Data Mining para Predição de Performance Estudantil

Nicholas Richers[1] e Bruno Almeida[2]

***Abstract*—**

I. DESCRIÇÃO DO PROBLEMA

Educação é reconhecidamente um fator chave para se alcançar uma economia próspera de longo prazo. Nas últimas décadas do século XX, o nível educacional entre os portugueses melhorou, no entanto, as estatísticas da primeira década do século XXI mantiveram Portugal entre os países com maiores taxas de insucesso e abandono escolar. Por exemplo, em 2006, 40% dos jovens portugueses entre 18 e 24 anos abandonavam precocemente a escola, enquanto a média da União Europeia foi de apenas 15% [1].

A área de educação é um terreno fértil para aplicações de BI devido as diversas fontes de dados e muitos grupos de interesse como alunos, professores, administradores, os próprios familiares ou ex alunos [2]. Existem várias questões interessantes que podem ser respondidas usando técnicas de Data Mining (DM) nesse âmbito [3]: Quem são os estudantes que recebem mais créditos por horas de aula? Quem é provável que retorne para as classes? Quais cursos podem ser oferecidos para atrair mais alunos? Quais são as principais razões para transferências de estudantes? É possível prever o desempenho do aluno? Quais são os fatores que afetam o desempenho do aluno?

Neste trabalho, analisaremos dados do mundo real de duas escolas públicas da região do Alentejo, em Portugal, durante o ano letivo de 2005 – 2006. Em Portugal, o ensino médio é constituído por 3 anos de escolaridade, precedendo 9 anos de ensino fundamental e seguido de ensino superior. A maioria dos alunos adere ao sistema de ensino público e gratuito. Existem vários cursos (por exemplo, Ciências e Artes Visuais) que compartilham disciplinas centrais, como a Língua Portuguesa e a Matemática. Uma escala de classificação de 20 pontos é usada, onde 0 é a nota mais baixa e 20 é a mais alta. Durante o ano letivo, os alunos são avaliados em três períodos e a última avaliação (G3) corresponde à nota final.

Duas diferentes fontes foram utilizadas para obtenção de dados: relatórios e questionários. O primeiro continha informações escassas (ou seja, apenas as notas e o número de ausências estavam disponíveis), foi complementado com o segundo, que agregou uma coleção maior de dados como informações demográficas, sociais e atributos escolares (por exemplo, idade do aluno, consumo de álcool e educação da mãe).

[1]N. Richers - Programa de Engenharia de Produção COPPE/UFRJ nicholasrichers@gmail.com

[2]B. Andrade - Programa de Engenharia de Produção COPPE/UFRJ bruno-andrade@hotmail.com

O objetivo é prever o desempenho do aluno e, se possível, identificar as principais variáveis que afetam o sucesso e fracasso educacional. As duas classes analisadas Matemática e Português serão modeladas com três objetivos de DM:

1) Classificação binária (aprovação / reprovação);
2) Classificação com cinco níveis (de muito bom ou excelente a insuficiente);
3) Regressão, com uma saída numérica que varia entre zero (0%) e vinte (100%).

Para cada uma dessas abordagens, serão testadas três configurações de entrada (por exemplo, com e sem as notas do período escolar) e quatro algoritmos de DM (por exemplo, Árvores de decisão e Random Forest). Além disso, uma análise explicativa será realizada sobre os melhores modelos, a fim de identificar as características mais relevantes.

II. PESQUISA BIBLIOGRÁFICA

Data mining é o campo da descoberta de novas informações potencialmente aproveitáveis a partir de grandes quantidades de dados [4]. Nesse contexto, técnicas de DM aplicadas ao ramo da educação ainda estão nos primeiros passos [5]. O artigo de Minaei-bigdoli [7], é citado por diversos autores [8][6], como um dos primeiros trabalhos a utilizarem algoritmos genéticos para prever performance acadêmica de estudantes.

Ao longo dos anos diversos artigos foram publicados com esse tema, contudo, nota-se divergências quanto ao modelo de melhor performance nesses estudos. Em [9] Naive Bayes é citado como a melhor performance, já em [10] o melhor resultado encontrado foi com o SVM e em [8] Random Forest.

Essa diferença pode ser causada por diferenças nos atributos considerados e no tamanho do conjunto de dados, onde em [11] três métodos foram usados para lidar com o problema de desequilíbrio de classes e todos eles mostram resultados satisfatórios. Primeiro, balancearam os conjuntos de dados e usaram o o SVM para os pequenos conjuntos de dados e Decision Tree para os conjuntos de dados maiores.

Nesse contexto se destaca uma revisão sistemática da literatura [12] que realiza um levantamento quantitativos das principais técnicas de data mining e também procura identificar quais os atributos mais importantes nos dados dos estudantes. Dez dos trinta artigos avaliados consideram o histórico de notas do aluno o principal atributo para predição de desempenho [12], seguidos por dados demográficos como a escolaridade dos pais.

O levantamento das técnicas mais usadas [12] revela que Decision Tree (DT) é a técnica mais utilizada, presente em dez dos trinta estudos avaliados, seguido por Neural Network

TABLE I
MY CAPTION

Atributo	Descrição
gênero	gênero do estudante (binário: feminino ou masculino)
idade	idade do estudante (numérico: 15 a 22 anos)
escola	de qual escola é o estudante (binário: Gabriel Pereira ou Mousinho da Silveira)
endereço	o tipo de endereço do estudante (binário: urbano ou rural)
Pstatus	binário: se os pais moram juntos ou separados
Medu	educação da mãe (numérico: de 0 a 4ª)
Mjob	trabalho da mãe (nominal)
Fedu	educação do pai (numérico: 0 a 4ª)
Fjob	trabalho do pai (nominal)
guardian	responsável pela guarda do estudante (nominal: mãe, pai, outro)
famsize	tamanho da família (binário: <= 3 ou >3
famrel	qualidade das relações familiares (numérico: 1 - muito ruim até 5 - excelente)
reason	razão pela qual escolheu esta escola (nominal: próximo de casa, reputação da escola, preferência do curso, ou outro)
traveltime	tempo do percurso de casa até a escola (numérico: 1 - <15 min , 2 - 15 a 30 min, 3- 30 min a 1 hora, ou 4 - >1 hora)
studytime	tempo de estudo por semana (numérico: 1 - <2 horas, 2 2 - 2 a 5 horas, 3 - 5 a 10 horas, 4 - >10 horas)
failures	numero de reprovações (numérico: n, se 1<= n <3, se não 4)
schoolsup	apoio escolar extra (binário: sim ou não)
famsup	suporte educacional da família (binário: sim ou não)
activities	atividades extracurriculares (binário:sim ou não)
paidclass	aulas particulares (binário: sim ou não)
internet	acesso a internet em casa (binário: sim ou não)
nursery	frequentou o maternal (binário: sim ou não)
higher	pretende cursar o ensino superior (binário: sim ou não)
romantic	está em algum relacionamento (binário: sim ou não)
freetime	tempo livre fora da escola (numérico: 1 - bastante tempo livre até 5 - pouco tempo livre)
goout	sai com amigos (numérico: 1 - muito pouco, até, 5 - bastante)
Walc	consumo de álcool semanal (numérico: 1 - muito pouco até 5 - bastante)
Dalc	consumo de álcool diário (numérico: 1 - muito pouco até 5 - bastante)
health	estado de saúde atual (numérico: 1 - muito ruim até 5 - muito bom)
absences	número de faltas na escola (de 0 até 93)
G1	nota do primeiro período (numérico: 0 a 20)
G2	nota do segundo período (numérico: 0 a 20)
G3	nota do terceiro período (numérico: 0 a 20)

(NN) com oito, Naive Bayes (NB) com quatro e K-Nearest Neighbor (kNN) com apenas três.

Levando em conta o melhor desempenho desempenho de cada técnica [12] considerando todos os artigos temos: NN com 98% como o melhor resultado, isso ocorreu devido a influência de um híbrido de dois dos principais atributos que eram o sistema de avaliação interno e externo da escola. Em seguida DT com 91%, SVM e kNN com 83% e por fim NB com 76%.

III. DESCRIÇÃO DOS DADOS

O banco de dados foi construído a partir de duas fontes: relatórios escolares, incluindo certos atributos (as três séries do período e o número de faltas escolares) e questionários utilizados para complementar a informação anterior. No questionário foram feitas perguntas fechadas relacionadas a vários dados demográficos (educação da mãe, renda familiar), sociais/emocionais (consumo de álcool) e relacionados à escola (número de falhas de classe anteriores), e demais variáveis que deveriam afetar o desempenho dos alunos.

O questionário foi revisado por profissionais da escola e testado em um pequeno grupo de 15 alunos, a fim de obter um avaliação prévia. A versão final continha 37 questões e foi respondida em aula por 788 alunos. Posteriormente, 111 respostas foram descartadas devido à falta de detalhes de identificação (necessário para se fundir com os relatórios da escola). Por fim, os dados foram integrados em dois conjuntos de dados relacionados à Matemática (com 395 exemplos) e à língua portuguesa (649 registros).

Durante o estágio de pré-processamento, algumas características foram descartadas devido ao excesso de valores ausentes. Por exemplo, poucos entrevistados responderam sobre sua renda familiar (provavelmente devido a questões de privacidade), enquanto quase 100% dos estudantes moram com os pais e têm um computador pessoal em casa. Os atributos restantes são mostrados na Tabela 1, onde as últimas quatro linhas denotam as variáveis retiradas dos relatórios escolares.

IV. APRESENTAÇÃO TECNOLÓGICA

As experiências aqui relatadas foram realizadas usando a biblioteca **RMiner**[1], uma biblioteca de código aberto para o ambiente R que facilita o uso de técnicas de DM. R é uma linguagem de programação matricial gratuita e de alto nível com um poderoso conjunto de ferramentas para análise estatística e de dados. A biblioteca RMiner apresenta um conjunto de funções coerentes (mineração, saveMining) para tarefas de classificação e regressão. Em particular os pacotes Rpart (DT), Random Forest (RF), Neural Networks (NN) e Kernlab (SVM). Também foram utilizados os pacotes **fBasics**, **grid**, **gridExtra** e **corrplot**.

[1]http://www.dsi.uminho.pt/˜pcortez/R/rminer.zip

	nobs	NAs	Minimum	Maximum	1. Quartile	3. Quartile	Mean	Median	Stdev
age	649	0	15	22	16	18	16.744222	17	1.218138
Medu	649	0	0	4	2	4	2.514638	2	1.134552
Fedu	649	0	0	4	1	3	2.306626	2	1.099931
traveltime	649	0	1	4	1	2	1.568567	1	0.74866
studytime	649	0	1	4	1	2	1.930663	2	0.82951
failures	649	0	0	3	0	0	0.22188	0	0.593235
famrel	649	0	1	5	4	5	3.930663	4	0.955717
freetime	649	0	1	5	3	4	3.180277	3	1.051093
goout	649	0	1	5	2	4	3.1849	3	1.175766
Dalc	649	0	1	5	1	2	1.502311	1	0.924834
Walc	649	0	1	5	1	3	2.280431	2	1.28438
health	649	0	1	5	2	5	3.53621	4	1.446259
absences	649	0	0	32	0	6	3.659476	2	4.640759
G1	649	0	0	19	10	13	11.399076	11	2.745265
G2	649	0	0	19	10	13	11.570108	11	2.913639
G3	649	0	0	19	10	14	11.906009	12	3.230656

Fig. 1. Parâmetros Estatísticos do conjunto de dados

V. AVALIAÇÃO PRELIMINAR DOS DADOS

Nessa seção é feita uma análise preliminar dos dados, ressaltando que para essa etapa não foi feita nenhuma etapa de pré processamento dos dados e os códigos desenvolvidos se encontram no apêndice desse estudo.

A. Análise Exploratória

1) Estatísticas Básicas A Figura 2 apresenta estatísticas básicas das variáveis numéricas do conjunto de dados de forma a auxiliar a análise dos dados.

2) Histogramas
A Figura 3 dispões os respecitivos histogramas das variáveis numéricas do conjunto de dados.

3) Verificação de Outliers
Na Figura 4 verificamos a existência de Outliers especialmente na variável referente as faltas.

4) Correlações
Analisando a matriz de correlação verificamos uma correlação próxima a 0.6 entre a educação do Pai (Fedu) com a da Mãe (Medu), o consumo de álcool em dias de semana (Dalc) com o consumo em finais de semana (Walc), e uma correlação próxima a 0.8 entre as primeiras notas (G1 e G2) e a variável de saída (G3).

B. Conclusões

A Pesquisa Bibliográfica indica que a variável de maior importante na predição do desempenho dos alunos são as notas anteriores (G1 e G2) para prever a nota final (G3). Matriz de correlação do conjunto de dados desse estudo dá a indicação que também seguirá esse caminho.

school	GP:423	MS:226	NA	NA	NA
sex	F:383	M:266	NA	NA	NA
address	R:197	U:452	NA	NA	NA
famsize	GT3:457	LE3:192	NA	NA	NA
Pstatus	A: 80	T:569	NA	NA	NA
Mjob	at_home :135	health : 48	other :258	services:136	teacher : 72
Fjob	at_home : 42	health : 23	other :367	services:181	teacher : 36
reason	course :285	home :149	other : 72	reputation:143	NA
guardian	father:153	mother:455	other : 41	NA	NA
schoolsup	no :581	yes: 68	NA	NA	NA
famsup	no :251	yes:398	NA	NA	NA
paid	no :610	yes: 39	NA	NA	NA
activities	no :334	yes:315	NA	NA	NA
nursery	no :128	yes:521	NA	NA	NA
higher	no : 69	yes:580	NA	NA	NA
internet	no :151	yes:498	NA	NA	NA
romantic	no :410	yes:239	NA	NA	NA

Fig. 2. Dados Não numéricos

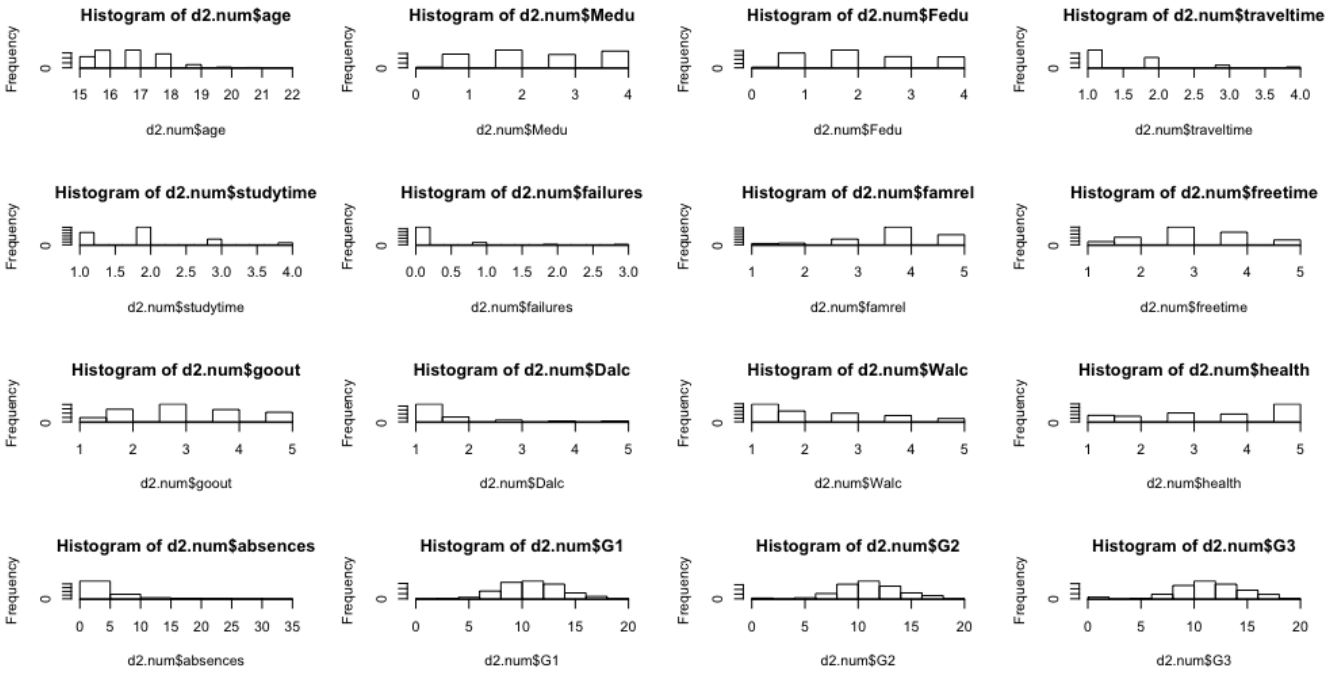

Fig. 3. Histogramas

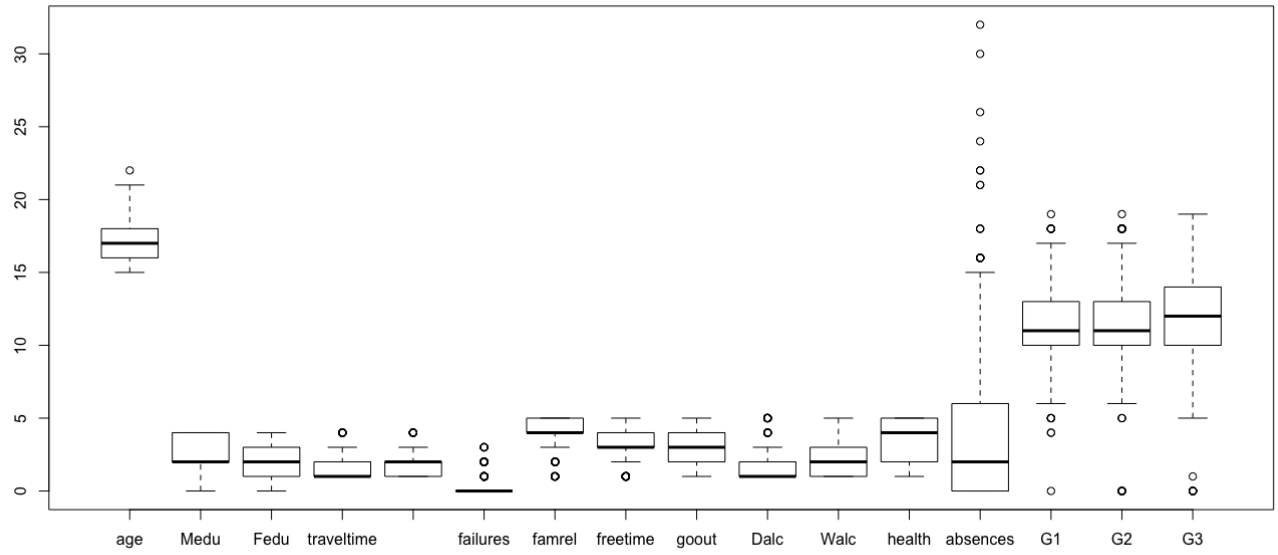

Fig. 4. Gráfico Boxplot

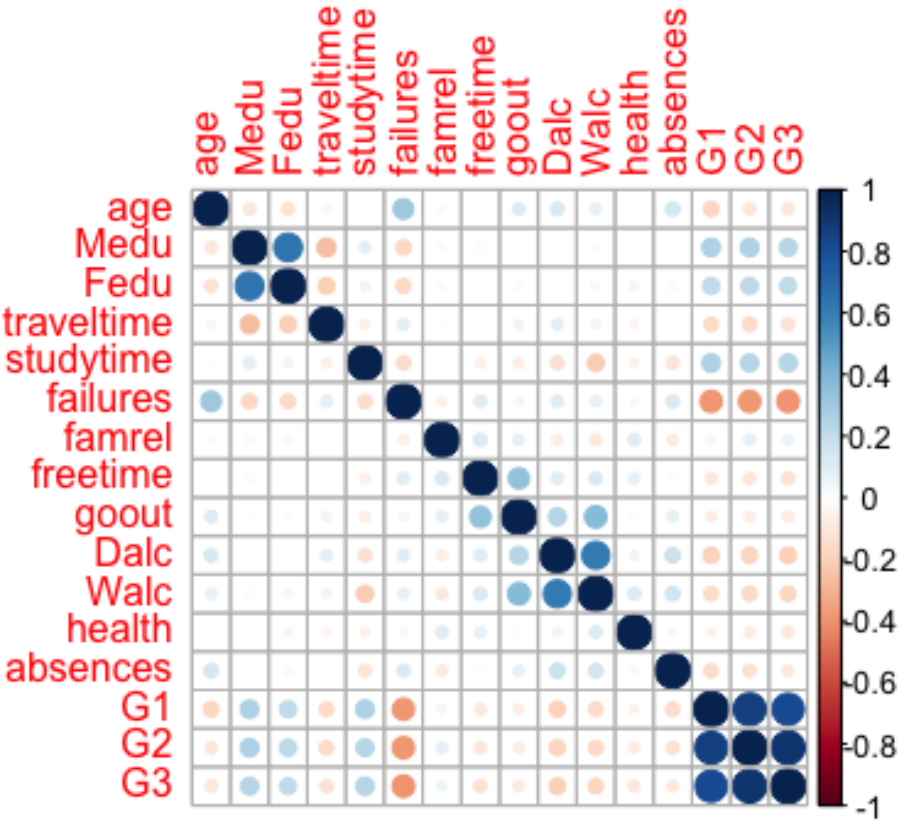

Fig. 5. Matriz de Correlação

```
library("fBasics")
library("grid")
library("gridExtra")
library("corrplot")

d1=read.table("student-mat.csv",sep=";",header=TRUE)
d2=read.table("student-por.csv",sep=";",header=TRUE)

#tabela com valores numericos
num.columns <- sapply(d2, is.numeric)
tab1 <- basicStats(d2[,num.columns])
numeric.table <- t(tab1[c(1:8,14),])
grid.table(numeric.table)



#tabela com valores nao numericos
not.num.columns <- !num.columns
tab2 <-t(summary(d2[,not.num.columns]))
not.numeric.table <- t(tab2[1:8,])
grid.table(not.numeric.table)

#Matrix Correlacao
COR <- cor(d2[,num.columns])
corrplot(COR, method="circle")


#Matriz de Projecao
pairs(d2[,num.columns])


#histogramas
par(mfrow=c(4,4))
hist(d2.num$age) hist(d2.num$Medu) hist(d2.num$Fedu) hist(d2.num$traveltime)
hist(d2.num$studytime) hist(d2.num$failures) hist(d2.num$famrel)
hist(d2.num$freetime) hist(d2.num$goout) hist(d2.num$Dalc) hist(d2.num$Walc)
hist(d2.num$health) hist(d2.num$absences) hist(d2.num$G1) hist(d2.num$G2)
hist(d2.num$G3)
```

Fig. 6. Código para gerar as imagens no R

REFERENCES

[1] Eurostat, 2007. Early school-leavers. http://epp.eurostat.ec.europa.eu/ (Acesso em Abr. 20, 2018).

[2] 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.

[3] Luan J., 2002. Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 113, 17–36.

[4] Kaur P., Singh M., Josan G. S., Classification and prediction based data mining algorithms to predict slow learners in education sector, Procedia Computer Science 57 (2015) 500 – 508.

[5] V.Ramesh, P.Parkavi, K.Ramar(2013), "Predicting student performance: A statistical and datamining approach",International journal of computer applications , Volume 63- no. 8, pp 35-39.

[6] Ahmed M. A., Rizaner A., Ulusoy A. H., Using data mining to predict instructor performance, Procedia Computer Science 102 (2016) 137 – 142.

[7] Minaei-Bidgoli B, Punch WF. Using genetic algorithms for data mining optimization in an educational web based system. Genetic and Evolutionary Computation, Springer Berlin Heidelberg, 2003, p. 2252-2263.

[8] P.Cortez, A.Silva (2008),"Using Data Mining To Predict Secondary School Student Performance", InEUROSIS, A.BritoandJ. Teixeira (Eds.), pp 5-12.

[9] 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.

[10] Sorour SE, Mine T, Goda K, Hirokawa S. Estimation of student performance by considering consecutive lesson. 4th International Congress on Advanced Applied Informatics, 2015, p.121-126.

[11] Nguyen Thai-Nghe, Andre Busche,LarsSchmidt Thieme(2009), "Improving Academic Performance Prediction by Dealing with Class Imbalance", Ninth International Conference on Intelligent Systems Design and Applications.

[12] Shahiri A. M, Husain W., Rashid N. A., The Third Information Systems International Conference A Review on Predicting Student's Performance using Data Mining Techniques, Procedia Computer Science 72 (2015) 414 – 422.