Data Mining Mini Project

A FOURTH YEAR COMPUTER MINI-PROJECT

SUBMITTED BY

1. BIKASH BANJARA

2. SABIN SHRESTHA

SCHOOL OF SCIENCE

KATHMANDU UNIVERSITY

DHULIKHEL, NEPAL

DECEMBER 2021

CONTENTS

ABSTRACT

Classification is the process where we try to classify or predict the target variables. For classification we have used KNN and Decision tree . Regression is the process where we try to find out the relations between independent variables and dependent variables. We have used linear regression for this purpose. Using classification and regression we have predicted different class values of our data set .I.e. Student performance dataset and we have interpreted them.

CHAPTER 1

INTRODUCTION

Data mining is the process of discovering insightful,interesting and novel patterns and knowledge from a large amount of data.

Data preprocessing is a technique done before datamining. It includes better understanding of data,detection of mistakes,Accessing the direction and rough size of relationship between explanatory and outcome variables.

Regression: regression is an approach for modeling the relationship between a scalar dependent variable Y and one more explanatory variable (or independent variable) denoted X.The earliest form of the regression was the method of the least square which was published by Adrien Marie Legendre (1752-1833) in 1805 and later gauss developed it in 1809. Regression analysis is the type of the statistical analysis that enables 3 things:

1. **Description**: Relationship between the dependent variable and independent variable can be statistically described by means of regression analysis. It is also known as situation regression.
2. **Estimation**: The values of the dependent variable can be estimated from the observed value of the independent variable.
3. **Prognostication**: Risk factors that influence the outcome can be identified and individual prognosis can be determined.

Classification:

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output. In this project, we have used the KNN and Decision tree for classification.

CHAPTER 2

METHODOLOGY

2.1 Data Mining Task :

In this project, the assigned task was to do regression and classification. And we are asked to do the necessary things like EDA, data cleaning etc.and finally to visualize and interpret the data.

2.2 Data Mining Steps

Firstly we have chosen the data from the UCI ML repository. Then we have done preprocessing tasks like data cleaning and data reduction to perform prediction for G3 based on various parameters such as study time, health status, absence, etc.

2.3 Data Mining Algorithm

To do classification, we have chosen KNN and Decision tree. And for regression we have chosen linear regression algorithms. We have implemented all these algorithms in orange.

CHAPTER 3

Result and Discussion

As our assigned task is to perform classification and regression. We have used a student performance dataset and we have chosen it from the uci machine learning repository.It is a multivariate Data, it has 33 attributes,649 instances.it has 2 data of 2 different subjects but we have taken data of only one subject.i.e we have taken data of math.

Data Set Information:

This data approaches 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).

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

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)

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)

these grades are related with the course subject, Math or Portuguese:

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)

In classification we have tried to predict whether the student goes for higher study or not and for this purpose, various parameters taken are age, failure, absence, health, G1, G2, G3, Studytime, Fjob as independent variables. And the result obtained is shown below.

Predictions Sun Dec 19 21, 13:27:57

Info

11 instances
1 model
Showing probabilities for: no, yes

Data & Predictions

	kNN	higher	age	failures	absences	health	G2	G1	studytime	G3	Fjob
1	0.00 : 1.00 → yes	yes	16	0	4	2	6	7	2	6	other
2	0.00 : 1.00 → yes	yes	15	0	2	5	15	15	2	14	other
3	0.00 : 1.00 → yes	yes	17	0	1	5	14	12	3	15	services
4	0.00 : 1.00 → yes	yes	17	0	2	3	11	12	2	12	services
5	0.00 : 1.00 → yes	yes	17	0	4	2	12	12	1	13	other
6	0.17 : 0.83 → yes	no	18	1	0	2	7	7	2	0	services
7	0.00 : 1.00 → yes	yes	15	1	2	3	7	7	2	7	teacher
8	0.00 : 1.00 → yes	no	17	0	2	2	10	10	1	10	services
9	0.00 : 1.00 → yes	yes	15	0	6	1	13	10	1	13	other
10	0.00 : 1.00 → yes	yes	16	0	2	5	13	13	2	11	other
11	0.00 : 1.00 → yes	yes	16	0	8	3	9	9	1	10	other

Scores

Model	AUC	CA	F1	Precision	Recall
kNN	0.750	0.818	0.736	0.669	0.818

Fig: classification using KNN

Similarly we have used decision tree for classification and its result is shown below:

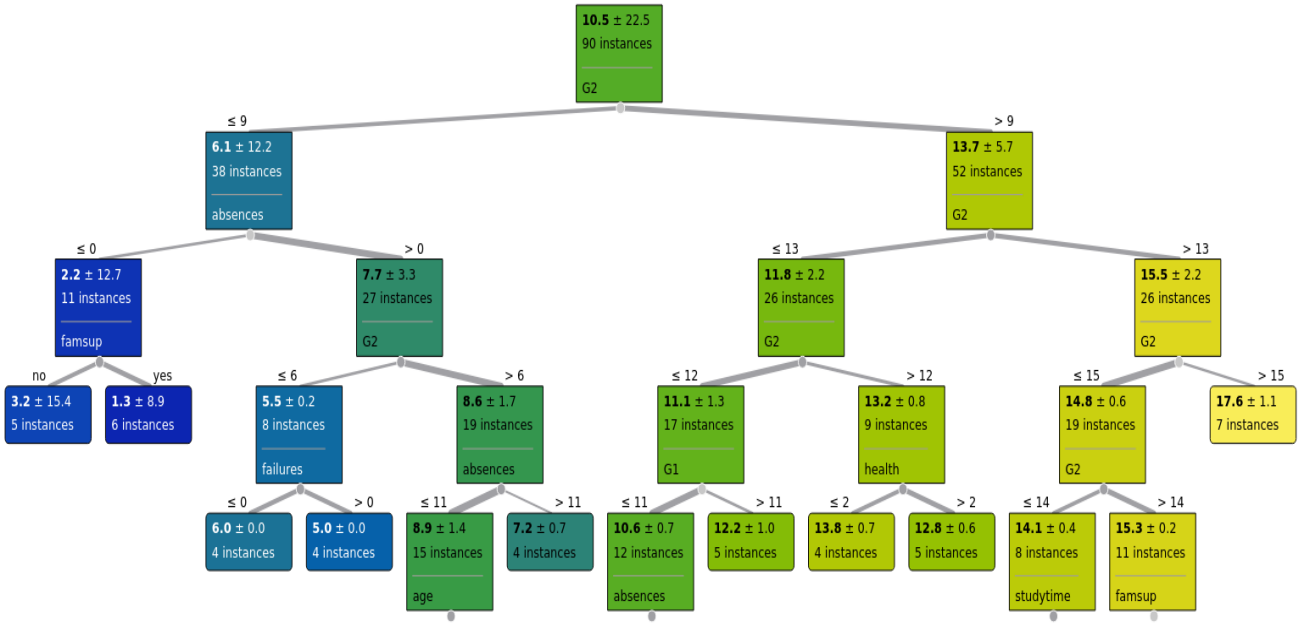

Fig: Classification With Decision Tree

Considering 20 as base marks for every internal, we can say 10 is average marks and the figure above distributed list of students on the basis of marks where right tree shows the list of students who are sure candidate for boards and can score good whereas student distributed on left tree are those of average and poor students who might undergo further process for evaluation purpose to appear for boards.

And for the regression task we have used linear regression and we have predicted G3 . We have used age,study time,failure,absence,health,G2,G1 as the independent variable and G3 is dependent on all these variables.and the result is displayed below:

Predictions Sun Dec 19 21, 13:58:16

Info

75 instances
1 model

Data & Predictions

	Linear Regression	G3	age	studytime	failures	absences	health	G2	G1
1	14	13	17	2	0	23	5	13	13
2	9	9	17	2	0	12	3	9	11
3	11	11	17	3	0	3	3	11	11
4	14	15	17	3	0	1	5	14	12
5	15	15	17	3	0	0	2	15	16
6	11	11	18	3	0	3	3	12	9
7	15	16	17	1	0	3	5	15	14
8	10	10	17	1	0	8	4	10	11
9	9	9	17	3	0	7	4	9	10
10	14	14	17	3	0	4	4	14	14
11	7	8	18	4	0	2	5	8	9
12	14	14	17	3	0	7	5	14	12
13	-1	0	18	2	0	0	4	0	7
14	7	0	18	2	0	0	2	8	8
15	8	0	18	4	0	0	4	9	10
16	16	15	17	3	0	16	5	15	16
17	13	13	19	3	1	12	5	13	14

58	8	10	18	2	0	4	2	9	8
59	15	15	18	2	0	0	1	15	15
60	10	10	17	2	0	17	1	10	10
61	14	14	18	2	0	4	2	14	15
62	6	7	18	1	0	5	5	6	7
63	11	10	17	2	0	2	3	11	11
64	4	0	19	1	1	0	5	5	6
65	4	5	18	1	1	14	3	5	6
66	8	10	18	3	0	2	4	9	10
67	4	6	18	1	0	7	5	5	6
68	4	0	19	3	1	0	5	5	7
69	8	8	18	2	0	0	1	9	7
70	4	0	18	2	1	0	5	5	6
71	8	9	16	2	2	11	4	9	9
72	16	16	17	1	0	3	2	16	14
73	6	7	21	1	3	3	3	8	10
74	12	10	18	1	0	0	5	12	11
75	8	9	19	1	0	5	5	9	8

Scores

Model	MSE	RMSE	MAE	R2
Linear Regression	5.850	2.419	1.376	0.746

Fig: Linear Regression

CHAPTER 4

Conclusion:

For this mini project we have taken a student performance data set and we have done the data mining tasks like data cleaning, data reduction, data preprocessing and finally we have used the classification and regression algorithm which was our motive in this project. Similarly, we have understood how these algorithms work and where they should be used.finally we also learned how to use Data mining tools like weka and orange.

REFERENCES

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[2] data from UCI machine learning repository.