Machine Learning Lab Report

Topic: **Predicting A Student's Performance**

Prepared by:

Md. Nowshad Hasan

2012331052

Submitted to:

Ayesha Tasnim

Assistant Professor, Computer Science & Engineering, SUST

Introduction:

Prediction of student performance using machine learning process is very important and interesting for every educational institute like school, college or university. When we've enough data of every student of a class of their previous course performance, then we can predict his or her next performance for a particular course or last final exam. It can help course teacher to classify student from best to good and bad to worst. After that, they can monitor students to improve their performance.

Data Set:

We used two datasets for our analysis. One from UCI and another from Kaggle.

Dataset UCI:

i) sex - student's sex (binary: female or male)
ii) age - student's age (numeric: from 15 to 22)
iii) school- student's school (binary: Gabriel Pereira or Mousinho da Silveira)
iv) address student's- home address type (binary: urban or rural)
v) Pstatus- parent's cohabitation status (binary: living together or apart)
vi) Medu- mother's education (numeric: from 0 to 4a)
vii) Mjob- mother's job (nominalb)
viii) Fedu- father's education (numeric: from 0 to 4a)
ix) Fjob father's job (nominalb)
x) guardian- student's guardian (nominal: mother, father or other)
xi) famsize- family size (binary: ≤ 3 or > 3)
xii) famrel- quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
xiii) reason- reason to choose this school (nominal: close to home, school reputation, course preference or other)
xiv) 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).
xv) studytime- weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
xvi) failures- number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
xvii) schoolsup- extra educational school support (binary: yes or no)

xviii) famsup- family educational support (binary: yes or no)
xix) activities- extra-curricular activities (binary: yes or no)
xx) paidclass- extra paid classes (binary: yes or no)
xxi) internet- Internet access at home (binary: yes or no)
xxii) nursery- attended nursery school (binary: yes or no)
xxiii) higher- wants to take higher education (binary: yes or no)
xxiv) romantic- with a romantic relationship (binary: yes or no)
xxv) freetime- free time after school (numeric: from 1 – very low to 5 – very high)
xxvi) gout- going out with friends (numeric: from 1 – very low to 5 – very high)
xxvii) Walc- weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
xxviii) Dalc- workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
xxix) health- current health status (numeric: from 1 – very bad to 5 – very good)

xxx) absences- number of school absences (numeric: from 0 to 93)
xxxi) G1- first period grade (numeric: from 0 to 20)
xxxii) G2- second period grade (numeric: from 0 to 20)

Output class is for Final Grade is G3, numeric from 0 to 20.

Here, this dataset contains almost 400 student data.

Dataset Kaggle:

1. Gender - student's gender (nominal: 'Male' or 'Female')

2. Nationality- student's nationality (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')

3. Place of birth- student's Place of birth (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')

4. Educational Stages- educational level student belongs (nominal: 'lowerlevel','MiddleSchool','HighSchool')

5. Grade Levels- grade student belongs (nominal: 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12 ')

6. Section ID- classroom student belongs (nominal:'A','B','C')

7. Topic- course topic (nominal:' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology')

8. Semester- school year semester (nominal:' First',' Second')

9. Parent responsible for student (nominal:'mom','father')

10. Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)

11. Visited resources- how many times the student visits a course content(numeric:0-100)

12. Viewing announcements-how many times the student checks the new announcements(numeric:0-100)

13. Discussion groups- how many times the student participate on discussion groups (numeric:0-100)

14. Parent Answering Survey- parent answered the surveys which are provided from school or not (nominal:'Yes','No')

15. Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:'Yes','No')

16. Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

Output is classified in three catagories:-

Low-Level: interval includes values from 0 to 69,

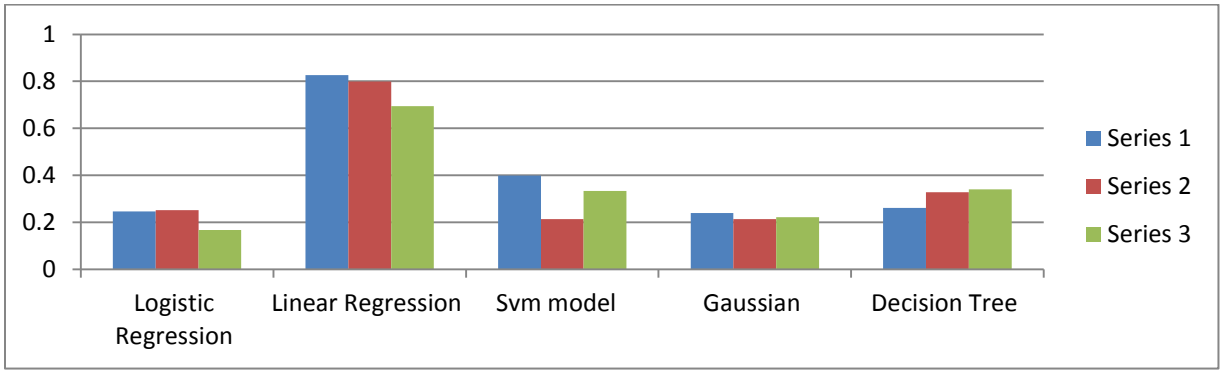
Middle-Level: interval includes values from 70 to 89,

High-Level: interval includes values from 90-100.

Analysis Process:

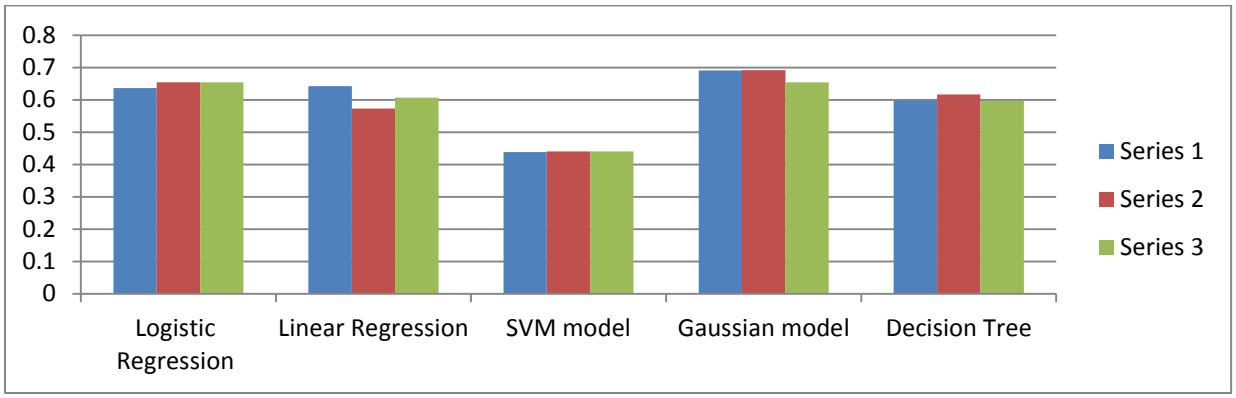
We used 5 machine learning models on this data set. Here, we used cross validation score that is 70% for training and 30% for testing for scoring on that particular model randomly. Here is chart for every dataset we used.

Dataset UCI:



Here, linear regression works fine.

Dataset Kaggle:



Here, we see that almost all model work fine but Gaussian model is the best.