**PROFESSIONAL TRAINING REPORT**

**at**

**Sathyabama Institute of Science and Technology (Deemed to be University)**

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

By

**SUJEETH S.R**

**REG. NO. 39110982**

****

**DEPARTMENT OF COMPUTER SCIENCE ANDENGINEERING**

**SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

**CHENNAI – 600119, TAMILNADU**

**NOVEMBER 2021**

|  |  |  |
| --- | --- | --- |
|  | **SATHYABAMA**  **INSTITUTE OF SCIENCE AND TECHNOLOGY**  **(DEEMED TO BE UNIVERSITY)**  **Accredited with Grade “A” by NAAC**  (Established under Section 3 of UGC Act, 1956)  JEPPIAAR NAGAR, RAJIV GANDHI SALAI  CHENNAI– 600119  [**www.sathyabama.ac.in**](http://www.sathyabama.ac.in) |  |



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the Bonafide work of **SUJEETH S.R (Reg. No: 39110982)** who carried out the project entitled “**PREDICT SCHOOL STUDENT PERFORMANCE**” under my supervision from June 2021 to November 2021.

**Internal Guide**

## Mr. Amandeep Singh K B.tech, M.tech (Ph.D.),

**Head of the Department**



**Submitted for Viva-voce Examination held on**

**InternalExaminer ExternalExaminer**

**DECLARATION**

I, **SUJEETH S.R** hereby declare that the project report entitled **Predict Student Performance was** done by me under the guidance of **Mr. Amandeep Singh, B.tech, M.tech, (Ph.D.).** submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

**DATE OF SUBMISSION: 10/11/2021**

**SUJEETH SR**

**SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

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

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

I would like to express my sincere and deep sense of gratitude to my Project Guide

**Mr. Amandeep Singh K, Assistant Professor, B.tech, M.tech, (Ph.D.).,** for her valuable guidance, suggestions, and constant encouragement paved way for the successful completion of my project work.

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

**TRAINING CERTIFICATE**



**ABSTRACT**

Although the educational level of the Portuguese population

has improved in the last decades, the statistics

keep Portugal at Europe’s tail end due to its high student

failure rates. In particular, the lack of success in the

core classes of Mathematics and the Portuguese language

is extremely serious. On the other hand, the

fields of Business Intelligence (BI)/Data Mining (DM),

which aim at extracting high-level knowledge from raw

data offer interesting automated tools that can aid the

education domain. The present work intends to approach

student achievement in secondary education using

BI/DM techniques. Recent real-world data (e.g.

student grades, demographic, social, and school-related

features) was collected by using school reports and questionnaires.

The two core classes (i.e. Mathematics and

Portuguese) were modeled under binary/five-level classification

and regression tasks. Also, four DM models

(i.e. Decision Trees, Random Forest, Neural Networks

and Support Vector Machines) and three input

selections (e.g. with and without previous grades) were

tested. The results show that a good predictive accuracy

can be achieved, provided that the first and/or second

school period grades are available. Although student

achievement is highly influenced by past evaluations, an

explanatory analysis has shown that there are also other

relevant features (e.g. number of absences, parent’s job

and education, alcohol consumption). As a direct outcome

of this research, more efficient student prediction

tools can be developed, improving the quality of education

and enhancing school resource management.

TABLE OF CONTENTS

CHAPTER TITLE PAGE-NO

ABSTRACT

1. INTRODUCTION

1.1 INTRODUCTION

1.2 RESEARCH MOTIVATION AND

STATE OF THE ART

* 1. RESEARCH APPROACH, DATA

SELECTION, AND PREPROCESSING

* 1. Python Data Analytics

2. Development Environment Software

2.1 AIM OF THE PROJECT

2.2 BACKGROUND OF THE PROJECT

2.3 SCOPE

2.4 OPERATING SYSTEM

2.5 Development tools and

programming language

2.5.1 Jupyter Notebook

2.5.2 Python

2.6 HARDWARE

2.6.1 PROCESSOR

2.6.2 RAM

2.7 OPERATION ENVIRONMENT

3. MATERIALS AND METHODS

3. STUDENT DATA

3.1 NAVIE BAYER

3.2 DECISION TREE

3.4 DATA SET INFORMATION

3.5 DATA SET DESCRIPTION

TABLE 1. THE PREPROCESSED STUDENT

RELATED VARIABLES

TABLE 2. THE FIVE-LEVEL

CLASSIFICATION SYSTEM

3.6 DATA PROCESSING

3.7 EXPERIMENTAL RESULTS

3.8 READING DATA

3.8.1 STUDY DATASET

3.8.2 SPLITTING THE DATA

3.8.3 LINEAR REGRESSION

4. RESULT AND DISCUSSION, PERFORMANCE ANALYSIS

4.1 PREDICT THE TEST SET

4.2 EVALUATE THE MODEL

4.3 PLOT THE RESULT

4.4 PREDICTED VALUES

5. SUMMARY AND CONCLUSIONS

REFERENCES

APPENDIX:

A.SCREENSHOTS

B.SOURCE CODE

**CHAPTER 1**

***1.1* INTRODUCTION**

Recently, online systems in education have increased, and student digital data has come to big data size. This makes pit possible to draw rules and predictions about the students by processing educational data with data mining techniques. All kinds of information about the student’s socioeconomic environment, learning environment, or course notes can be used for prediction, which affects the success or failure of a student.

In this study, the successes of the students at the end of the semester are estimated by using the student data obtained from secondary education of two Portuguese schools. The aim of this study is to predict the students’ final grades to support the educators to take precautions for the children at risk. A number of data pre-processing processes were applied to increase the accuracy rate of the prediction model. A wrapper method for feature subset selection was applied to find the optimal subset of features. After that, three popular data mining algorithms (decision tree, random forest, and Naive Bayes) were used and compared in terms of classification accuracy rate. In addition, this study also investigates the effects of two different grade categorizations on data mining: five-level grade categorization and binary grade categorization.

The two core classes (i.e. Mathematics and Portuguese) will be modeled under three DM goals:

* binary classification (pass/fail);
* classification with five levels (from I very good or excellent to V - insufficient); and
* regression, with a numeric output that ranges between

zero (0%) and twenty (100%).

For each of these approaches, three input setups (e.g.

with and without the school period grades) and four

DM algorithms (e.g. Decision Trees, Random Forest)

will be tested. Moreover, an explanatory analysis will

be performed over the best models, in order to identify

the most relevant features.

***1.2***  **RESEARCH MOTIVATION AND STATE OF THE ART**

The rationale behind the research work described in this paper is based on the great potential that is seen in using data mining methods and techniques for the effective usage of university data. The discussions with high-level managers and administrators of a famous and prestigious Bulgarian university have I level lead to the identification of existing needs for better knowing the students and performing a more effective university marketing policy. The literature review reveals that these problems have been of interest to various researchers during the last few years. The development of data mining models for predicting student performance at various levels, and comparison of those models, are discussed in a number of research papers. In 2000 the results of a study are described [4] aimed at finding weak students and involving them in additional courses for advanced support by extracting association rules from data. The retention of students is a problem discussed also by Luan, who implemented clustering, neural network, and decisions tree methods to predict the students at risk of failure [5], [6]. Data mining methods are implemented for modeling online student grades [7], using three classification approaches used (binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). Kotsiantis et al. [8] also deal with predicting student performance, recognizing dropout-prone students based on demographic characteristics (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment). Pardos et al. [9] use data from an online tutoring system for teaching Math and implement a regression approach for predicting the math test score based on individual skills. Superby et al. [10] predict students at risk of drop-out, determining factors influencing the achievement of the first-year university students, classifying students into three classes – low-risk, medium-risk and high-risk, using Decision trees, Random forest method, Neural networks, and Linear discriminant analysis. Vandamme et al. [11] also deal with the early identification of three categories of students: low, medium, and high-risk students using Decision trees, Neural networks, and Linear discriminant analysis. Cortez and Silva in [12] attempt to predict student failure by applying and comparing four data mining algorithms, Decision Tree, Random Forest, Neural Network, and Support Vector Machine. The implementation of predictive modeling for maximizing student recruitment and retention is presented in the study of Noel-Levitz [13]. The development of enrolment prediction models based on student admissions data by applying different data mining methods (Decision trees, Rule induction, Feature subset selection) is the research focus of Nandeshwar [14]. Dekker et al. [15] focus on predicting students’ dropout. Kovačić in [16] uses data mining techniques (feature selection and classification trees) to explore the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block) that may influence persistence or dropout of students, identifying the most important factors for student success and developing a profile of the typical successful and unsuccessful students. Ramaswami et al. in [17] focus on developing a predictive data mining model to identify the slow learners and study the influence of the dominant factors on their academic performance, using the popular CHAID decision tree algorithm.

*1.3* RESEARCH APPROACH, DATA SELECTION, AND PREPROCESSING :

The performed research is based on the CRISP-DM (CrossIndustry Standard Process for Data Mining) model, a nonproprietary, freely available, and application-neutral standard for data mining project implementation, widely used by researchers in the data mining field during the last ten years [18]. It is a cyclic approach, including six main phases – Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment, with a number of internal feedback loops between the phases, resulting from the very complex non-linear nature of the data mining process and ensuring the achievement of consistent and reliable results. The open-source software WEKA, offering a wide range of classification methods for data mining [19], is used as a data mining tool for research implementation. First of all, the business problem is identified – it is the growing need of university management for better knowing the university students and predicting their performance in order to approach in the marketing campaigns exactly those students that will be most successful in the university education process. The stated business problem is transformed into a data mining task – the task for classifying students into two categories – successful and unsuccessful, by analyzing the available student data with selected data mining methods for classification. The next phase in the research implementation includes the data selection and pre-processing, crucial activities within each data mining project, highly influencing the quality of the final results. After studying the application process for student enrollment at the University and reviewing the procedures for collecting and storing data about the academic performance of the university students, it is established that the university data is generally organized in two databases. All the data related to the university admission campaigns are stored in the University Admission database, including personal data of university applicants (names, addresses, secondary education scores, selected admission exams, etc.), data about the organization and performance of the admission exams, scores achieved by the applicants at the admission exams, data related to the final classification of applicants and student admission, etc. All the data concerning student performance at the university is stored in the University Students Performance database, including student personal and administrative data, the grades achieved at the exams on the different subjects, etc. For the purposes of the study, student data from both databases is carefully selected, extracted and combined in a new flat file (in this case Excel file) used for the data mining analysis in the WEKA software tool The provided flat file contains data about 10330 students that have been enrolled as university students during the period between 2007 and 2009, described by 20 parameters, including gender, birth year, birthplace, living place and country, type of previous education, profile and place of previous education, total score from previous education, university admittance exam and achieved score, total university score at the end of the first year, etc. The data is carefully studied and subjected to many transformations. Some of the parameters are removed, e.g. the “Birthplace” and the “Place of living” fields containing data that is of no interest to the research, the “Country” field containing only one value (Bulgaria) because the data concerns only Bulgarian students, the “Type of previous education” field which has only one value as well because concerns only students who have finished secondary education. Some of the variables, containing important data for the research, are text fields where free text is being entered at the data collection stage. Therefore, these variables are processed and turned into nominal variables with a limited number of distinct values. Such a parameter is the “Profile of the secondary education” which is turned into a nominal variable with 9 distinct values (e.g. language, math, natural sciences, economics, technical, sports, arts, etc.). The “Place of secondary education” field is also preprocessed and transformed into a nominal variable with 7 distinct values, corresponding to the capital city and the 6 geographic regions in Bulgaria – North-East, North-Central, North-West, South-East, South-Central, and South-West. A new numeric variable is added – the “Student age at enrollment”, calculated by subtracting the values contained in the “Admission year” and “Birth year” fields. Another important operation during the preprocessing phase is also the transformation of some variables from numeric to nominal (e.g. age, admission year, current semester, total university score, etc.) because they are much more informative when interpreted with their nominal values. The data is also being studied for missing values, which are very few and could not affect the results, and obvious mistakes, which are corrected. Essentially, the challenge in the presented data mining research is to predict the student university performance based on the available student pre-university and university performance data. This is achieved by solving a classification data mining task. A binary categorical target variable is constructed, based on the original numeric parameter “University average score” (the average numeric score achieved by the students at the end of the first year at the University). The predicted variable has two distinct values, corresponding to the two classes in which the students are classified – Weak and Strong. Since a six-level scale is used in the Bulgarian educational system for evaluation of student performance at schools and universities, the students with average university score that is lower than 4.50 are classified as “Weak”, and the students with average university score equal to or higher than 4.50 are classified as “Strong”. The final dataset, on which the selected classification data mining algorithms are applied, contains 10067 instances and 14 attributes (summarized in Table 1). activities.

***1.4 Python Data Analytics:***

Data Analysis can help us to obtain useful information from data and can provide a solution to our queries. Further, based on the observed patterns we can predict the outcomes of different business policies.

## *1.4.1Understanding the basic of Data Analytics*

### *1.4.1. Data:*

The kind of data on which we work during the analysis is mostly of the CSV (comma separated values) format. Usually, the first row in the CSV files represents headers.

### *1.4.2 Packages Available:*

There's a diversity of libraries available in [Python](https://www.javatpoint.com/python-tutorial) packages that can facilitate easy implementation without writing a long code.

**CHAPTER 2**

*2.1 Aim of the project:*

The aim of the study is to empirically investigate and compare the use of multiple data sources, different classifiers, and ensembles of classifiers techniques in predicting student academic performance. The study will compare the performance and efficiency of ensemble techniques that make use of different combinations of data sources with that of base classifiers with a single data source Design/methodology/approach.

For each of these approaches, three input setups (e.g. with and without the school period grades) and four DM algorithms (e.g. Decision Trees, Random Forest) will be tested. Moreover, an explanatory analysis will be performed over the best models, in order to identify the most relevant features.

***2.2 Background of the project:***

As mentioned in the previous section, estimating the final grade of the students based on different attributes related to their learning activities can be viewed as a typical regression problem. Currently, there are many available models and several collections of machine learning algorithms, among which the best known is arguably Weka [22]. In this paper, we aim to compare the performance of classical algorithms with an original one, called Large-Margin Nearest Neighbor Regression (LMNNR). Following our experience from previous work [31], here we only consider two classical models that provided the best results for our particular problem, i.e., Random Forest (RF) and k-Nearest Neighbors (kNN), which we briefly describe below.

A random forest [9] is composed of a collection of classification or regression trees. Each tree is generated using random split tests on slightly different data, using bagging. The output of a new instance is computed by aggregating the outputs of the individual trees, by voting or averaging.

***2.3 Scope:***

The scope of the present study is limited in the investigation of the effects of cognitive factors (i.e., the above-stated eight predictor variables) on student academic performance in secondary school students.

The effects of a student’s non-cognitive factors (such as learning style, self-efficacy, motivation and interest, time devoted to learning, family background, race, and many others 14), the instructor’s teaching effectiveness and preparation 15, as well as teaching and learning environment 16 on student academic performance is beyond the scope of the present study and will be dealt with in the future study

***2.4 Development Environment Software***

*2.4.1 Operating System:*

*Windows 10*

**Windows 10 is selected as the developing operating system because** Windows has the biggest selection of software available for its platform than any other operating system. The benefit of this is that users get to choose from a wider variety of options. This creates healthy “competition” for users, where software developers really have to push boundaries to produce the best program possible. Anything less than the best will result in the user’s picking the next program on the list. This alone does wonders in motivating software developers to deliver excellent solutions that meet users’ needs.

*2.5 Development tools and programming language:*

*Jupyter Notebook and Python:*

*2.5.1 Jupyter Notebook:*

### Jupyter Notebook combines live code, graphics, visualizations, and text in shareable notebooks that run in a web browser. Originally developed for data science applications written in Python, R, and Julia, Jupyter Notebook is useful in all kinds of ways for all kinds of projects.

Most people have their first exposure to Jupyter Notebook by way of a data visualization, a shared notebook that includes a rendering of some data set as a graphic. Jupyter Notebook lets you author visualizations, but also share them and allow interactive changes to the shared code and data set.

*2.5.2 Python:*

One of the most common uses for Python is in its ability to create and manage data structures quickly — Pandas, for instance, offers a plethora of tools to manipulate, analyze, and even represent data structures and complex datasets.

This includes time series and more complex data structures such as merging, pivoting and slicing tables to create new views and perspectives on existing sets. Elsewhere, tools like Scikit-Learn (also known as Sklearn) provide advanced analytics tools combined with complex machine learning capabilities.

This allows you to build more sophisticated models, perform more complex and multivariate regressions, as well as data preprocessing. Combined with libraries such as iPython and NumPy itself, these tools can form the foundation of a powerful data analytics suite. Additionally, you can use Python to write your own data analysis algorithms that can be directly integrated into your business intelligence tools via API.

***2.6 Hardware****:**2.6.1 Processor*

Intel Core i5 10 thGen Processor provides better processing capabilities and better cooling technology to our CPU. With an Intel processor, we can run our laptop for a long time without the need to switch it off. Besides that, intel processors can help us to boost up the CPU processing power. By using this, we can keep developing the Library Management System without the need to worry that the laptop cannot support it.

*2.6.2 Ram: 8 Gb*

In order to support Anaconda Navigator and Server, we use 8Gb Ram to avoid any problem occurring during the development phase. Besides that, the Local Server can process faster when running Additional statements with 8Gb ram. It can save a lot of time if it totals up the processing time.

***2.7 Operation Environment***

The table shown below is the minimum requirement :

**Table** **2.1 Table for operation environment**

|  |  |
| --- | --- |
| Processor | Intel Pentium 233Ghz or better performance |
| Operating System | Microsoft Window XP, Vista or Window 7 |
| Memory | 2GB RAM |
| Screen Resolution | Monitor with screen resolution minimum 1024 x 768 |
| Hard disk Space | Minimum 5GB to include database usage for future |

**CHAPTER-3**

MATERIALS AND METHODS

*3. Student Data :*

In Portugal, secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. There are several courses (e.g. Sciences and Technologies, Visual Arts)that share core subjects such as the Portuguese Language and Mathematics. Like several other countries

(e.g. France or Venezuela), a 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, students are evaluated in three periods and the last evaluation (G3 of Table 1) There are basically three data mining methods: classification, clustering, and association rule mining. In this study, we focus on the classification task.

The methods to be used in data mining may differ depending on the field of study and the nature of the data we have. In this study, three well-known classification algorithms (decision tree, random forest, and naive Bayes) were employed on the educational datasets to predict the final grades of students.

3.1 Naive Bayes:

Naive Bayes classifiers are a family of algorithms. These classifiers are based on Bayes’ Theorem, which finds the possibility of a new event based on previously occurring events. Each classification is independent of one another but has a common principle.

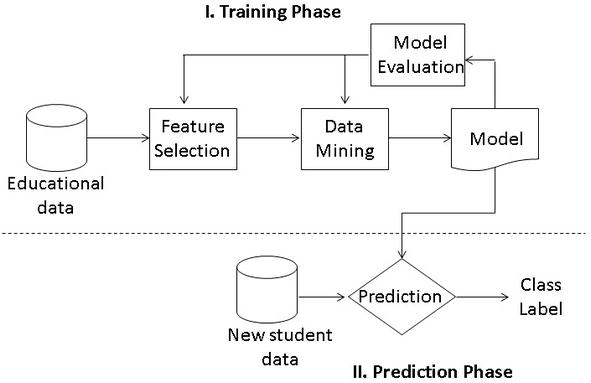
3.2 Decision tree:

A decision tree uses a tree-like graph. Decision trees are like flowcharts but not noncyclic. The tree consists of nodes and branches. Nodes and branches are arranged in a row. The root node is on the top of a tree and represents the entire dataset. Entropy is calculated when determining nodes in a tree. It models decisions with efficacy, results, and resource costs. In this study, the decision tree technique is preferred because it is easy to understand and interpret.

3.3 Random forest:

Random forest is an ensemble learning algorithm. It is a supervised classification method. It consists of randomly generated many decision trees. The established forest is formed by the decision trees community trained by the bagging method, which is one of the ensemble methods. The random forest creates multiple decision trees and combines them to achieve more accuracy rates and stable predictions.

Figure 1 illustrates the workflow of the data mining model for classification. In the first step, feature selection algorithms are applied to the educational data. Next, classification algorithms are used to build a good model which can accurately map inputs to desired outputs. The model evaluation phase provides feedback to the feature selection and learning phases for adjustment to improve classification performance. Once a model is built, then, in the second phase, it is used to predict label of new student data.



**3.4 Data Set Information:**

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. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were 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 (see paper source for more details).

**3.5 Dataset description:**

In this study, two publically available datasets [19] were used to predict student performances. Both datasets were collected from the secondary education of two Portuguese schools. Dataset attributes are about student grades and social, demographic, and school-related features. All data were obtained from school reports and questionnaires. The first dataset has information regarding the performances of students in Mathematics lessons, and the other one has student data taken from Portuguese language lessons. Both datasets have 33 attributes as shown in Table 1.

## **Table 1: The preprocessed student-related variables**

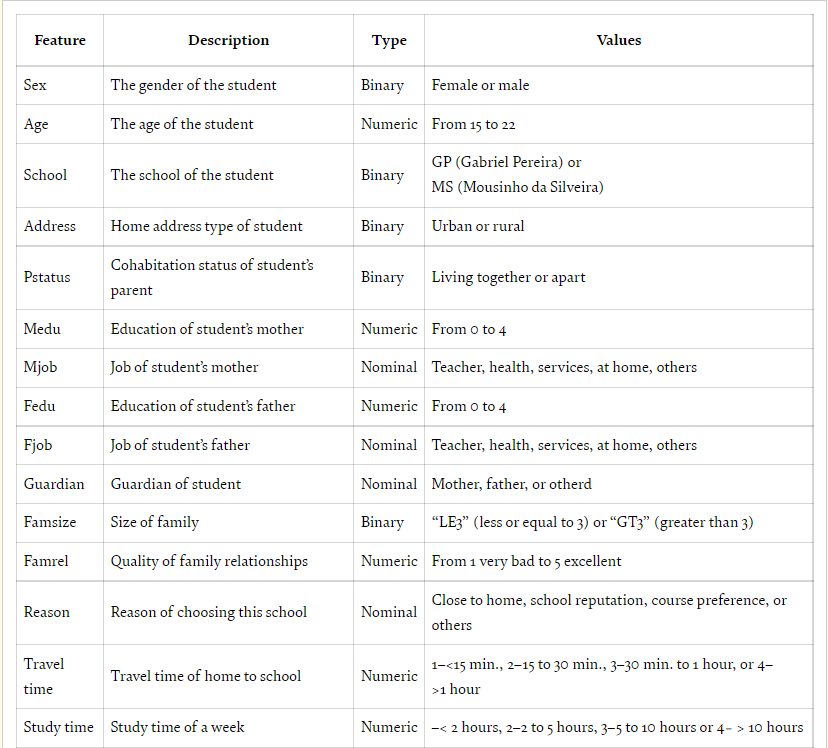




Table 2: The five-level classification system

Five-level grading categories.

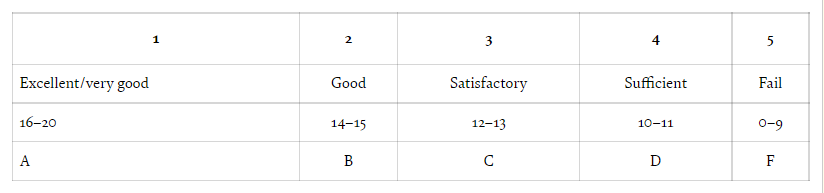
To compare the results, we also categorized the final grade as “passed” and “fail.” As shown in Table 3, the range of 0–9 corresponds to F, and it means “fail”; the range of 10–20 refers to A, B, C, and D, and it means “pass.”



3.6 Data preprocessing:

In the raw dataset, the final grade is in the range of 0–20 as with many European countries, where 0 is the worst grade and 20 is the best score. Since the final grade of the students is in the form of an integer, the predicted class should be in the form of categorical values, the data needed to be transformed into categories according to a grading policy. In the study, we used and compared two different grading systems: five-level grading and binary grading systems.

We first categorized the final grade in five groups. These ranges are defined based on the Erasmus system. As shown in Table 2, the range 0–9 refers to grade F, which is the worst grade and corresponds to the “fail” label. The others (10–11, 12–13, 14–15, and 16–20) correspond to D (sufficient), C (satisfactory), B (good), and A (excellent/very good) class labels, respectively. Binary fail/pass category.



3.7 Experimental results:

As a pre-processing operation, the final grade attribute was categorized according to two different grading systems, before classification. As a result, we have created two versions of both datasets. Both math and Portuguese datasets were available in both five-level and binary grading versions. Hence, we have the chance to compare the results of these versions.

In the first experiment, three algorithms [decision tree (J48), random forest, and naive Bayes] were compared on the five-level grading version and binary version of the Portuguese dataset. the best performance for the five-level grading version for this dataset was obtained with an accuracy rate of 73.50% with the random forest algorithm. However, this accuracy rate was increased with the binary grading version of this dataset. In the dataset, where the final grade is categorized in binary form (passing or failing), the accuracy rate was increased to 93.07%.

| **Algorithm** | **Five-level grading** | **Binary grading (P/F)** |
| --- | --- | --- |
| Decision tree (J48) | 67.80% | 91.37% |
| Random forest | **73.50%** | **93.07%** |
| Naive Bayes | 68.26% | 88.44% |

Classification accuracy rates for the Portuguese lesson dataset.

(accuracy values, bold – best model).

The performances of three classification algorithms on mathematics datasets (five-level and binary label dataset versions) are shown in Table 5. The best results for the five-level grading version were obtained with the decision tree (J48) algorithm with an accuracy rate of 73.42%. The best accuracy rate of 91.39% for the binary dataset version was obtained with the random forest ensemble method.

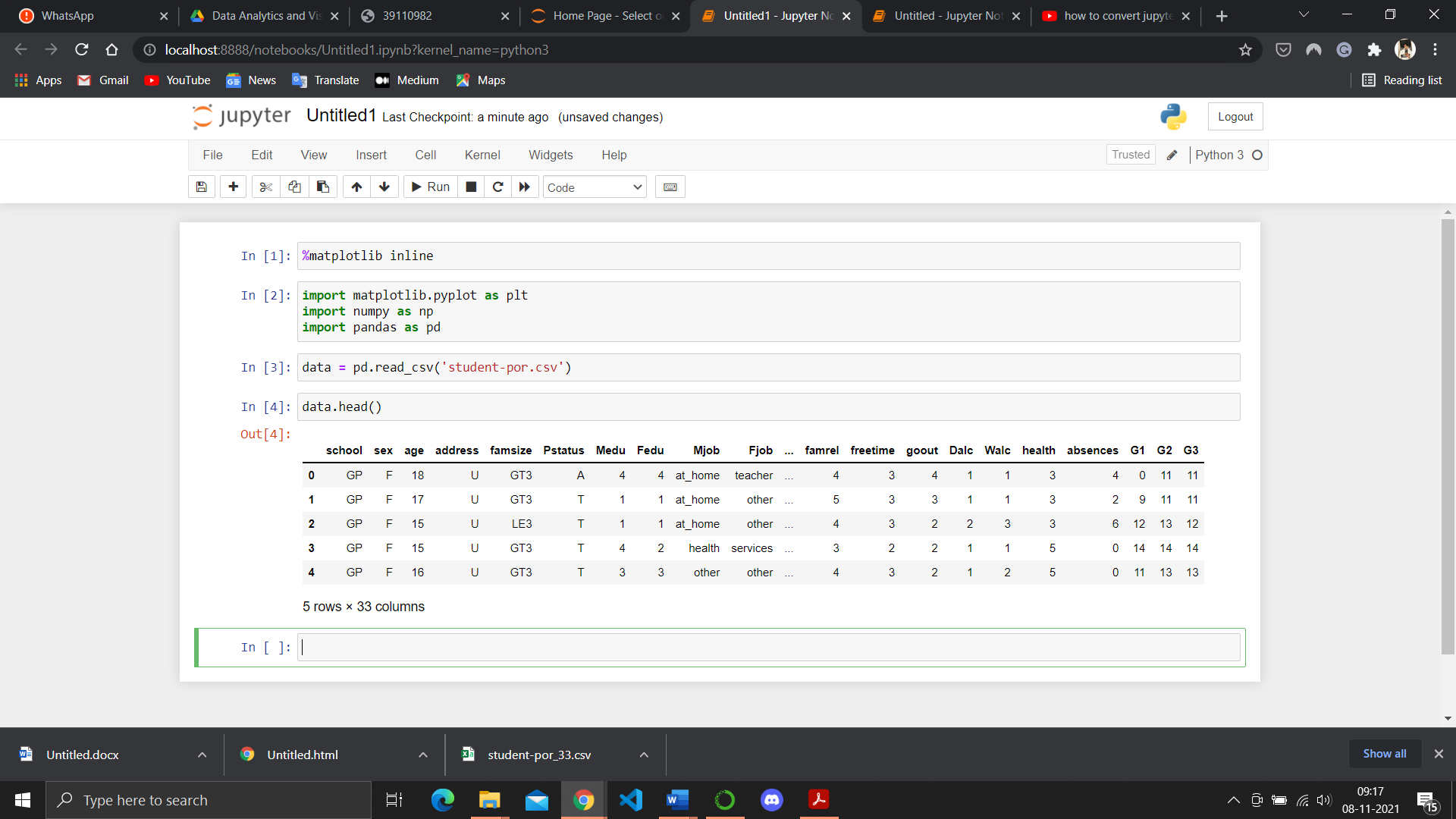
| **Mathematics** | **Five-level grading** | **Binary grading (P/F)** |
| --- | --- | --- |
| Decision tree (J48) | **73.42%** | 89.11% |
| Random forest | 71.14% | **91.39%** |
| Naive Bayes | 70.38% | 86.33% |

***3.8 Reading data:***

Generally, we use a dataset in the form of a CSV file, for reading this CSV file we will use the panda’s library, let’s see:

data = pd.read\_csv('student-por.csv ')

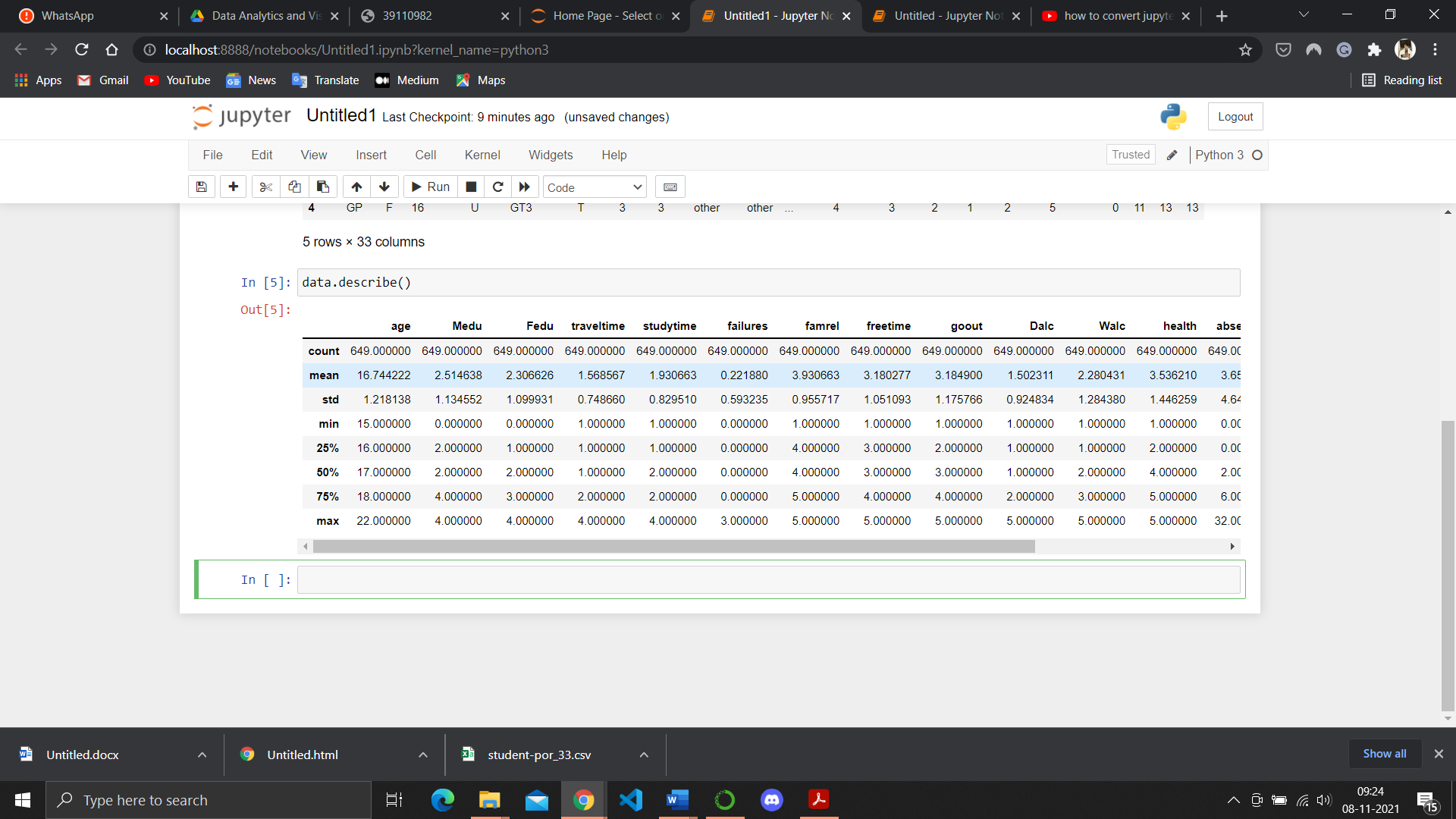
data.head()



## *3.8.1Study Dataset*

describe() -method to calculate the various calculation of each data point in the feature.

data.describe()



*3.8.2 Spliting the data:*

Now we use the scikit-learn module train\_test\_split, which is used for splitting the training and testing parts.

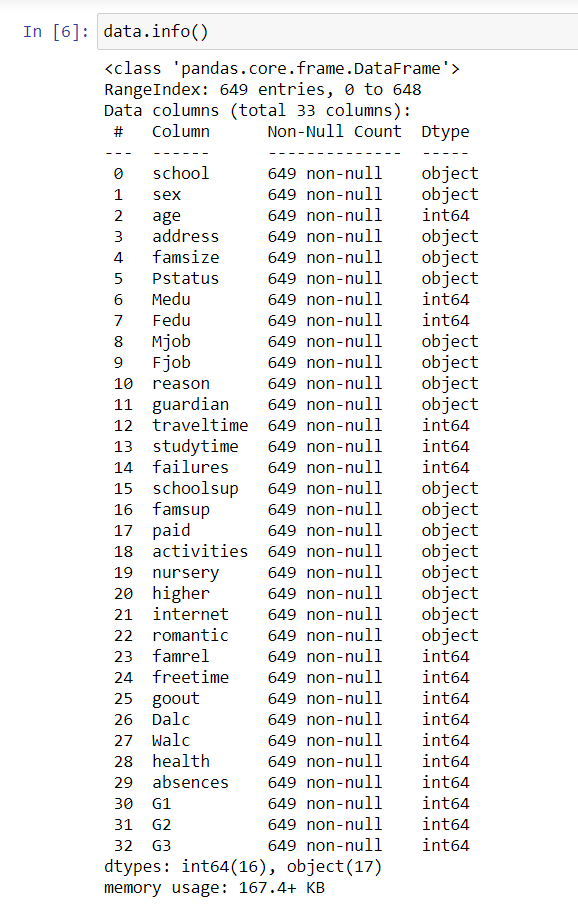
# Split the dataset in training set and test set

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size = 0.2)

After reading the dataset we have to extract information from the data, for that we use the certain function:

data.info()



***Figure 3.8 Student dataset info***

Here we notice that the count of null values in each feature and see what is the data type of features present in the dataset.

### *3.8.3 Linear regression:*

Linear Regression (LR) Linear regression is an old method to evaluate the correlation between two or more features. After deciding the relationship between features (input) and target (output), the learning process will be run to minimize the loss function value (like Mean Squared Error). The parameters that minimize the loss function are exactly the optimal parameters for the regression. Due to its simplicity, the accuracy of this model is not high. The general form of a multiple linear regression model is given in Equation (1) below :

yˆ = a0 + m ∑ j = 1 ajXj (1)

where yˆ is the predicted result, Xj is the features (input) of the dataset, and a0, a1, . . . , am are the parameters to train. In this study, this model is employed to fit multiple linear equations that relate the compressive strength and the given features.

According to previous research, the relation between features and compressive strength is complex and nonlinear. Therefore, in order to improve the prediction accuracy of the LR model, polynomial features are created using original features with different polynomial degrees.

from sklearn.linear\_model import LinearRegression

ml=LinearRegression()

ml.fit(X\_train,y\_train)

**CHAPTER 4**

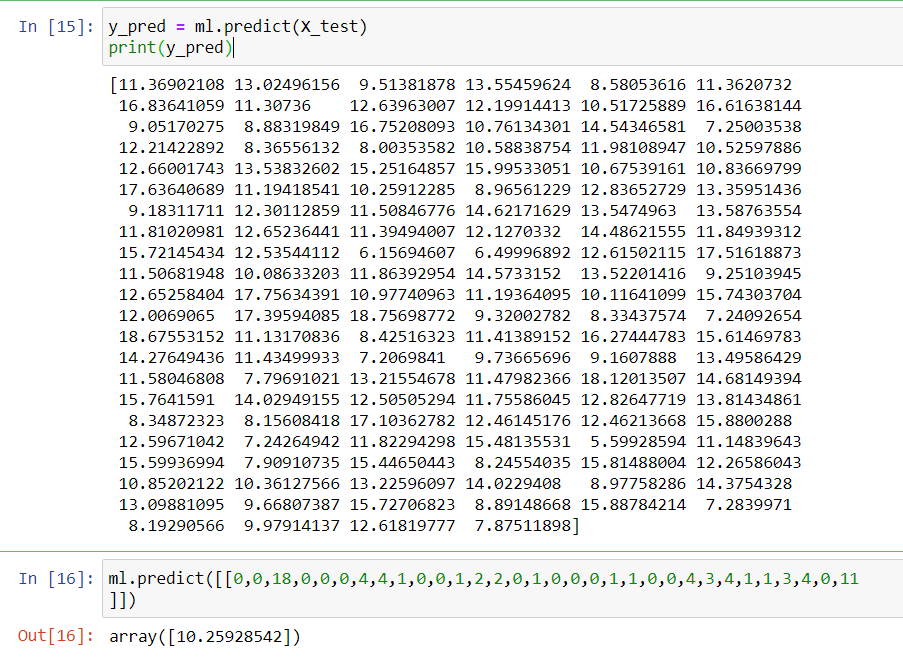
**RESULTS AND DISCUSSION****, PERFORMANCE ANALYSIS**

*4.1. PREDICT THE TEST SET RESULTS:*

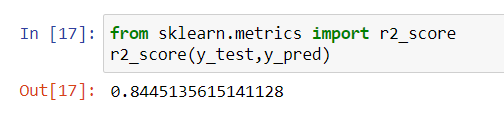
Now, we take a compression between the predicted values of the dependent variable and the original values of variable

y\_pred = ml.predict(X\_test)

print(y\_pred)



***4.2 Evaluate the model:***

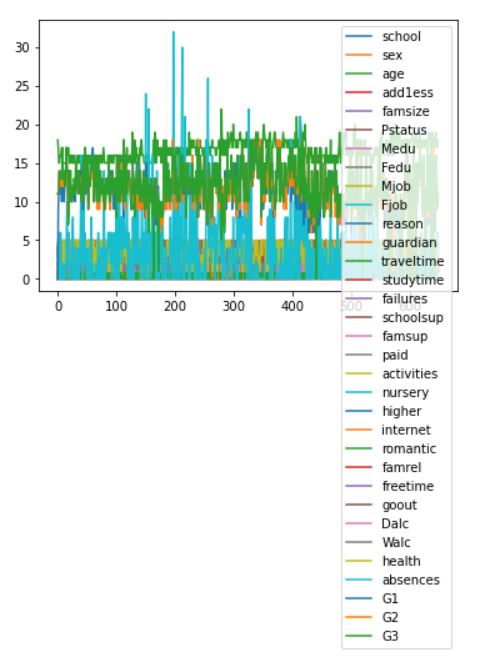


***4.3 Plot the Result:***

plt.show()

data.plot()

This command can also be used within a Notebook - for instance, to display multiple figures if several are created by a single cell.

******

import matplotlib.pyplot as plt

plt.scatter(y\_test,y\_pred)

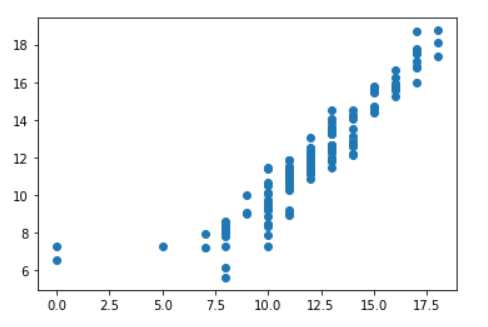
plt. figure(figsize=(15,10))

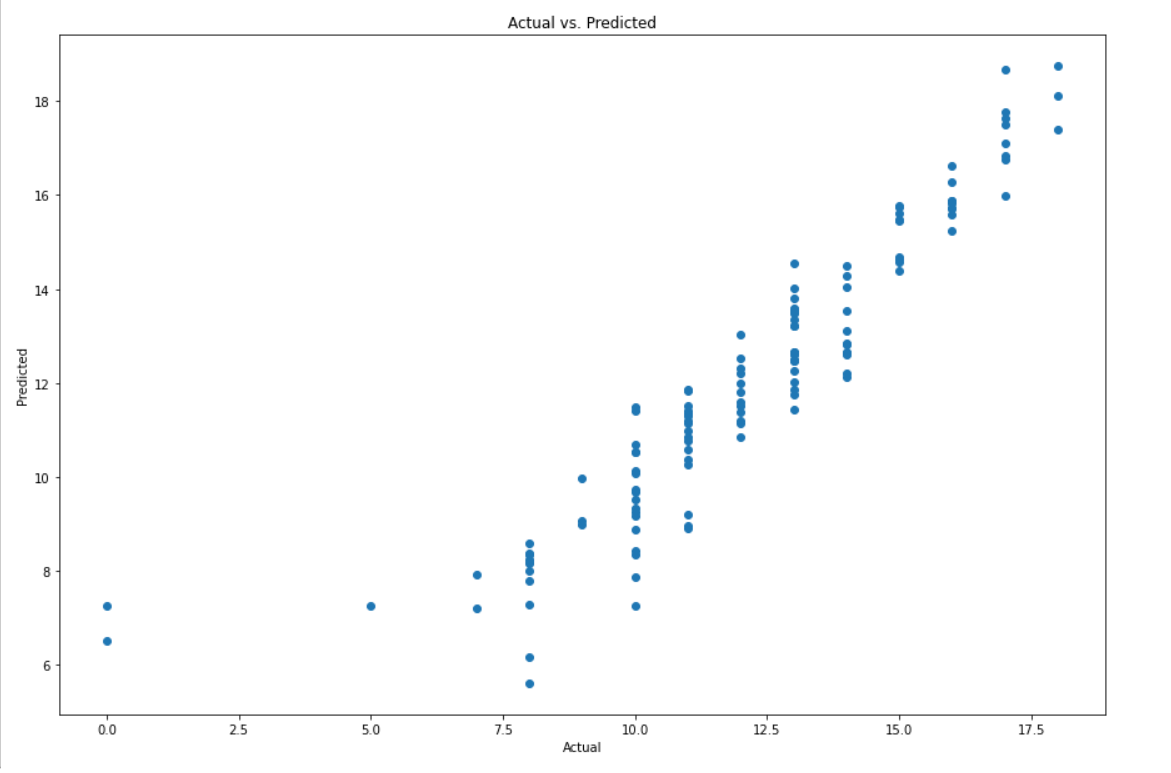
plt.scatter(y\_test,y\_pred)

plt.xlabel('Actual')

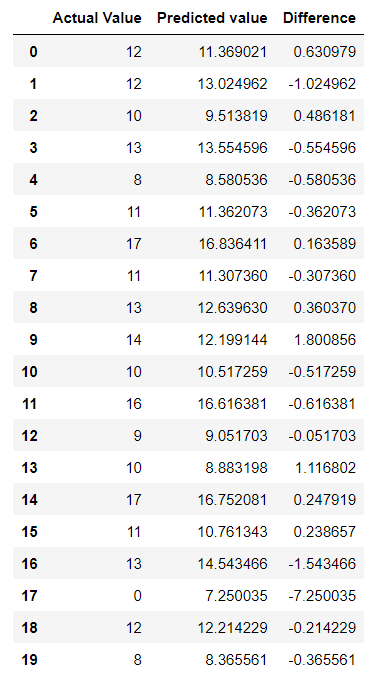
plt.ylabel('Predicted')

plt.title('Actual vs. Predicted')



****

***4.4 PREDICTED VALUES:***

******

**CHAPTER 5**

**CONCLUSION AND FUTURE WORKS**

|  |
| --- |
|  |

This paper proposes the application of data mining techniques to predict the final grades of students based on their historical data. Three well-known classification techniques (decision tree, random forest, and Naive Bayes) were compared in terms of accuracy rates. The wrapper feature subset selection method was used to improve the classification performance. Pre-processing operations on the dataset, categorizing the final grade field into five and two groups, increased the percentage of accurate estimates in the classification. The wrapper attribute selection method in all algorithms has led to a noticeable increase in inaccuracy rate. Overall, better accuracy rates were achieved with the binary class method for both mathematics and Portuguese datasets.

In the future, different feature selection methods can be used. In addition, different classification algorithms can also be utilized on the datasets.

**REFERENCES:**

Breiman L., 2001. Random Forests. Machine Learning,

45, no. 1, 5–32.

Breiman L.; Friedman J.; Ohlsen R.; and Stone C., 1984.

Classification and Regression Trees. Wadsworth,

Monterey, CA.

Cortez P., In press. RMiner: Data Mining with Neural

Networks and Support Vector Machines using R. In

R. Rajesh (Ed.), Introduction to Advanced Scientific

Softwares and Toolboxes.

Eurostat, 2007. Early school-leavers.

http://epp.eurostat.ec.europa.eu/.

Flexer A., 1996. Statistical Evaluation of Neural

Networks Experiments: Minimum Requirements and

Current Practice. In Proceedings of the 13th Euro-

pean Meeting on Cybernetics and Systems Research.

Vienna, Austria, vol. 2, 1005–1008.

Hastie T.; Tibshirani R.; and Friedman J., 2001. The

Elements of Statistical Learning: Data Mining, Infer-

ence, and Prediction. Springer-Verlag, NY, USA.

Kotsiantis S.; Pierrakeas C.; and Pintelas P., 2004. Pre-

dicting Students’ Performance in Distance Learning

Using Machine Learning Techniques. Applied Artifi-

cial Intelligence (AAI), 18, no. 5, 411–426.

Luan J., 2002. Data Mining and Its Applications in

Higher Education. New Directions for Institutional

Research, 113, 17–36.

Ma Y.; Liu B.; Wong C.; Yu P.; and Lee S., 2000.

Targeting the right students using data mining. In

Proc. of 6th ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining. Boston,

USA, 457–464.

Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and

Punch W., 2003. Predicting student performance: an

application of data mining methods with an educa-

tional web-based system. In Proc. of IEEE Frontiers

in Education. Colorado, USA, 13–18.

Pardos Z.; Heffernan N.; Anderson B.; and Heffernan C.,

2006. Using Fine-Grained Skill Models to Fit Student

Performance with Bayesian Networks. In Proc. of 8th

Int. Conf. on Intelligent Tutoring Systems. Taiwan.

Pritchard M. and Wilson S., 2003. Using Emotional and

Social Factors To Predict Student Success. Journal of

College Student Development, 44, no. 1, 18–28.

R Development Core Team, 2006. R: A language and

environment for statistical computing. R Foundation

for Statistical Computing, Vienna, Austria, ISBN 3-

900051-00-3, http://www.R-project.org.

Turban E.; Sharda R.; Aronson J.; and King D.,

2007. Business Intelligence, A Managerial Approach.

Prentice-Hall.

Witten I. and Frank E., 2005. Data Mining: Practical

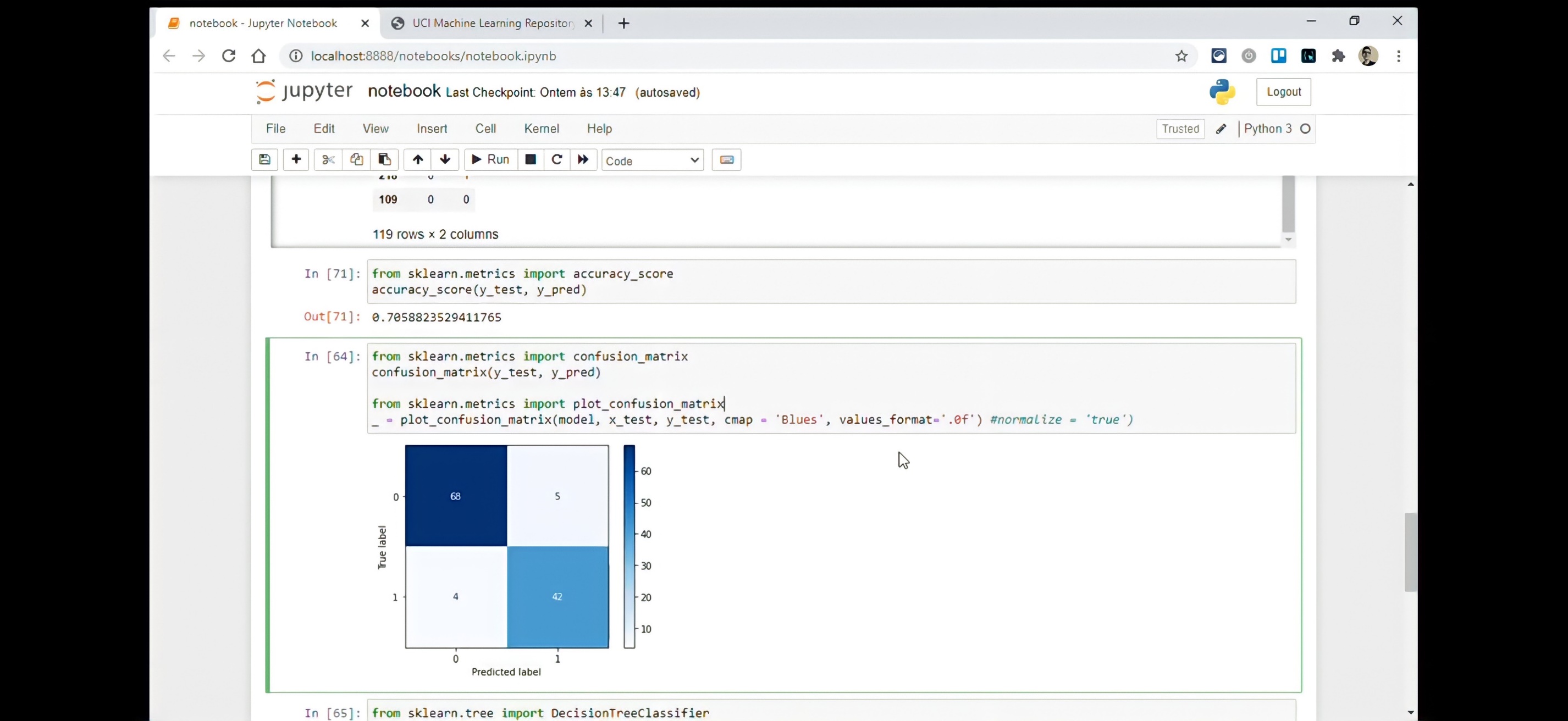
Machine Learning Tools and Techniques with Java

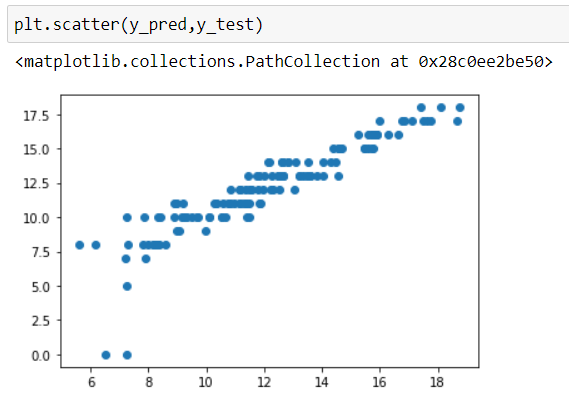
Implementations. Morgan Kaufmann, San Francisco,

CA.

**APPENDIX**

**A.SCREENSHOTS:**

****



1. **SOURCE CODE:**

**# Import Libraries**

%matplotlib inline

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

**# Import DataSet**

data = pd.read\_csv('student\_por.csv')

data.head()

data.shape

**# Define x and y**

x=data.drop(['G3'],axis=1).values

y=data['G3'].values

print(x)

print(y)

data.describe()

**# Split the dataset in training set and test set**

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size = 0.2)

**# Train the model on the training set**

from sklearn.linear\_model import LinearRegression

ml=LinearRegression()

ml.fit(X\_train,y\_train)

**# Predict the test set result**

y\_pred = ml.predict(X\_test)

print(y\_pred)

ml.predict([[0,0,18,0,0,0,4,4,1,0,0,1,2,2,0,1,0,0,0,1,1,0,0,4,3,4,1,1,3,4,0,11

]])

**# Evaluate the Model**

from sklearn.metrics import r2\_score

r2\_score(y\_test,y\_pred)

**# Plot the result**

import matplotlib.pyplot as plt

plt.scatter(y\_test,y\_pred)

plt. figure(figsize=(15,10))

plt.scatter(y\_test,y\_pred)

plt.xlabel('Actual')

plt.ylabel('Predicted')

plt.title('Actual vs. Predicted')

data.plt()

plt.scatter(y\_pred,y\_test)

**# Predicted Values**

pred\_y\_data\_df=pd.DataFrame({'Actual Value':y\_test, 'Predicted value':y\_pred, 'Difference': y\_test-y\_pred})

pred\_y\_data\_df[0:20]