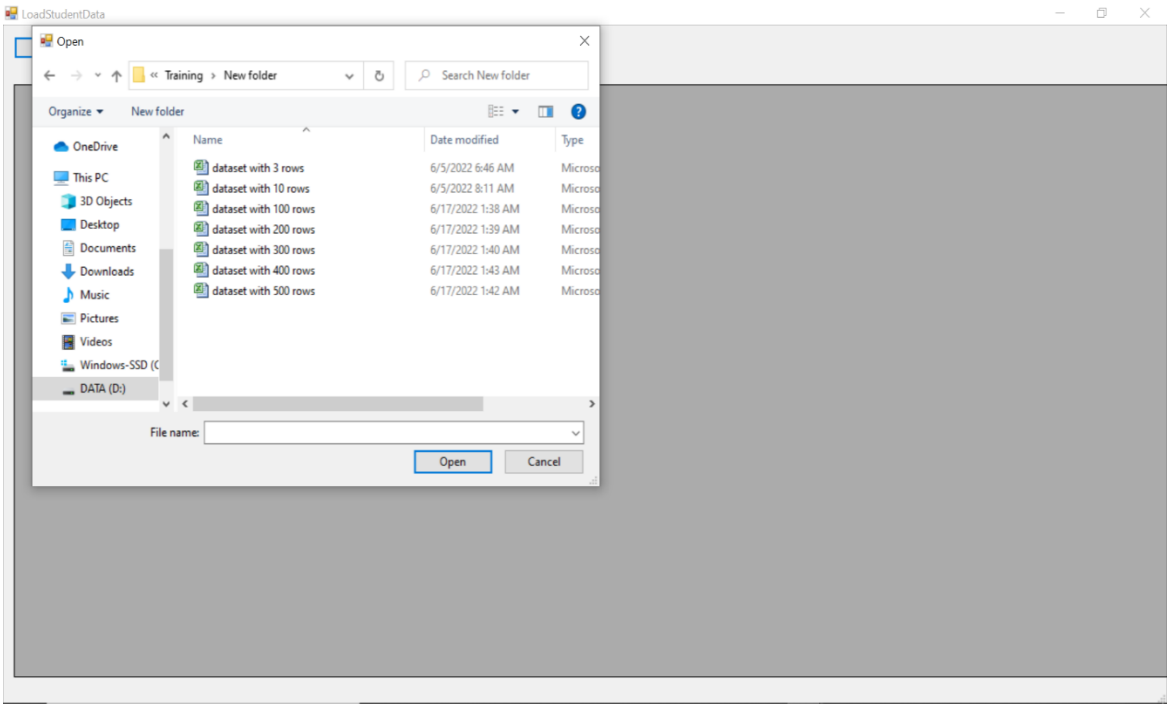
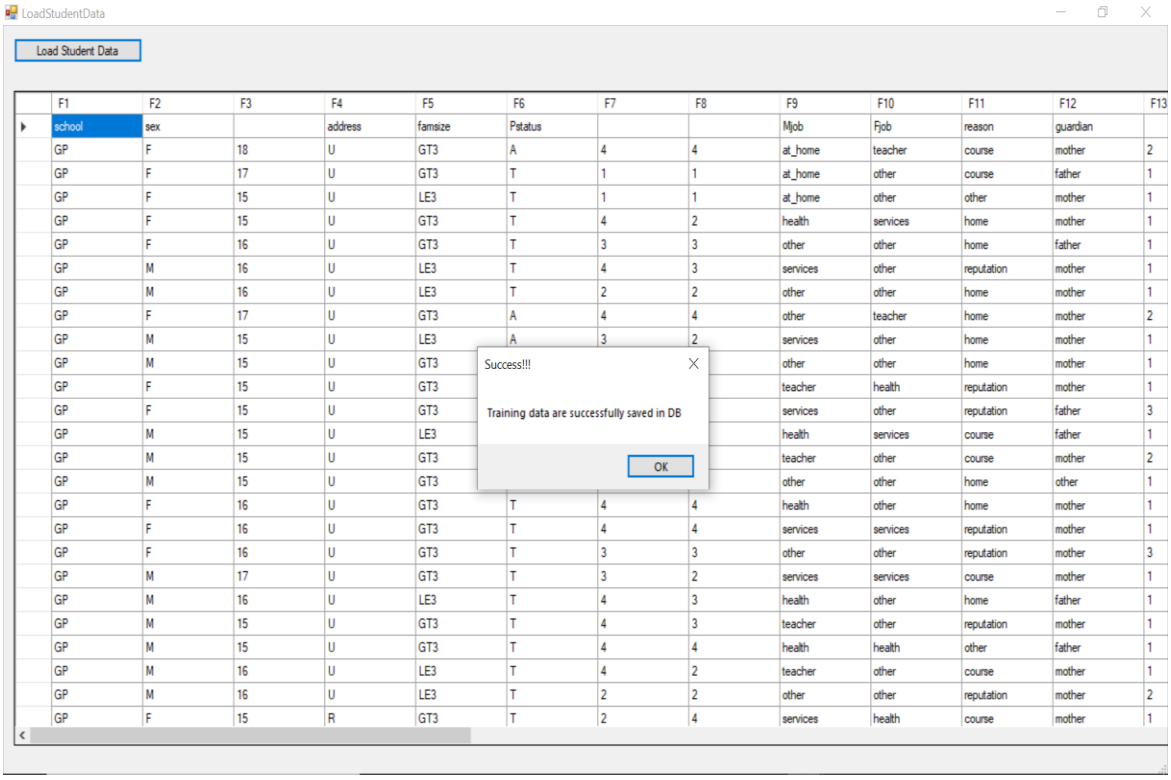


Figure 4.5 Choose and Load Dataset



F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
school	sex		address	famsize	Pstatus			Mjob	Fjob	reason	guardian	
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1
GP	F	16	U	GT3	T	3	3	other	other	home	father	1
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1
GP	M	15	U	GT3				other	other	home	mother	1
GP	F	15	U	GT3				teacher	health	reputation	mother	1
GP	F	15	U	GT3				services	other	reputation	father	3
GP	M	15	U	LE3				health	services	course	father	1
GP	M	15	U	GT3				teacher	other	course	mother	2
GP	M	15	U	GT3				other	other	home	other	1
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1
GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1
GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3
GP	M	17	U	GT3	T	3	2	services	services	course	mother	1
GP	M	16	U	LE3	T	4	3	health	other	home	father	1
GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother	1
GP	M	15	U	GT3	T	4	4	health	health	other	father	1
GP	M	16	U	LE3	T	4	2	teacher	other	course	mother	1
GP	M	16	U	LE3	T	2	2	other	other	reputation	mother	2
GP	F	15	R	GT3	T	2	4	services	health	course	mother	1

Figure 4.6 The Loaded Training Dataset

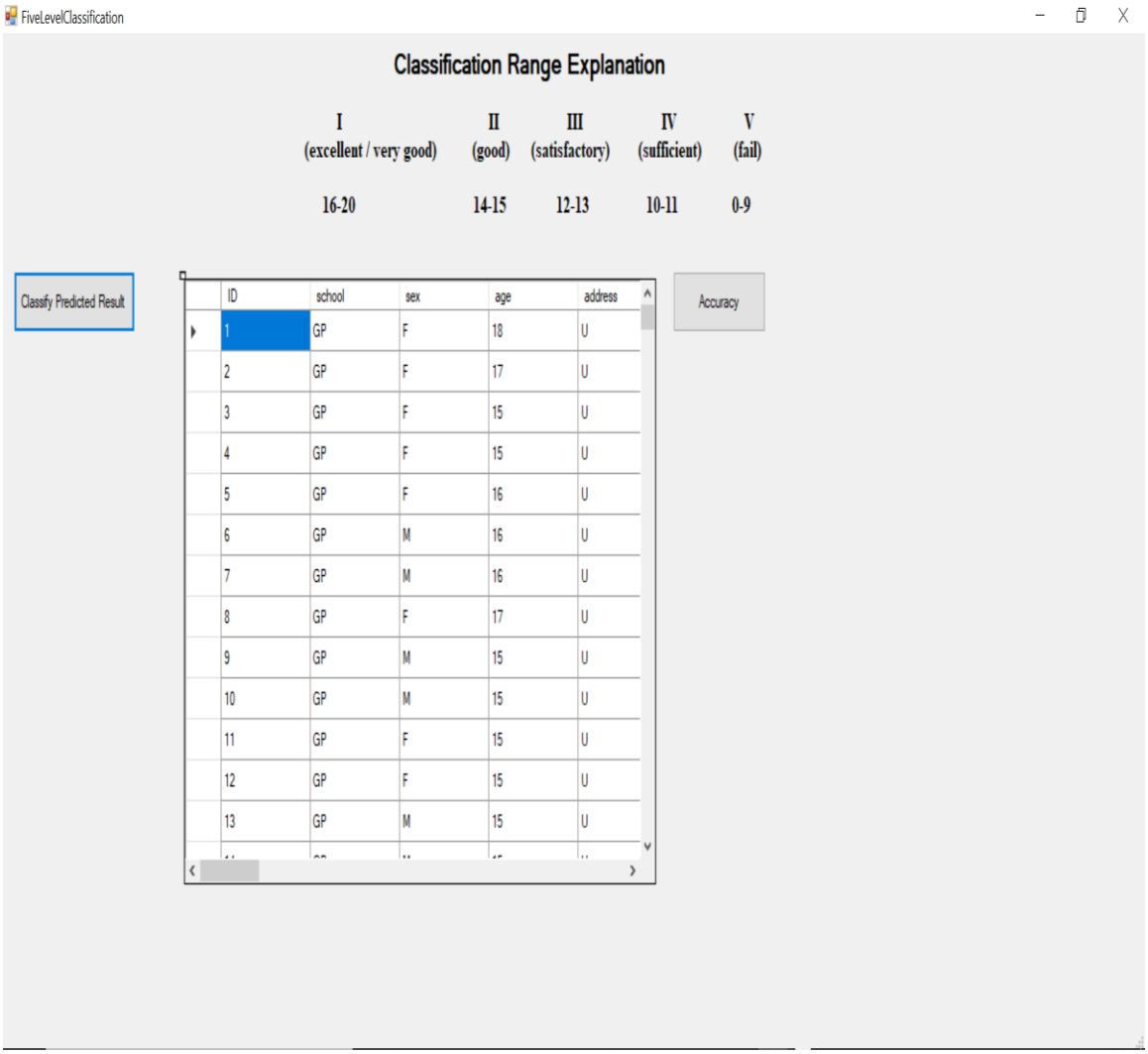


Figure 4.7 Five Level Classification (MLR based on 32 variables)

The theory background and detail processing steps are already discussed in previous chapters and then the system interface design of five level classifications which is based on 32 variables are as shown in figure 4.7. The accuracy value of the MLR prediction without concerning feature selection which is based on 32 variables is as shown in figure 4.8.

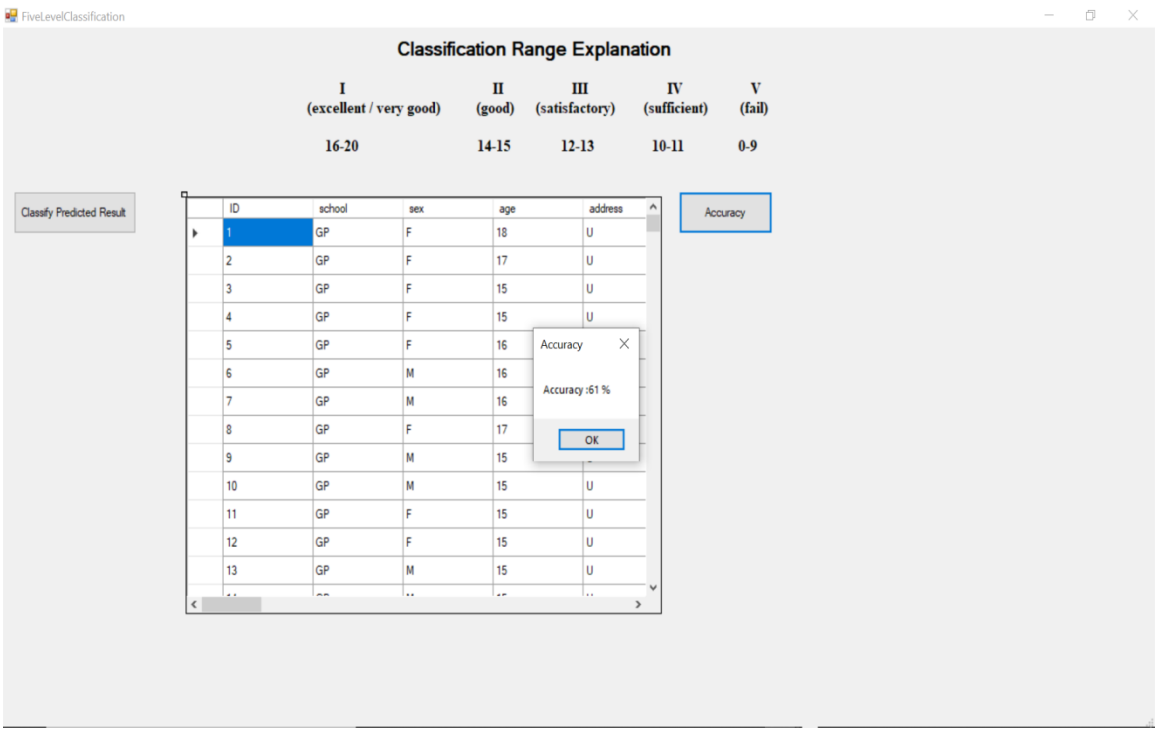
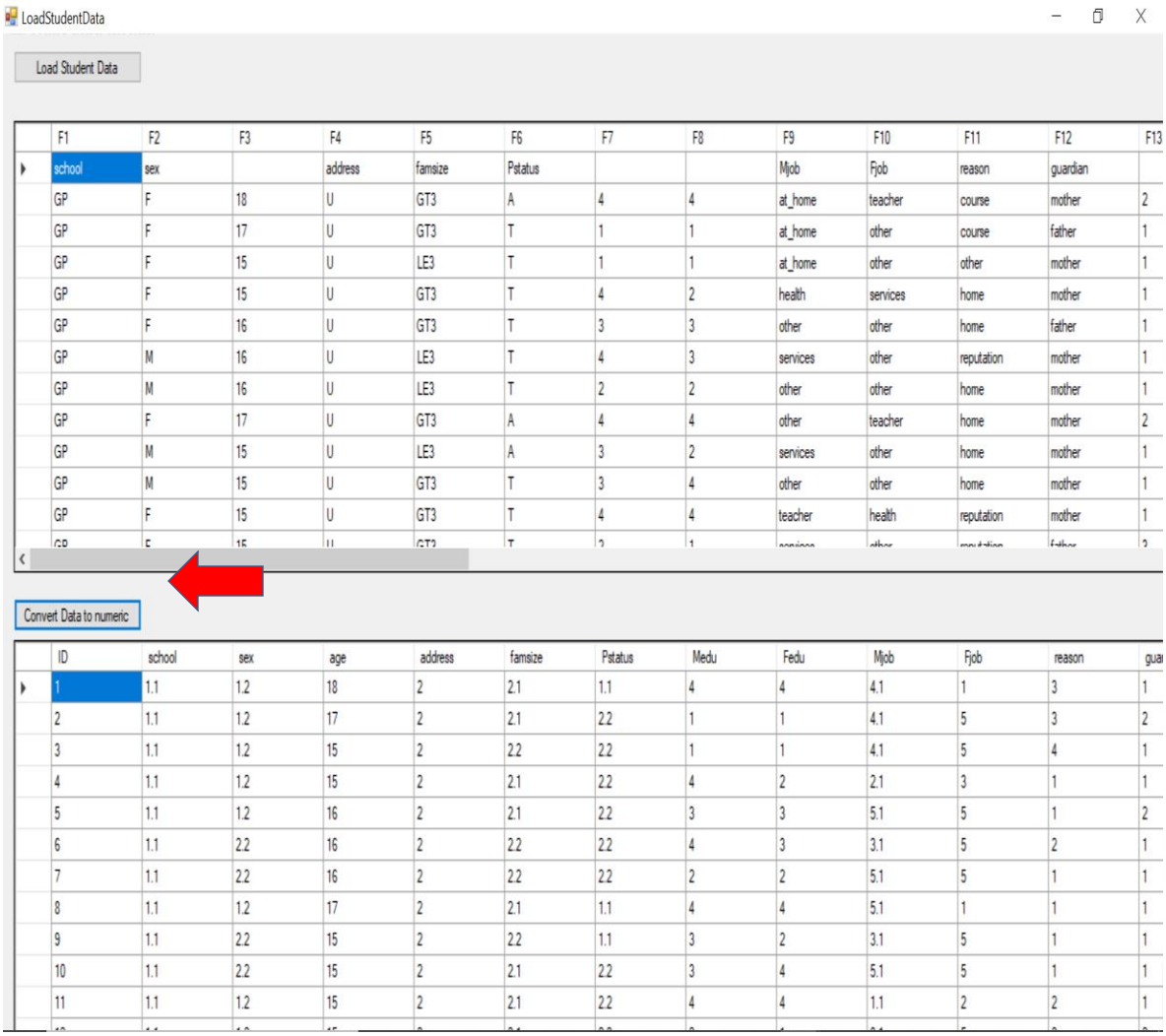


Figure 4.8 Accuracy (MRL Prediction without Feature Selections)

4.2 Academic Student Performance Evaluation with Feature Selection

The central premise when using a feature selection technique is that the data contains some features that are either *redundant* or *irrelevant*, and can thus be removed without incurring much loss of information.[9] *Redundant* and *irrelevant* are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Feature selection techniques should be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).



F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
school	sex		address	famsize	Pstatus			Mjob	Fjob	reason	guardian	
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1
GP	F	16	U	GT3	T	3	3	other	other	home	father	1
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1

ID	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	gua
1	1.1	1.2	18	2	2.1	1.1	4	4	4.1	1	3	1
2	1.1	1.2	17	2	2.1	2.2	1	1	4.1	5	3	2
3	1.1	1.2	15	2	2.2	2.2	1	1	4.1	5	4	1
4	1.1	1.2	15	2	2.1	2.2	4	2	2.1	3	1	1
5	1.1	1.2	16	2	2.1	2.2	3	3	5.1	5	1	2
6	1.1	2.2	16	2	2.2	2.2	4	3	3.1	5	2	1
7	1.1	2.2	16	2	2.2	2.2	2	2	5.1	5	1	1
8	1.1	1.2	17	2	2.1	1.1	4	4	5.1	1	1	1
9	1.1	2.2	15	2	2.2	1.1	3	2	3.1	5	1	1
10	1.1	2.2	15	2	2.1	2.2	3	4	5.1	5	1	1
11	1.1	1.2	15	2	2.1	2.2	4	4	1.1	2	2	1

Figure 4.9 Training Data loading and Numerical Conversion

In the training data loading phase of Academic Student Performance Evaluation with Feature Selection, the loaded data are converted to numerical values to calculate each data column as shown in figure 4.9.

After the feature selection determination on 32 variables is complete, the system will generate message alert for the completion of feature selection in figure 4.10.

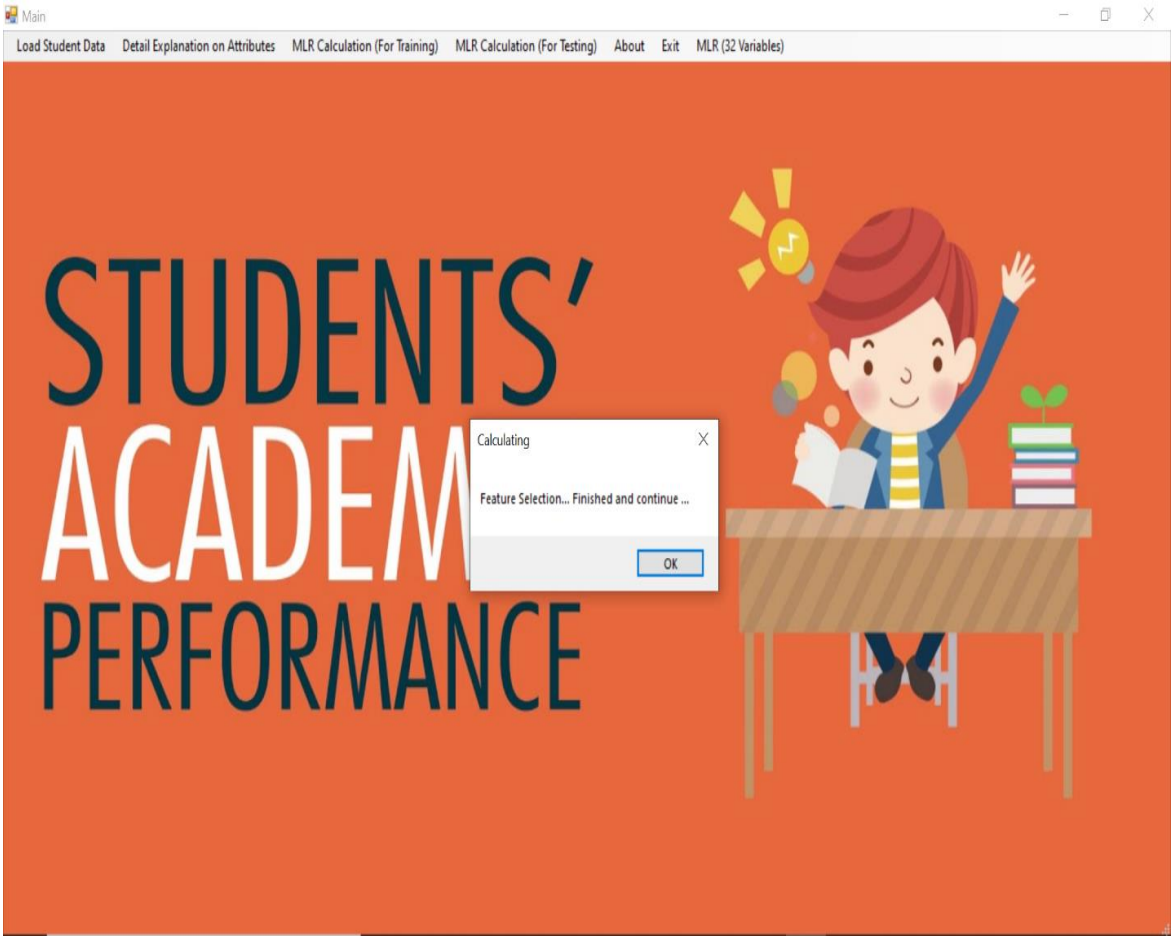


Figure 4.10 Message Alert for Feature Selection Completion

After the feature selection phase, the system will be proceeded the performance evaluation on the feature selected variables and the evaluation results are shown in grid view of the figure 4.10. And then the accuracy on prediction for the feature selected variables is also described with message alert. By visual evaluation on the accuracy values described in figure 4.8 (Accuracy 61%) and figure 1.1 (Accuracy 74.4%), the MLR on feature selection based prediction is more favorable than with concerning process of feature selection.

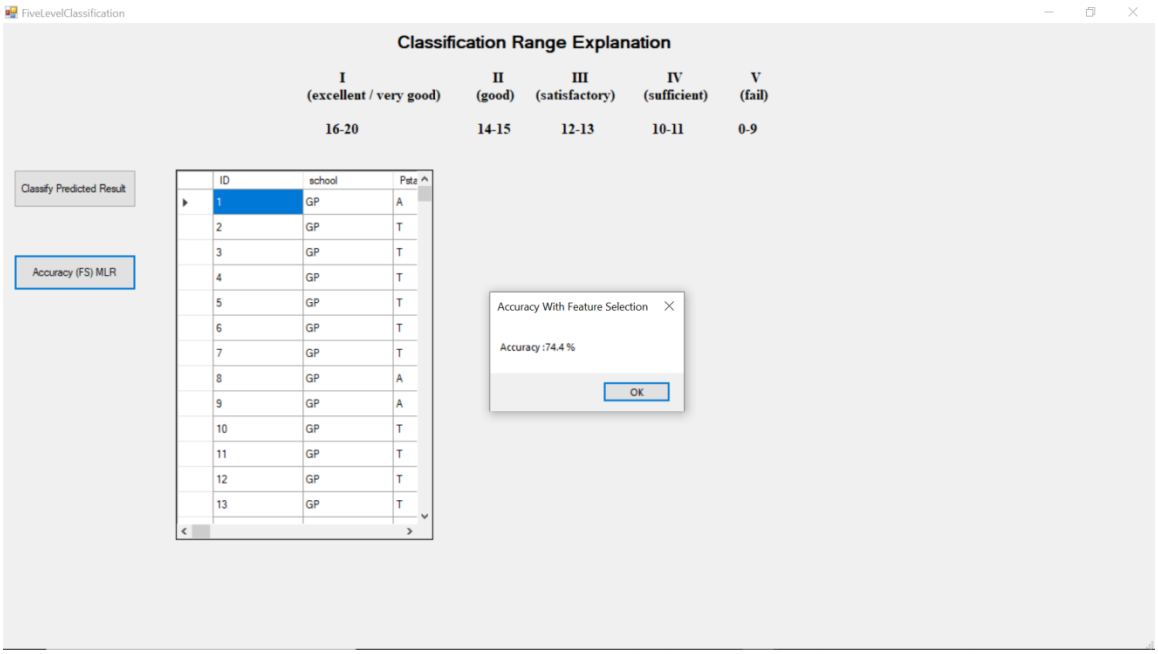


Figure 4.11 Accuracy Evaluation (MRL with Feature Selection)

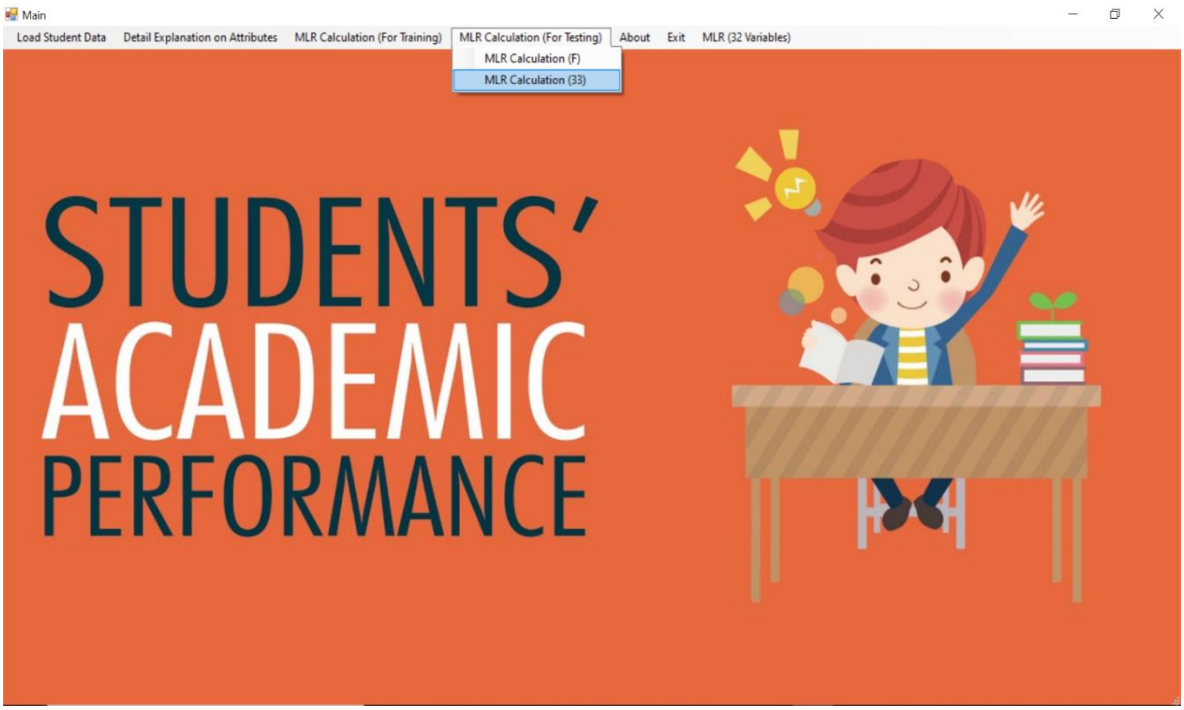


Figure 4.12 Menu Supported For Testing

In this page, the menu is supported for MRL calculation with feature selection and without feature selection sub menus. The evaluation process of these two different

calculations is analyzed on testing data sample for 200 records. The respective comparable accuracy values can be study in following figures.

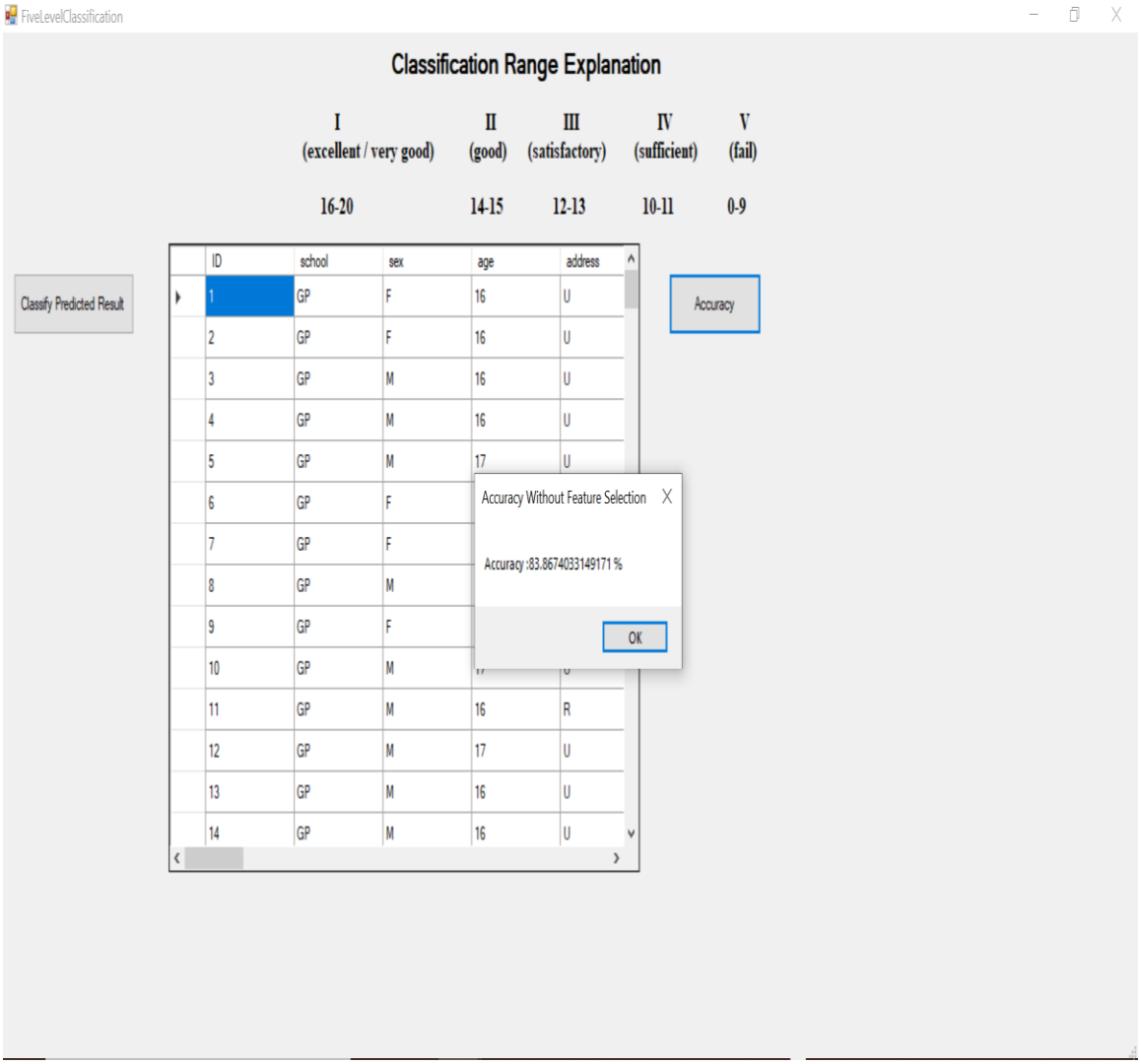


Figure 4.13 Accuracy on 200 Testing Data Sample (Without Feature Selection)

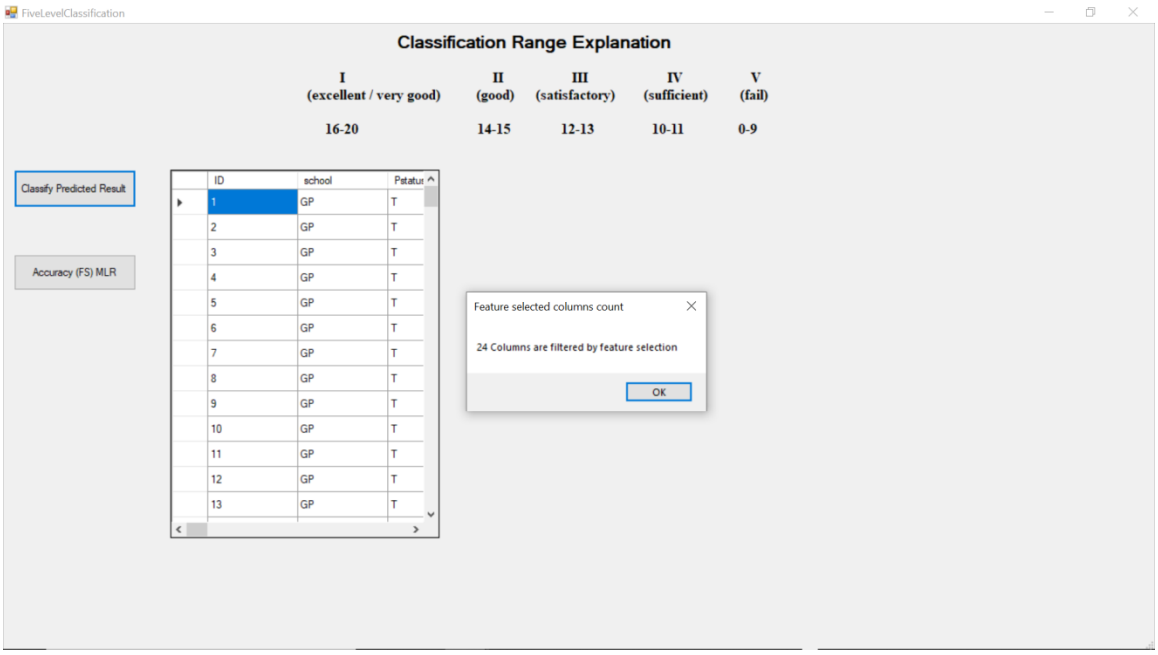


Figure 4.14 Feature Selected Columns Count

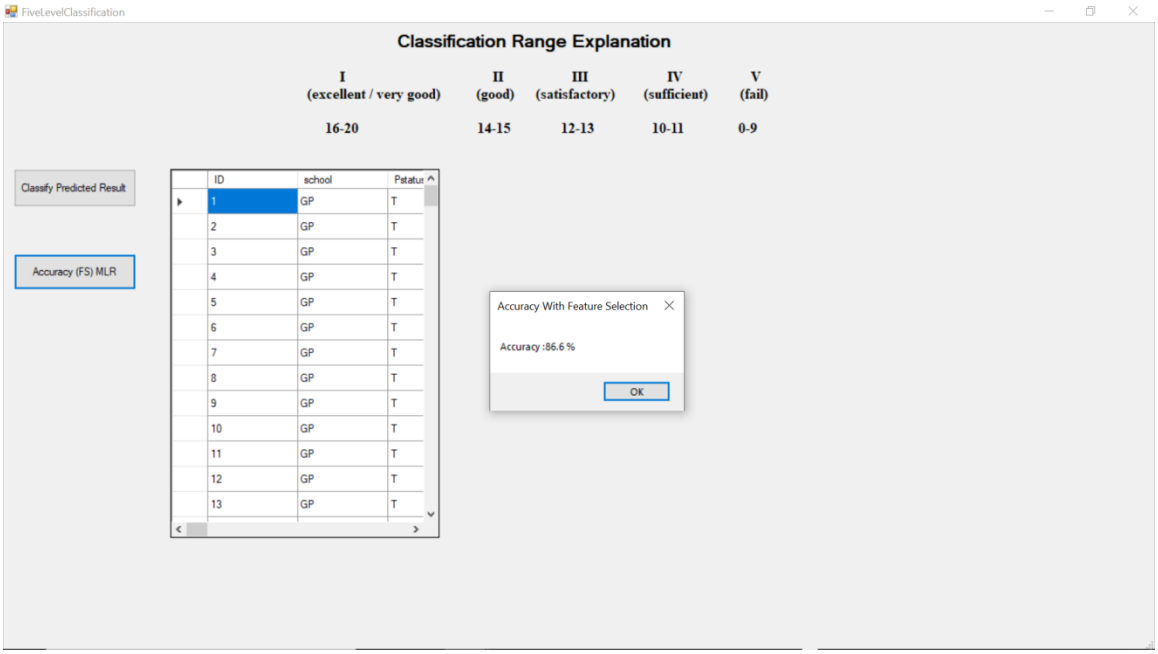


Figure 4.15 Accuracy on 200 Testing Data Sample (With Feature Selection)

4.3 Evaluation of the system

Assuming all of the assumptions for a multiple linear regression have been met, this can be done by generalizing to unseen data. The test data should be set aside and does not looked at until ready to determine how well proposed regression model is generalizing. The train data is used to fit proposed model. The following figure 5.1 and figure 5.2 show the performance analysis of the system with different data size.

Results are evaluated by four evaluation methods (accuracy, precision, recall and F-measure) called confusion matrix.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1) \qquad (4.1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2) \qquad (4.2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3) \qquad (4.3)$$

$$\text{F- measure} = \frac{Precision*Recall}{Precision+Recall} \quad (4) \qquad (4.4)$$

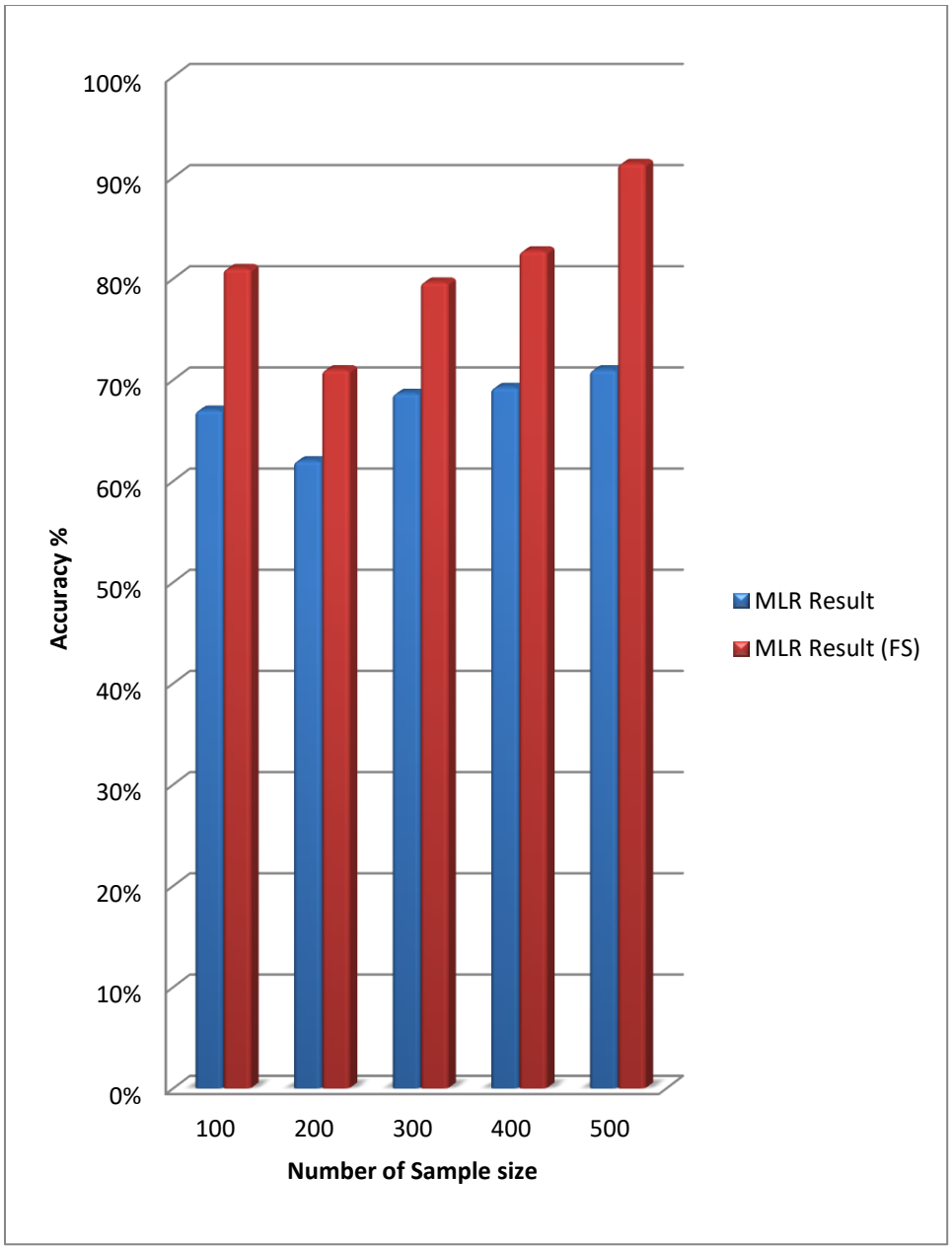


Figure 4.16 Comparison Results of MLR (without feature selection) and MLR (with feature selection)

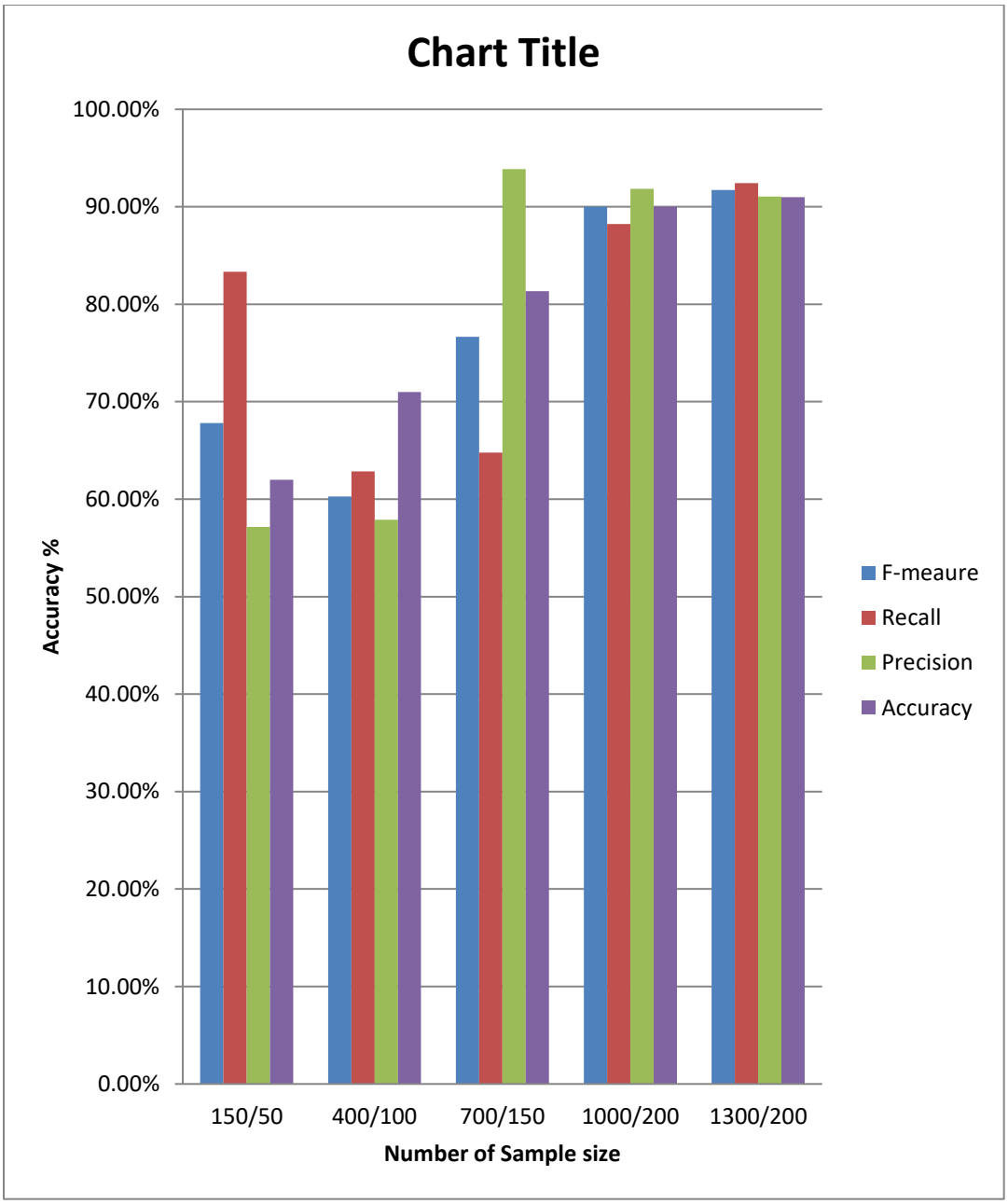


Figure 4.17 System Evaluation with Different Data Size

CHAPTER 5
CONCLUSION

Academic performance of students is a pillar for students' successful future and becomes a big area of interest for all academic institutions over the world. Student's academic performance is achieved with numerous factors of students. The factors considered on UCI focus on demographic attributes and school performance over past years. Several studies have used business intelligence (BI)/ data mining (DM) methods to improve the quality of education and enhance school resource management.

The various datamining techniques are employed in analysing the academic performance of students and one of them is Multiple Linear Regression (MLR). This study detailed the prediction of students' performance of final grades by using first grades, second grades and related featurs. The developed system upgraded when Multiple Linear Regression Method (MLR) with feature selection (Chi-Square) is applied and this system will assist the teacher educators and school related administrator of education system to forecast the students' academic performance and avoid the failure rate of the school and that can be accomplished with numerous factors of students.

The aim of this learning analytics is to improve learning performance by predicting at risk students and providing with the necessary intervention. With the increasing the complexity of the learning environment and diversity of available learning tools, traditional prediction methods have some limitations.

5.1 Benefits of Using Multiple Linear Regression Method

According to the goodness-of-fit and residual analysis results for the established regression model, MLR is suitable for building a student academic performance prediction model for the student and school management system. According to the analysis results, the predictive performance accuracy of MLR with feature selection (Chi-Square) is higher than that of without feature selection approach.

5.2 Limitation and Further Extension

This system performs the prediction process only based on the academic information variables 32. This is not concern on emotion of each student on the academic student age. The psychological point of views also support to raise the educational performance level of student. So, this system can be extended as the depression analysis or emotional on the student mental condition by sentimental analysis to be a perfect system from various point view.

AUTHOR'S PUBLICATIONS

[1] Chan Myae Myint Zu, Kyi Lai Lai Khine, "Prediction of Students' Academic Performance Using Multiple Linear Regression", Parallel & Soft Computing, University of Computer Studies, Yangon, 2022.

REFERENCES

[1] Alaf A A, Thair H and Ibrahim A, Analyzing Students' Academic Performance Through Educational Data Mining, IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[2] Oyerinde O D and Chia P A, Using data mining to predict secondary school student performance, International Journal of Computer Applications (0975 – 8887) Volume 157 – No 4, January 2017.

[3] Paulo C and Alice S , Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression, www3.dsi.uminho.pt, 2008.

[4] R R Rajalaxmi, P Natesan , N Krishnamoorthy ,S Ponni, Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance, 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[5] Sana, Isma F S and Qasim A A, Regression Model for Predicting Engineering Students Academic Performance, 3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019.

[6] Scott B H, Multiple Linear Regression Analysis: A Matrix Approach with MATLAB, Auburn University Montgomery, Alabama Journal of Mathematics, Spring/Fall 2009.

[7] Warey and Gregory, Multivariate Analysis of Variance (MANOVA): I. Theory. Retrieved March 22, 2011.

[8] W Rencher, Schaalje G Bruce, Linear Models In Statistics, Department of Statistics, Brigham Young University, Provo, Utah. second edition.

[9] Z Smyth**,** Nonlinear regression, Encyclopedia of Environmetrics (ISBN 0471 899976), 2002, Volume 3, pp 1405– 1411.

[10] Z Rolph, Tatham L R. and Black CW, Multivariate Data Analysis, fifth edition, chapter 6, pp.326- 352.