Predicting Student Performance Using Data Mining Techniques in Secondary School.

Student Performance Analysis

1. Abstract

This study tries to find out which factors affect the academic performance of a student. Data was collected for students in two subjects – Math and Portuguese. This included directly study related data such as hours spent studying, past failures and also data from a student's domestic life such as mother's job, parents' cohabitation status. The grades for these students in three tests for these two subjects were also collected. Using this data, we tried to find out which were the most useful predictors for grades. Using this information, teachers and schools may be able to intervene if they know students are heading towards bad grades and offer assistance in areas it is needed for the students. We found that while study-related variables such as hours spent studying and past failures are significant in predicting grades, factors such as a child having school support and how much time they spend going out also has a relationship with grades.

2. Introduction

Classification vs. Regression

Question: Your goal for this project is to identify students who might need early intervention before they fail to graduate. Which type of supervised learning problem is this, classification or regression? Why?
Answer: This supervised learning belongs to classification as by the definition of classification "given a known relationship, identify the class that the data belongs to". So the following project distinguishes for whether the student will graduate or not.

Education is an important element of society. Regression techniques, which allow high-level extraction of knowledge from raw data, can offer useful insights for the education domain. There are several interesting research questions in the area of school and college education that can be studied using regression techniques. How is Student final grade dependent upon the frequency of going out? How is the student performance impacted by relationships or alcohol consumption? Does gender affect student performance? What are the factors that affect student grades? This project will focus on some of these questions.

Modeling student performance is an important tool for both educators and students since it can help get a better understanding of this phenomenon and ultimately improve it. For instance, school professionals could perform corrective measures for weak students (e.g. remedial classes). They could detect if a student was heading towards a bad grade on a class and help them get on track to a better grade. Further, there are two reasons supporting the choice of such a project: (a) there are multiple sources of data available (e.g. traditional databases, online web pages), and (b) there are diverse interest groups (e.g. students, teachers, administrators or alumni) interested in the insights offered by such a project.

We have classified these students into three categories, **"good", "fair",** and **"poor",** according to their final exam performance. Then we analyzed a few features that have a significant influence on students' final performance, including Romantic Status, Alcohol Consumption, Parents Education Level, Frequency Of Going Out, Desire Of Higher Education and Living Area. Finally, leveraging available features, we have created various machine learning models to predict students' final performance classification and have compared models performance based on one-out sample accuracy score.

Dataset available at: http://archive.ics.uci.edu/ml/datasets/Student+Performance#

3. 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).

Number of Instances:	**649**

4. 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 - 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)
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 to the course subject, Math or Portuguese:

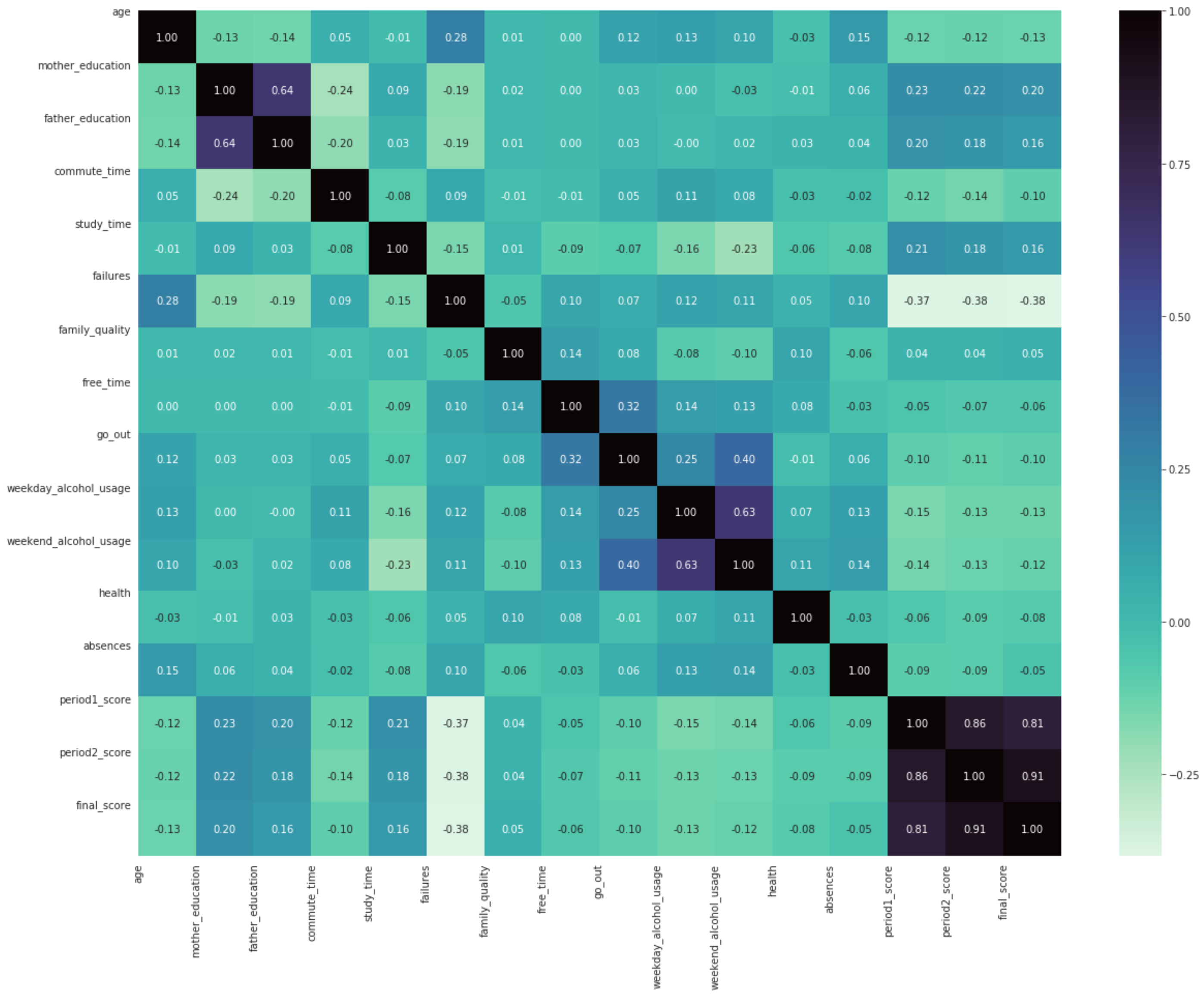
31. **G1 -** first-period grade
    (numeric: from 0 to 20)

31. **G2 -** second-period grade
    (numeric: from 0 to 20)

32. **G3 -** final grade
    (numeric: from 0 to 20, output target)

5. Exploratory Data Analysis

Correlation between variables through a correlation heatmap:



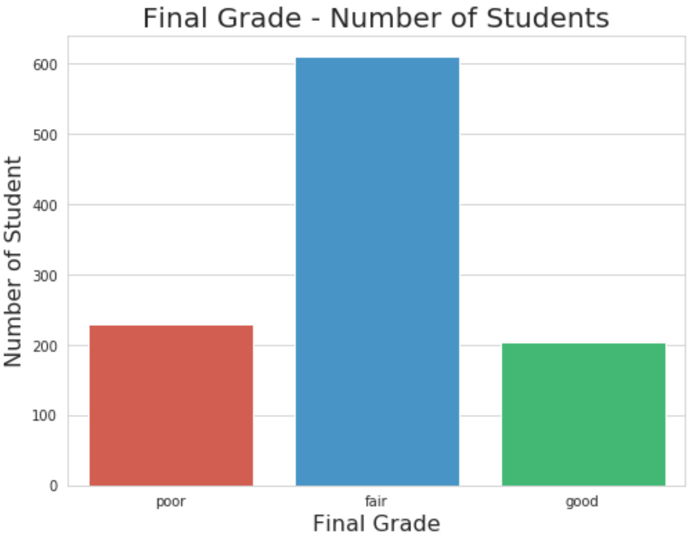
	age	mother_education	father_education	commute_time	study_time	failures	family_quality	free_time	go_out	weekday_alcohol_usage	weekend_alcohol_usage	health	absences	period1_score	period2_score	final_score
age	1.00	-0.13	-0.14	0.05	-0.01	0.28	0.01	0.00	0.12	0.13	0.10	-0.03	0.15	-0.12	-0.12	-0.13
mother_education	-0.13	1.00	0.64	-0.24	0.09	-0.19	0.02	0.00	0.03	0.00	-0.03	-0.01	0.06	0.23	0.22	0.20
father_education	-0.14	0.64	1.00	-0.20	0.03	-0.19	0.01	0.00	0.03	-0.00	0.02	0.03	0.04	0.20	0.18	0.16
commute_time	0.05	-0.24	-0.20	1.00	-0.08	0.09	-0.01	-0.01	0.05	0.11	0.08	-0.03	-0.02	-0.12	-0.14	-0.10
study_time	-0.01	0.09	0.03	-0.08	1.00	-0.15	0.01	-0.09	-0.07	-0.16	-0.23	-0.06	-0.08	0.21	0.18	0.16
failures	0.28	-0.19	-0.19	0.09	-0.15	1.00	-0.05	0.10	0.07	0.12	0.11	0.05	0.10	-0.37	-0.38	-0.38
family_quality	0.01	0.02	0.01	-0.01	0.01	-0.05	1.00	0.14	0.08	-0.08	-0.10	0.10	-0.06	0.04	0.04	0.05
free_time	0.00	0.00	0.00	-0.01	-0.09	0.10	0.14	1.00	0.32	0.14	0.13	0.08	-0.03	-0.05	-0.07	-0.06
go_out	0.12	0.03	0.03	0.05	-0.07	0.07	0.08	0.32	1.00	0.25	0.40	-0.01	0.06	-0.10	-0.11	-0.10
weekday_alcohol_usage	0.13	0.00	-0.00	0.11	-0.16	0.12	-0.08	0.14	0.25	1.00	0.63	0.07	0.13	-0.15	-0.13	-0.13
weekend_alcohol_usage	0.10	-0.03	0.02	0.08	-0.23	0.11	-0.10	0.13	0.40	0.63	1.00	0.11	0.14	-0.14	-0.13	-0.12
health	-0.03	-0.01	0.03	-0.03	-0.06	0.05	0.10	0.08	-0.01	0.07	0.11	1.00	-0.03	-0.06	-0.09	-0.08
absences	0.15	0.06	0.04	-0.02	-0.08	0.10	-0.06	-0.03	0.06	0.13	0.14	-0.03	1.00	-0.09	-0.09	-0.05
period1_score	-0.12	0.23	0.20	-0.12	0.21	-0.37	0.04	-0.05	-0.10	-0.15	-0.14	-0.06	-0.09	1.00	0.86	0.81
period2_score	-0.12	0.22	0.18	-0.14	0.18	-0.38	0.04	-0.07	-0.11	-0.13	-0.13	-0.09	-0.09	0.86	1.00	0.91
final_score	-0.13	0.20	0.16	-0.10	0.16	-0.38	0.05	-0.06	-0.10	-0.13	-0.12	-0.08	-0.05	0.81	0.91	1.00

Final Grade Distribution based on converting **final_score** to a categorical variable -

Good: 15~20

Fair: 10~14

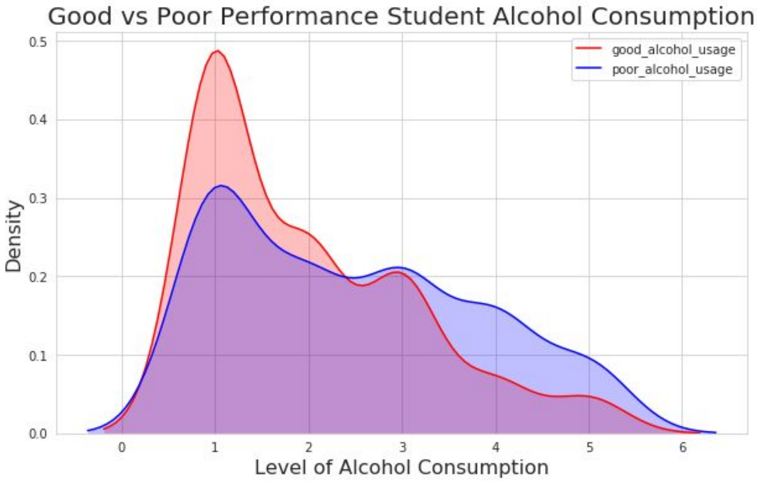
Poor: 0~9

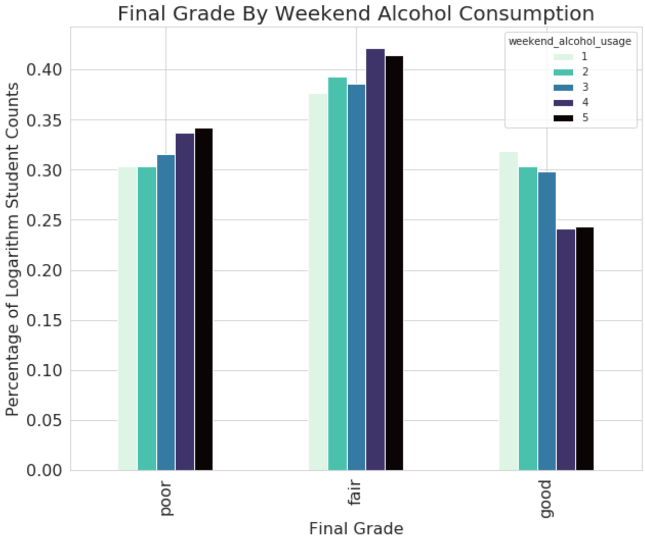


6. Exploring the relationships of features on the Final Grade

Final Grade by Alcohol Consumption

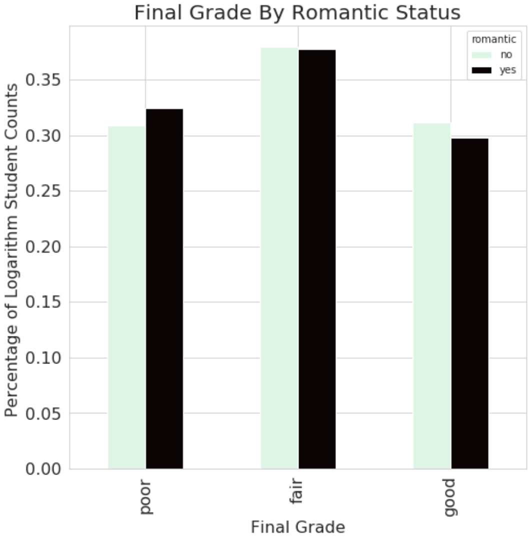
See how alcohol consumption influences students' final grade.





Weekend alcohol consumption has a significant correlation with the final grade.

Final Grade by Romantic Status



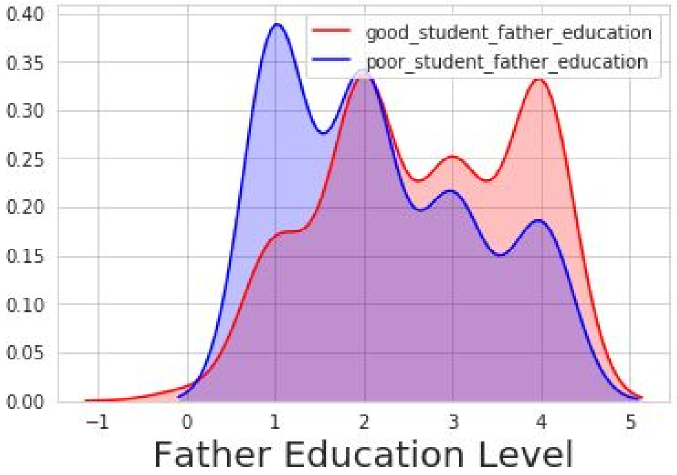
This depicts how students in a romantic relationship vs. those not in a relationship perform in their final exams.

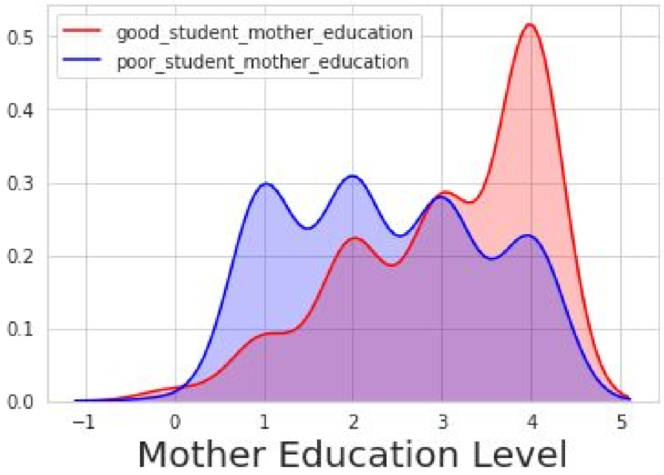
Romantic status affects the final grade.

In poor performing students, Students with relationships are more than with no relationships. The opposite is true for good performing students.

Final Grade by Parents Education Level

See how parents' education level influence student performance.





Ordinary Least Squares is a type of linear least squares method for estimating the unknown parameters in a linear regression model.

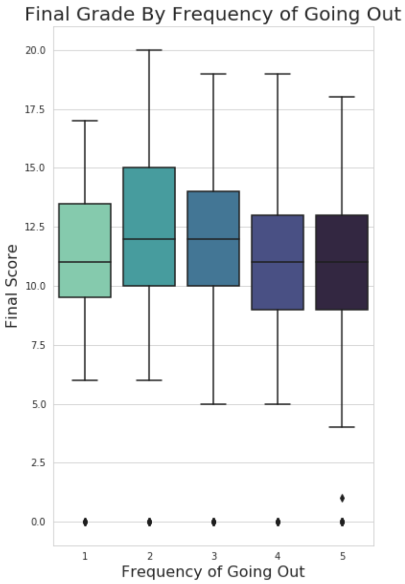
OLS tells that parents' education level has a **positive correlation** with students' final score.

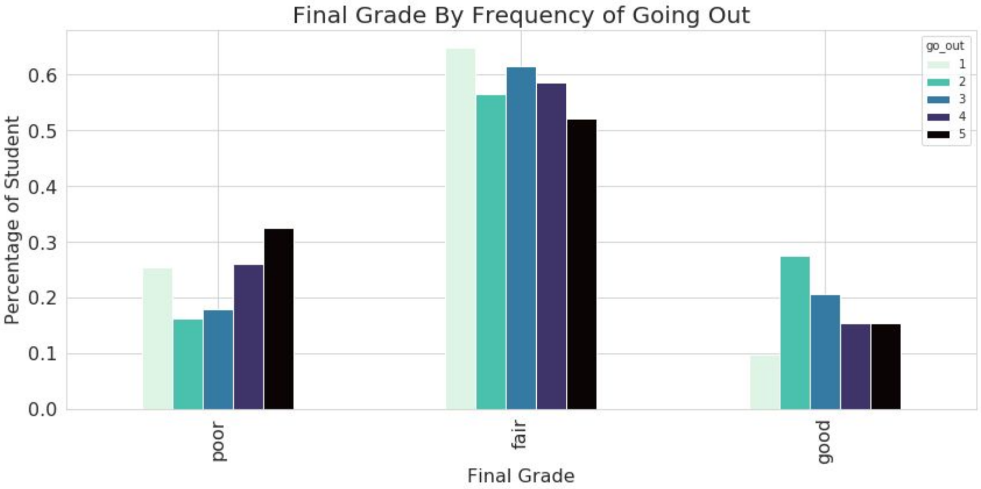
	coef	std err	t	P>|t|	[0.025	0.975]
mother_education	2.4078	0.166	14.527	0.000	2.083	2.733
father_education	1.5746	0.179	8.806	0.000	1.224	1.926

Comparatively, the mother's education level has a bigger influence than the father's education level.

Final Grade by Frequency Of Going Out

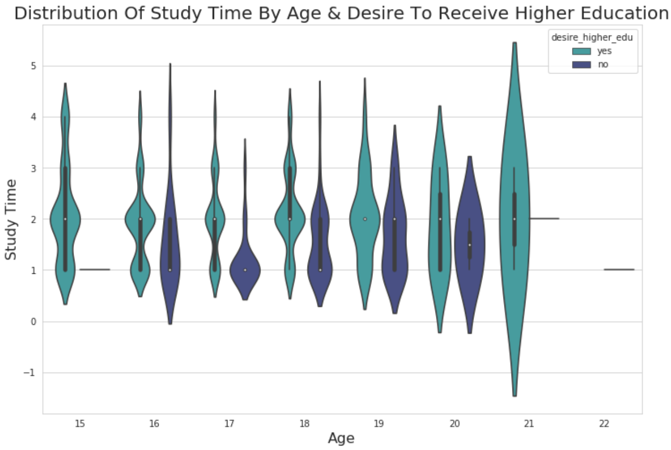
See how the frequency of going out with friend influence students' final performance.

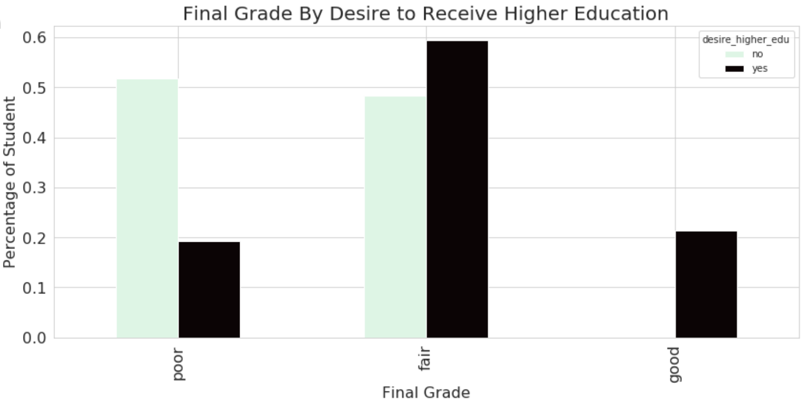




Final Grade by Desire To Go To College

See how the desire to go to college influence student final performance.

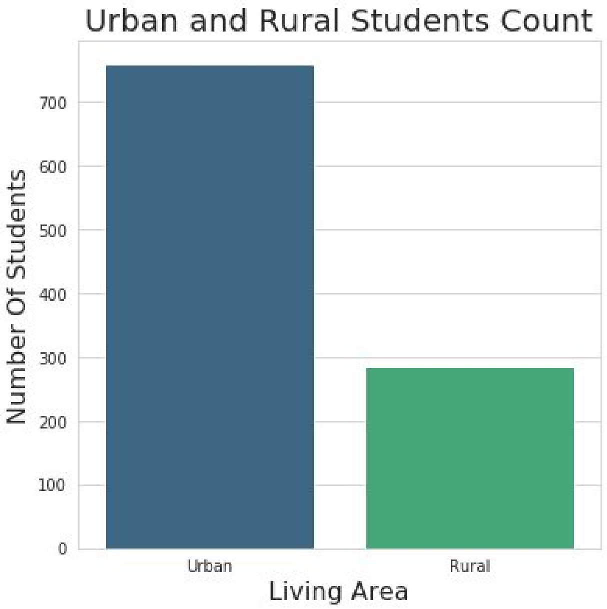


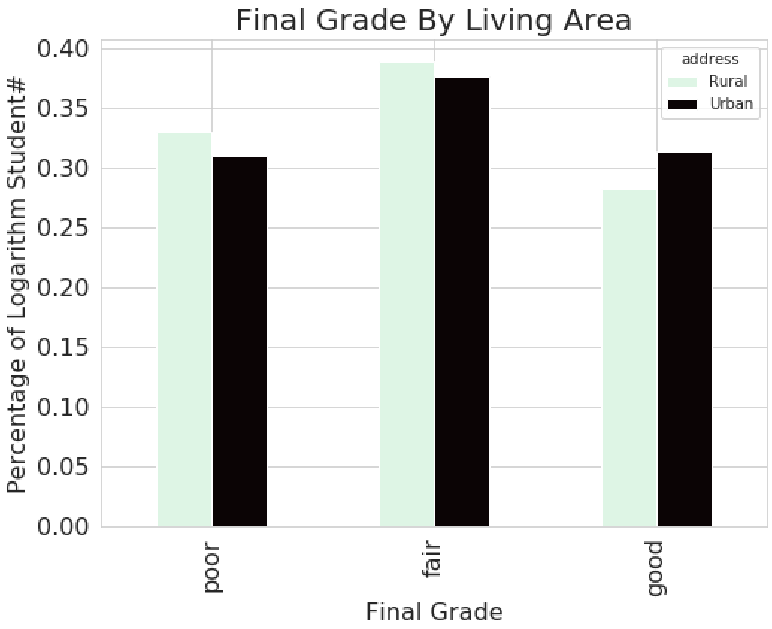


The desire of going to college has a significant correlation with students' final performance so the kids should be motivated to pursue higher education in the future.

Final Grade by Living Area

See how the final performance difference between students living in the city and those living in rural areas.

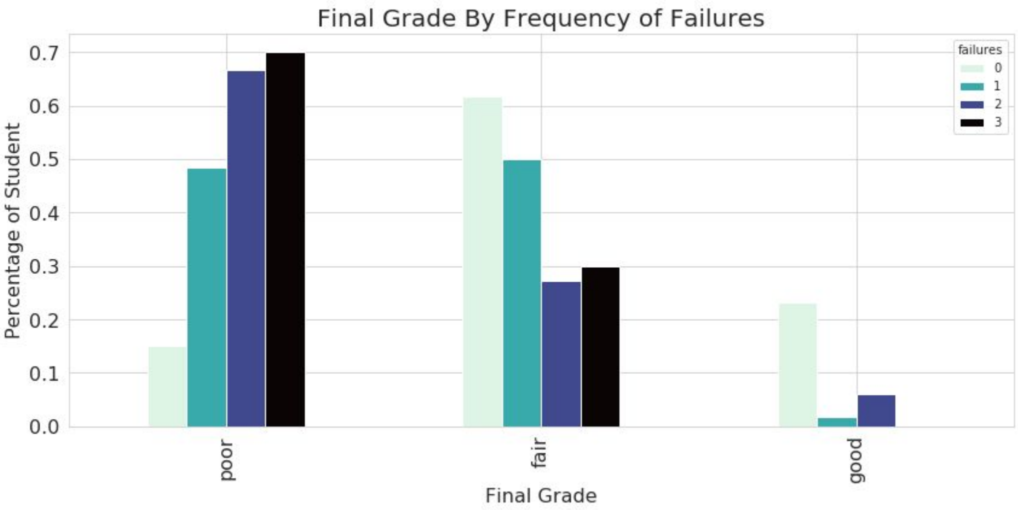




City students have an advantage over students living in rural areas.

Final Grade by Failures

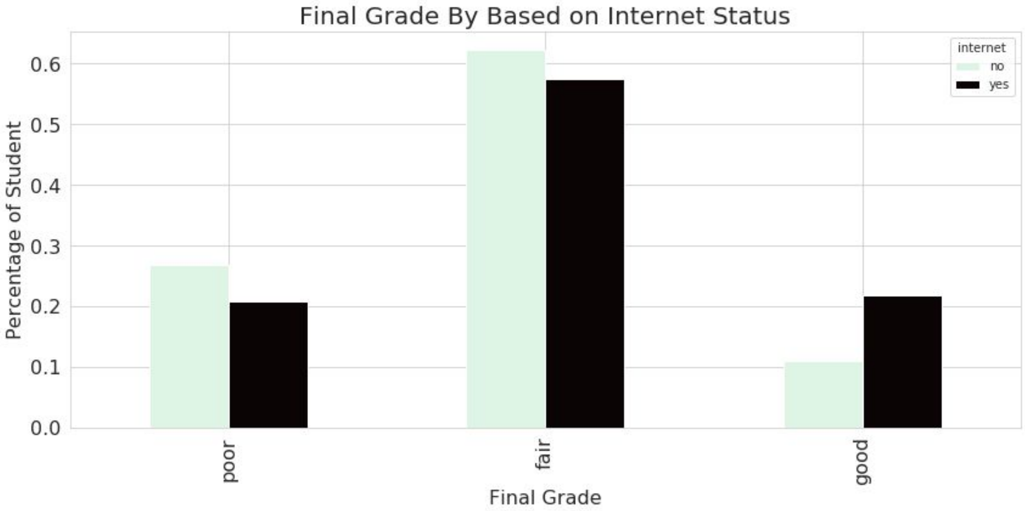
See how the final performance difference between students based on the number of times they have failed previously.



Students who have failed previously are likely to fail again.

Final Grade by Internet Status

See how the final performance difference between students based on their internet status



Internet status improves the students' final performance.

7. Classification

Aim: To identify a classification model that fits this problem well.

1) Support Vector Machines

- General Applications: SVMs have demonstrated highly competitive performance in numerous real-world applications, such as bioinformatics, text mining, face recognition, and image processing.
- Strengths: Some of its advantages are that SVM is very effective in high-dimensional spaces, and in situations when we have a non-linear separation problem. With SVM we have the possibility to apply new kernels that allows flexibility for our decision boundaries, leading to a better classification performance.
- Weaknesses: One major disadvantage of the SVM is the choice of the kernel and also it is kind of slower when compared to some other models like Decision trees or Naive .
- Why we chose this model: I chose this model as given a small number of data samples SVMs work well and applying different kernels provide flexibility for our decision boundaries thus leading to a better classification performance.

2) Decision Trees

- General Applications: Decision trees have long been important areas of application in the field of medical research and practice. Recent uses of automatic induction of decision trees can be found in diagnosis, cardiology, psychiatry, gastroenterology, for detecting microcalcifications in mammography, to analyze Sudden Infant Death(SID) syndrome and for diagnosing thyroid disorders.
- Strengths: The advantages of a decision trees are that nonlinear relationships between parameters do not affect our performance metrics and they give us a faster prediction as compared to some other models like SVMs.
- Weaknesses: Decision Trees do not work well if we have smooth boundaries. i.e they work best when we have a discontinuous piecewise constant model. If we truly have a linear target function decision trees are not the best.

- Why we chose this model: I chose this model as a decision tree provides a good performance metric for non-linear data. So, if our student data would come out to be non-linear then decision trees would be a better option.

3) Naive Bayes

- General Applications: Naive Bayes methods can be used to mark an email as spam or not spam, to check a piece of text expressing positive emotions, or negative emotions and also in face recognition softwares.
- Strengths: The advantages of naive Bayes is that an NB classifier will converge quicker than discriminative models like logistic regression, so we need less training data.
- Weaknesses: The same conditional independence assumption can be a disadvantage when we have no occurrences of a class label and a feature value together, what will give us a zero frequency-based value probability that affects any posterior probability estimate.
- Why we chose this model: I chose this model because if the NB conditional independence assumption actually holds, the Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so we'll need less training data which is in accordance with the number of data samples given to us.

4) Logistic Regression

- General Applications: You might then want to study how various factors or variables influence whether or not a person owns a foreign car. Variables you might consider are income, age, gender, marital status, children, political affiliation, and so on.
- Strengths: No distribution requirement, perform well with few categories categorical variables, compute the logistic distribution, good for few categories variables, easy to interpret, compute CI, suffer multicollinearity.
- Weaknesses: There should exist a strong correlation in the data between x and y. If there isn't a high correlation between the observed and predicted variables, then the LR model can add little predictive value.
- Why we chose this model: We use multinomial logistic regression model because it can define accurately the relationship between the group of explanatory variables and the response variable, identify the effect of each of the variables, and can predict the classification of any individual case

5) ANN

- General Applications: Image Processing, Computer Vision applications, Forecasting, Prediction problems.
- Strengths: ANNs have the ability to learn and model non-linear and complex relationships, which is really important to model real-life problems. ANNs can generalize well. Unlike many other prediction techniques, ANN does not impose any restrictions on the input variables
- Weaknesses: Extreme Hardware dependence, Unexplained behavior of the model, Determination of model structure, Unknown duration of the network operation.
- Why we chose this model: Given 58 total features surrounding the individual student, modeling this real-life problem of predicting the grade of the student would be best modeled and learned by the ANN. We wanted to check how the network learns this problem (How efficiently).

6) Random Forest

- General Applications: The banking sector consists of most users. There are many loyal customers and also fraud customers. To identify the great combination in the medicines and to determine whether the customer is a loyal or fraud random forest can be used. When you want to know the behavior of the stock market, with the help of this algorithm, the behavior of the stock market can be analyzed.
- Strengths: It provides higher accuracy. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, it won't allow overfitting trees in the model. It has the power to handle a large data set with higher dimensionality
- Weaknesses: Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data and that they may over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon your own data set.
- Why we chose this model: I chose this model because there are several criteria for predicting a student's performance and random tree classifier is best suitable classifier to handle large datasets with higher dimensionality with good accuracy.

Use Students' Information To Predict Their Final Grade:

Decision Tree Classification

```
Decision Tree Model Score : 0.9054794520547945 , Cross Validation Score : 0.8439490445859873
percentage of sensitivity = 82.45744963958023
percentage of precision = 81.57880819771148
Accuracy percentage = 89.59660297239914
```

Random Forest Classification

```
Random Forest Model Score : 0.9534246575342465 , Cross Validation Score : 0.8662420382165605
percentage of sensitivity = 81.43635669408866
percentage of precision = 86.92475732431339
Accuracy percentage = 91.0828025477707
```

Support Vector Classification

```
SVC Model Score : 0.9493150684931507 , Cross Validation Score : 0.856687898089172
percentage of sensitivity = 82.47827646453075
percentage of precision = 84.05896512424917
Accuracy percentage = 90.44585987261145
```

Logistic Regression

```
Logistic Regression Model Score : 0.9219178082191781 , Cross Validation Score : 0.8439490445859873
percentage of sensitivity = 82.68191653071378
percentage of precision = 81.84889755182768
Accuracy percentage = 89.59660297239914
```

Artificial Neural Network

(2 Hidden Layers | 5 nodes each)

```
ANN Model Score : 0.9123287671232877 , Cross Validation Score : 0.8280254777070064
percentage of sensitivity = 81.26511391116202
percentage of precision = 79.87166497804795
Accuracy percentage = 88.53503184713377
```

Naive Bayes

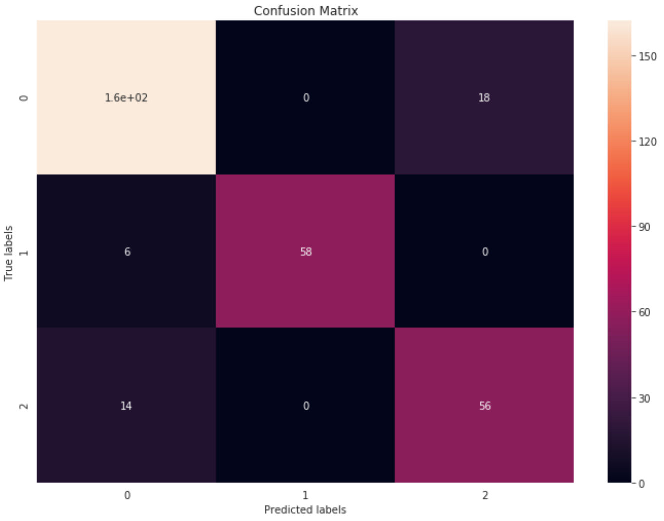
```
Naive Bayes Model Score : 0.6684931506849315 , Cross Validation Score : 0.6878980891719745
percentage of sensitivity = 65.36036192393581
percentage of precision = 63.70188370188371
Accuracy percentage = 79.19320594479831
```

8. Model Selection

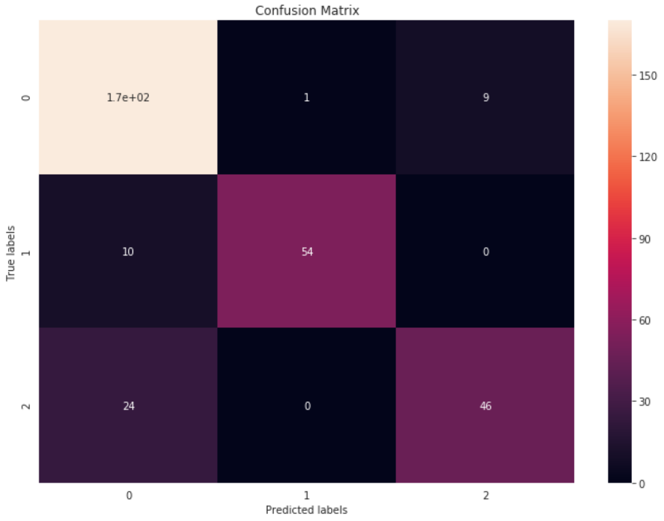
Comparison of the performance of each Model.

	Models					
	Decision Tree	Random Forest	SVM	Logistic Regression	ANN	Naive Bayes
Model Score	0.90	0.95	0.94	0.92	0.91	0.66
Cross-Validation Score	0.84	0.86	0.86	0.84	0.82	0.68
Accuracy Percentage	89.59	91.08	90.44	89.59	88.53	79.19

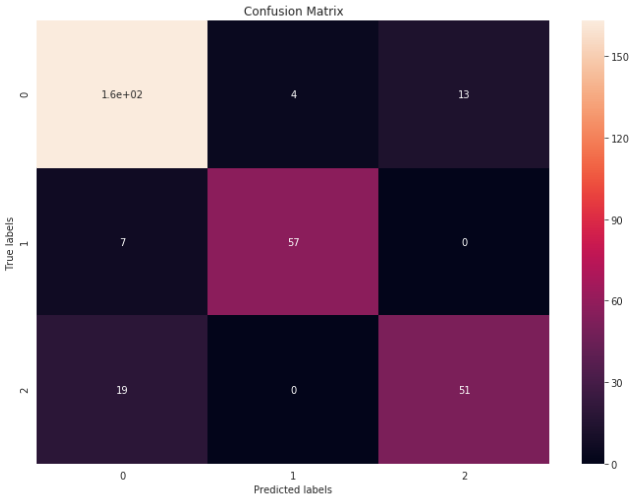
Confusion Matrices of all the models.

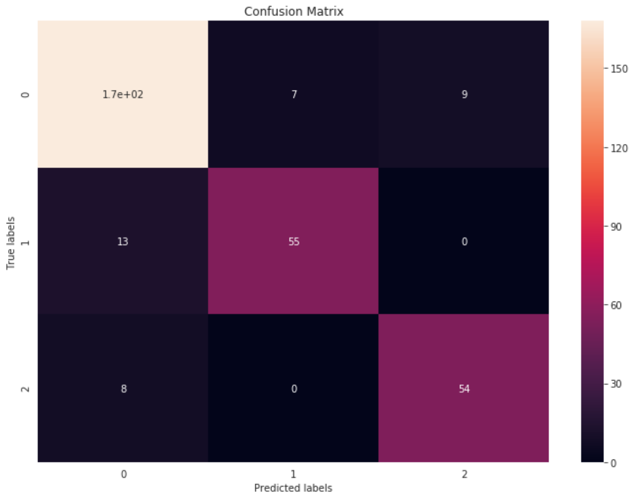


Decision Tree

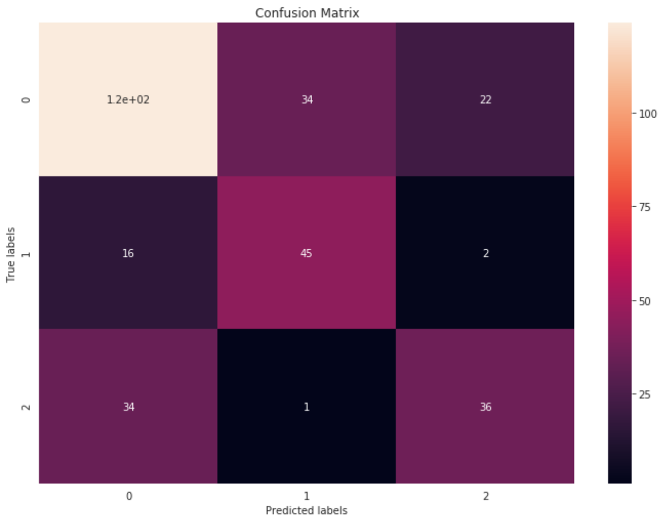


Random forest



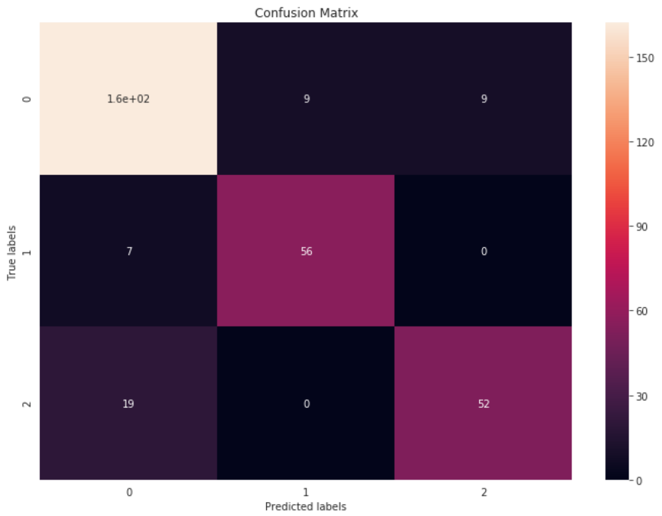


SVM



Naive Bayes

Logistic Regression



ANN

Report By-

Group 11

1. Arpit Jain, 2015047
2. Gautam Yadav, 2015093
3. Narosenla Longkumer, 2015165
4. Anuraag Singh, 2015043
5. Md. Arshad Siddiqui, 2015153
6. Harsh Agarwal, 2015102