STUDENT PERFORMANCE PREDICTOR

A Minor Project-II Report (IT-608)

Submitted by:

ABHISHEK KUMAR DAKSH (0103IT171006)
AYUSH THAKUR (0103IT171027)
DIVYANSH SHARMA (0103IT171035)

Group No.-14

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

at

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY

KALCHURI NAGAR, RAISEN ROAD, BHOPAL (INDIA) - 462021

SESSION JAN - JUNE 2020

DECLARATION

We/I hereby declare that the project entitled “**STUDENT PERFORMANCE PREDICTOR**” submitted for the B.Tech. (Information Technology) degree is **our/my** original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Name & Signature of the students with date

Place: **(1)ABHISHEK KUMAR DAKSH**

Date: **(2 AYUSH THAKUR**

(3) DIVYANSH SHARMA

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY, BHOPAL
श्रद्धावान लभते ज्ञानम्

CERTIFICATE

This is to certify that the project titled “**STUDENT PERFORMANCE PREDICTOR**” is the bonafide work carried out by **Student ABHISHEK KUMAR DAKSH(0103IT171006), AYUSH THAKUR(0103IT171027), DIVYANSH SHARMA(0103IT171035)** student/students of B.Tech. (Information Technology) of Lakshmi Narain College of Technology, Bhopal affiliated to Rajiv Gandhi Proudyogiki Vishwavidyal, Bhopal, Madhya Pradesh (India) during the academic year 2019-20, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (**Information Technology**) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

Signature & Seal of HOD, Information Technology **Signature of the Guide with Date**

Lakshmi Narain College of Technology, Bhopal

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY BHOPAL
श्रद्धावान् लभते ज्ञानम्

Table of Content

CHAPTER 1

INTRODUCTION

1.1PROBLEM STATEMENT:

The Educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation .The future of our society entirely depends upon the next generation , so our educational system must be capable enough to identify each students capabilities and provides them with special attention if needed .But the faculty cannot find out students abilities and their interest easily so that they can enhance them in it. Thus it may affect with poor school results and career of individual . The application predicts the academic performance of students based on the information of each student such as their G1 marks,G2 marks, Health , Absence , Study Time and many more The impact is it help us from fulfilling mission and vision of the institute. If the project get successful then it will be great help for faculty to enhance education system .

1.2 PROJECT OVERVIEW:

We are creating a project which aims to predict the performance of the student . We're using Python and some of its popular data science related packages , such as pandas to read our data from a CSV file and manipulate it for further use , numpy to convert out data into a format suitable to feed our classification model , seaborn and matplotlib for visualizations and few algorithm such as Linear Regression algorithm from sklearn. This algorithm will help us build our predictive model . The application provides user friendly interface and provides the graphical representation of the academic statistics of the students .

The overall activities are broadly categorized into the following steps:

- Data collection and Data set preparation.
- Data preprocessing.
- Data processing.
- Results & Analysis

1.3 OBJECTIVE:

The main objective of this project is to use data science and data mining methodology to study and predict students performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this research, the regression task is used to evaluate students performance and as there are many approaches that are used for performance prediction, the linear regression method is used here. Information like G1 marks, G2 marks, absence , health, etc was collected from the students management system, to predict the performance of the student . The overall vision for the Performance Prediction System is that it will fulfill the following objectives:

- To create a user friendly interface on which the system can be implemented.
- To be able to predict the student performance
- To be able to make the performance prediction methodology more efficient and accurate

1.4 SCOPE:

The scope of our project is to develop an application capable enough –

- To predict the academic performance of the student based on their previous records , mid term marks , attendance status etc .
- To classify student based on their previous academic performance and helps the teachers to provide special attention to those students who needs it .
- To generate the report of the student’s performance.

CHAPTER 2

LITERATURE SURVEY

2.1 Literature Survey

In this literature survey we are trying to analyze various techniques used for predicting student's performance. Student's bad performance is a common problem face today by the faculties , that can be predicted by using different techniques of machine leaning and data mining. In this literature survey we have tried to analyse these techniques and select an appropriate approach that can be applied in our project work . In below section we have organized sub section that gives you brief about these techniques and help us to understand different benefits .

In sub sections we have discussed these techniques like Classification, Clustering Logistic regression and other methods in order to predict the Student's performance . Prediction involves some fields in the data set to predict the values of other variables. On the other hand, Description focuses on finding patterns of the data that can be interpreted by humans. The different algorithm of data mining are used in the field of prediction are discussed in this project.

2.1.1Machine learning

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it learn for themselves.It is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

2.1.1.1 Logistic regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable although many more complex extensions exist. It is estimating the parameters of a logistic model (a form of binary regression). It is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events. Each object being detected in the model would be assigned a probability between 0 and 1 and the sum adding to one. It also measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

2.1.1.2 Linear regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Here the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data and such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values, less commonly, the conditional median or some other quantile is used. It focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

2.1.2Data mining

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods. It is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the knowledge discovery in databases process or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures , visualization, and online updating. The different data mining techniques are described below.

2.1.2.1 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. Classifications are discrete and do not imply order. In a general point of view, regression and classification are two types of predictive factors that regression is used for prediction of continuous data and classification is used for prediction of discrete and nominal data.

2.1.2.2 Clustering

Clustering is the process of grouping set of objects in such a way that objects in the same group(called a cluster) are more similar (in some sense or another) to each other to those in other than to those in other groups(clusters).

2.1.2.3 Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction.

2.1.2.4 Correlation rules

Correlation rules considered as the most important form of discovery and extraction of patterns. This method retrieves all possible patterns of databases. Each algorithm implement on different on different records, and each of them has indicated different functions according to implementation conditions and also data types.

2.2 Existing System

A background study is done to review similar existing systems used to perform student performance analysis. Three existing system are chosen because these systems are similar to the proposed system.

2.2.1. Faculty Support System (FSS)

Shana and Venkatacalam has proposed a framework named Faculty Support System (FSS) which is low in cost as it uses cost effective open source analysis software, WEKA to analyse the students' performance in a course offered by Coimbatore Institute of Technology of Anna University . FSS is able to analyse the students' data dynamically as it is able to update of students' data dynamically with the flow of time to create or add a new rule. The update of new rule is possible with the help from domain expert and the rule is determined by data mining technique such as classification technique. Classification technique is used to predict the students' performance. Besides, FSS focus on the identification of factors that contribute to performance of students in a particular course.

2.2.2. Student Performance Analyser (SPA)

SPA is existing secure online web-based software that enables educators to view the students' performance and keep track of the school's data. The SPA is a tool designed for analysing, displaying, storing, and getting feedback of student assessment data . It is a powerful analyser tool used by schools worldwide to perform analysis and displays the analysis data once raw student data is uploaded to the system. The analysis is done by tracking the student or class to get the overall performance of student or class. It helps to identify the students' performance which is below the expected level, at expected level or above the expected level. This would allow the educators or staffs to identify the current students' performance easily. Other than that, it enables various kinds of students' performance report such as progress report and achievement report to be generated.

2.2.3. Intelligent Mining and Decision Support System (InMinds)

InMinds helps Universitiy Malaysia Sarawak (UNIMAS) to monitor the performance of various areas in every UNIMAS's departments [2]. The system enables top and mid-management in UNIMAS to have a clear look on the areas that needed attention by looking at the figures, revenues and risks. The features, ease of use and flexibility provided by the system makes the performance analysis in UNIMAS to be performed in an ideal solution. Charts are provided by the system for ease of student performance's interpretation. From the reviews on these existing systems, useful techniques and features could be applied into the proposed system for a better system's performance. The WEKA is chosen as a tool for data mining because it is open source software.

Chapter-3

SYSTEM ANALYSIS AND DESIGN

3.1Requirement Specification

3.1.1 Functional Requirements

3.1.1.1 - System must be able to accept the student detail,to create dataset.

3.1.1.2 - Our system must be able to predict student performance by applying proper learning techniques

3.1.2 Non functional Requirements

3.1.2.1 Performance Requirements

It requires substantial CPU performance for proper machine learning.

3.2 Use-Case diagram

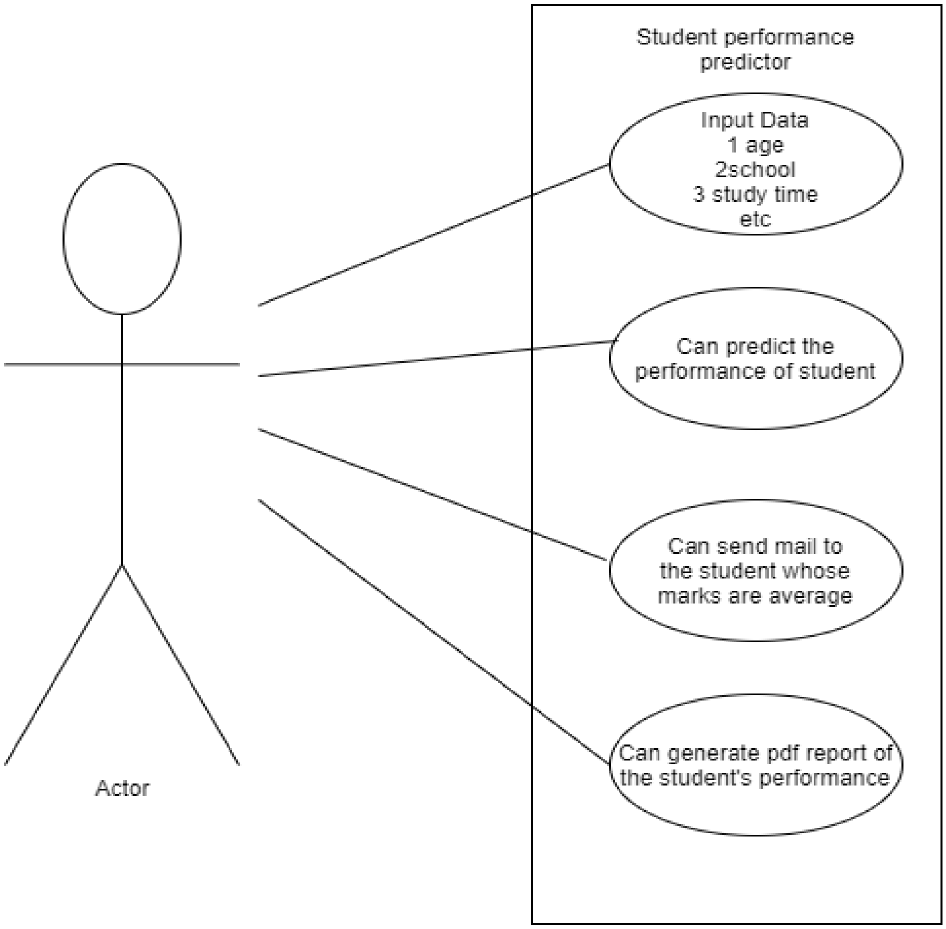

Student performance predictor
Input Data
1 age
2school
3 study time
etc
Can predict the performance of student
Can send mail to the student whose marks are average
Can generate pdf report of the student's performance
Actor

Fig1. Use Case Diagram for predict student’s performance

The above Fig 1.shows Use case diagram for predicting student performance. The Use case diagram comprises of one actor that is user .It describes that user needs to provide his/her basic information such as age, gender, G1 marks ,G2 marks, Absence , Mother’s education , Father’s education etc. based on these information our system predict that whether that person have diabetes or not .

3.4 Flow chart

The Flow chart includes the selection of the right attributes from the large database, based on the sensitivity of the dataset and the problem statement. The selection of optimal attributes for the problem, it requires an overall analysis of the attributes and ignoring the irrelevant attributes. The input dataset includes various attributes and its description. Selection of the right attributes adheres to the quality input dataset and quality results from the analyses can be expected. Our approach includes 5 steps.

1. Analysis of the attributes and importance of the attributes on the problem stated.
2. Assigning a sequence of the dataset attributes from $n_{i=0}$ to the $n_{k=max}$, where max is the total number of attributes, and i is the attribute-1 (main cause). Input: Attribute-1 (Main attributes responsible for the cause).
3. Process: Attribute-1 Co-relates the other attribute-n, and generates the value.

   Co-relation Value = [Attributemax−∑i=1NAttribute(xi)]2−1Co-relation

   Value = [Attributemax−∑i=1NAttribute(xi)]2−1

   The process is continued with other attributes, values are compared with each attribute, if the value difference is more than the other attribute, then attribute has less significance, i.e. value-1 is compared with value-n. The best attributes are selected and arranged in a significant order and the final optimal features-dataset is given to the classification techniques.
4. Output: Based on the best attributes selection, the results of the analyses techniques can be improved. The flow chart of the technique is given below Fig.2.

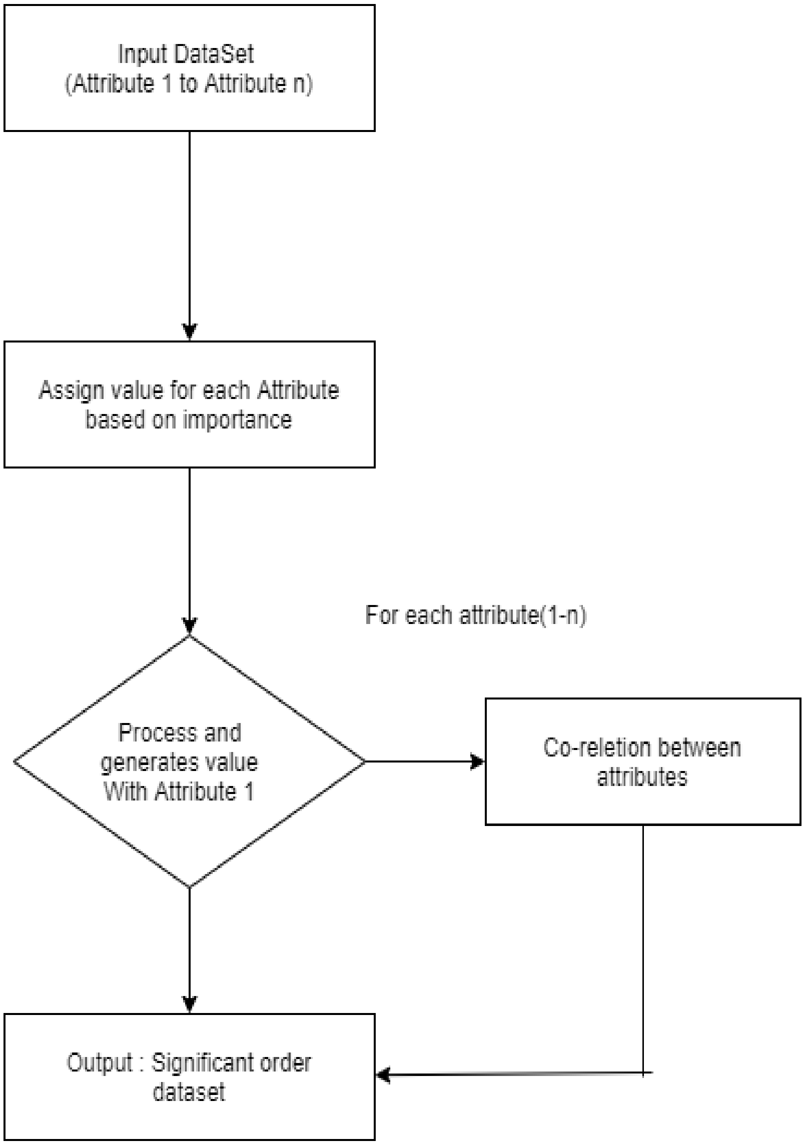

Input DataSet
(Attribute 1 to Attribute n)
Assign value for each Attribute based on importance
For each attribute(1-n)
Process and generates value With Attribute 1
Co-reletion between attributes
Output : Significant order dataset

Fig.2 Flow chart for attribute selection

3.3 Hardware Specification

Memory Space: Minimum – 32 MB, Recommended – 64 MB

HDD: To install the software at least 2 GB and the data storage is depending upon the organizational setup.

PROCESSOR: Intel Pentium IV, 1GHZ or above

RAM: 256MB or above.

KEYBOARD: Standard 104 Keys(QWERTY)

3.4 Software Specification

Operating System: Windows 10.

Language : Python

IDE : Spyder

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc). Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick. Python can be treated in a procedural way, an object-orientated way or a functional way.

Spyder is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

CHAPTER 4

PROPOSED WORK

4.1 Module Description

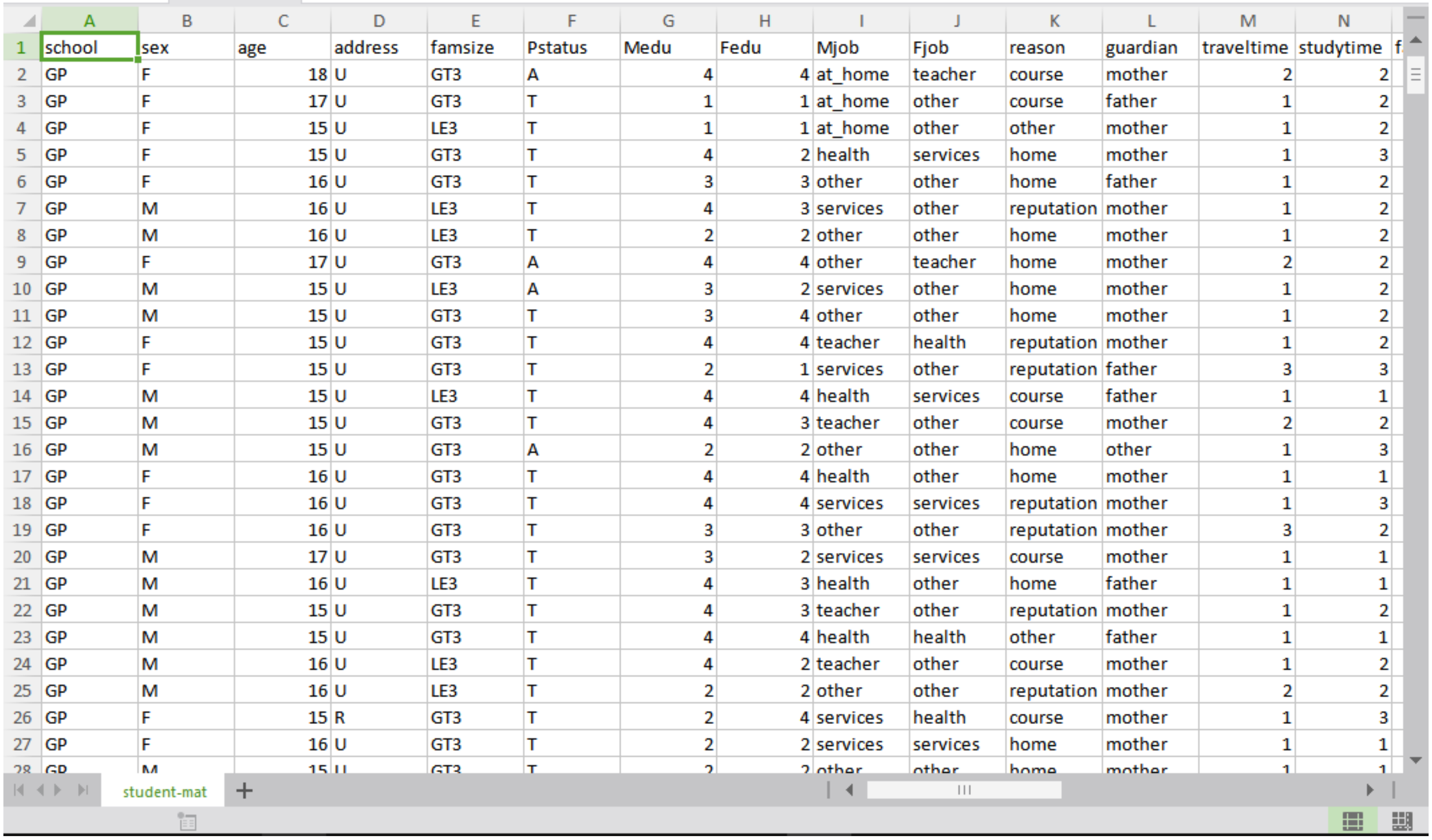
There are a few features from the existing systems that are employed during the design and implementation phase of the proposed system. These features and functionalities include the user interface, students' performance prediction, illustration displays and report generation. A good user interface provides an user-friendly interface as it is easy to be navigate and not complicated. Meanwhile, the students' performance prediction is included into the proposed system to make sure the objectives are achieved. Furthermore, the generation of reports which is saved in database. From these features found in proposed system, all the user requirements would be fulfilled.
E-mail would be send to the Student's parents whose result is predicted below the average.

The user requirements collected from lecturers of FCSIT during the system analysis phase are as follows:-

i. Able to help lecturers to automatically predict students' performance in courses.
ii. Able to keep track and retrieve students' performance in a particular course and semester .
iii. Able to view the factors that affect the students' prediction result.
iv. Able to generate students' reports.

4.2 Dataset Description

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features) and it was collected by using school reports and questionnaires. The dataset is provided regarding the performance in two distinct subjects: Mathematics and Portuguese language .The dataset was modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	f
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	
7	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	
11	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	
12	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	
13	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	
14	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	
15	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	
16	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	
17	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	
18	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3	
19	GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	
20	GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	
21	GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	
22	GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother	1	2	
23	GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	
24	GP	M	16	U	LE3	T	4	2	teacher	other	course	mother	1	2	
25	GP	M	16	U	LE3	T	2	2	other	other	reputation	mother	2	2	
26	GP	F	15	R	GT3	T	2	4	services	health	course	mother	1	3	
27	GP	F	16	U	GT3	T	2	2	services	services	home	mother	1	1	
28	GP	M	15	U	GT3	T	2	2	other	other	home	mother	1	1	

Fig: Dataset- student.csv

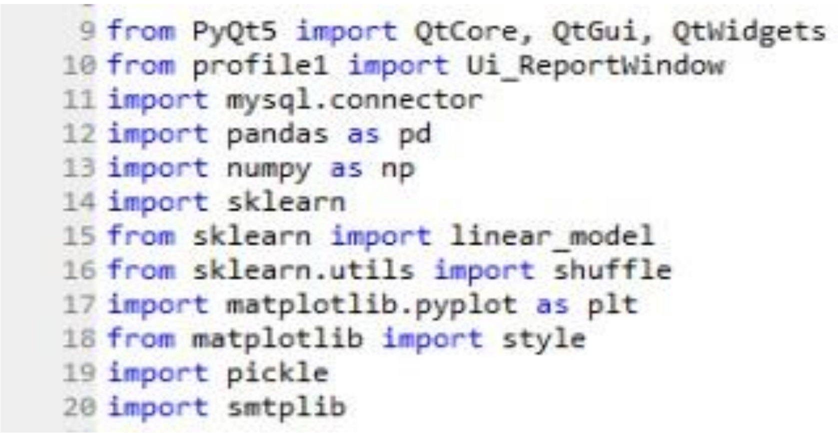
CHAPTER 5

ALGORITHMS AND PSEUDO CODE

5.1 Pseudo Code

1. Importing libraries numpy, pandas, matplotlib, math, scikit learn and cufflinks

- numpy is imported for support of multidimensional array and matrix.
- pandas is used for data analytics and for reading of dataset which is in .csv format.
- from math, sqrt library is used which has inbuilt function which returns the square root of the value.
- From scikit learn library the preprocessing method is used, from which StandardScaler is imported which standardizes a feature by subtracting the mean and then scaling to unit variance.
- From scikit learn library the metrics is used where precision_recall_fscore_support is imported for calculation of precision, recall, f1 score and support.
- cufflinks is imported for plotting the 2D and 3D graphs.

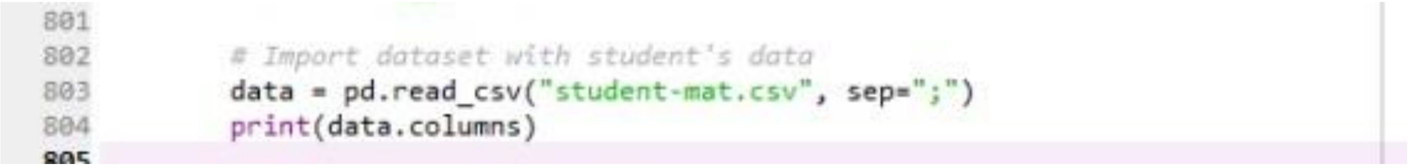


```
from PyQt5 import QtCore, QtGui, QtWidgets
from profile1 import Ui_ReportWindow
import mysql.connector
import pandas as pd
import numpy as np
import sklearn
from sklearn import linear_model
from sklearn.utils import shuffle
import matplotlib.pyplot as plt
from matplotlib import style
import pickle
import smtplib
```

Fig. Importing libraries numpy, pandas, matplotlib, sklearn and pickle

2. Loading data

The data is stored in a CSV format namely student-mat.csv. The dataset is read into a pandas dataframe called data as shown below. The data is loaded in the environment .



```

        # Import dataset with student's data
        data = pd.read_csv("student-mat.csv", sep=";")
        print(data.columns)
```

Fig. Data loaded

CHAPTER 6

TESTING PROCESS

6.1Testing Methodology

In order to predict the student marks that a student is passed or not. Firstly using The dataset consist of several marks predictor variables and one target variable, that is the outcome. The dataset used is Portugal student data. The dataset contains many attributes like school,sex,age,address,famsize,guardian,etc. Here we develop a system using techniques like multi linear regression. It is a predictive algorithm used to assign observations to a continious set of data.It is a Machine Learning and predictive analysis algorithm and is based on the concept of probability. We expect our system to give us a set of outputs based on probability when we pass the inputs from dataset and returns a probability score between 0 and 1.

6.2Test Cases and Test Steps

In the proposed Student Marks Prediction system, used to predict whether particular student will pass or not . The input is based on a person details having attributes like .

Step-1: Importing Libraries

Step-2: Loading Dataset

Step-3: All values are checking for null values and class distribution except the class variable which contains either “0”or “1”.

Step-4: After the data set is normalised, the class variable is excluded and check for Correlation that helps to know which attributes are highly dependent on the prediction variable Outcome.

Step-5: plotting the false positive rate and true positive rate in 2D graph using machine learning techniques.

CHAPTER 7

Result

7.1 Result

In the present study, the main objective is to find a system that predicts whether the student will pass or not. In order to fulfil the objective we are using different approaches of supervised learning like Logistic algorithms and other were used and implemented. This system when included in real time applications in education system can be used to predict marks with greater accuracy. The model can be enhanced by using real time dataset or school's student data. It would be beneficial if user gets a mobile application which not only predicts marks but also stores the student information.

Chapter 8

Conclusions and Future Enhancement

8.1Conclusions & Future Enhancement

Machine learning has the great ability to revolutionize the students marks prediction with the help of advanced computational methods and availability of large amount of dataset. This work has described a machine learning approach to predicting marks. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of teachers to help them make better decision about the student status.

The Results can be improved by applying the Data cleaning and Feature Scaling have to be done with the data. Then running the prepared data with the logistic regression to get the improved results. As a result it can help to improve the health conditions.