20/21

Data Mining para a Ciência de Dados

Universidade do Minho
Escola de Engenharia

João André de Castro Macedo
A68463

Marta Isabel Marinho de Andrade
A81483

Nelson José Marques Martins Almeida
A95652

Grupo 6

Estudo do Dataset “Student Performance”

Índice

Índice de Figuras

Índice de Gráficos

Índice de Tabelas

1. Introdução

No âmbito da unidade curricular optativa de Data Mining para a Ciência de Dados foi proposto o desenvolvimento de um projeto que utiliza técnicas de Data Mining para fazer análises descritivas, explicativas ou preditivas sobre um conjunto de dados do mundo real.

Neste caso, foram utilizados dados fornecidos pelo docente sobre o desempenho de alunos do Ensino Secundário no que toca a duas disciplinas: Português e Matemática. Estes dados incluem as notas dos alunos, assim como dados demográficos, sociais e escolares destes.

Usando a metodologia Crisp-DM como guia no desenvolvimento do trabalho, este projeto encontra-se dividido nas seguintes fases:

- Business Understanding: nesta fase inicial são identificados os requisitos do projeto, assim como identificado um objetivo principal do negócio, que irá guiar todo o restante processo de modo a que no fim, este objetivo seja cumprido;
- Data Understanding: aqui será feita a compreensão dos dados que irão ser trabalhados, sendo identificados problemas de qualidade e retirados insights, de modo a retirar informação ignorada;
- Data Preparation: com base nas conclusões retiradas da fase anterior, nesta fase serão realizadas todas as atividades de modo a criar o conjunto final de dados que serão utilizados na fase seguinte;
- Modeling: através da aplicação de várias técnicas de modelagem, será realizada a previsão de informação através do uso dos dados selecionados na fase anterior;
- Evaluation: nesta fase final, serão avaliados os resultados atingidos na fase anterior e comparados com as metas traçadas na primeira fase de modo a verificar se estes foram atingidos.

2. Execução do Projeto

De forma geral, a equipa sente que existiu um bom ambiente durante todo o desenvolvimento do projeto, existindo boa comunicação, troca de ideias e empenho por parte de todos os elementos do projeto.

Inicialmente a equipa era constituída por 4 elementos, mas um deles não efetuou qualquer contribuição para o projeto, tendo desistido da realização deste. Este fator exigiu uma flexibilidade por parte dos restantes elementos pois exigiu uma redistribuição das tarefas.

O fato da realização do projeto ser realizada totalmente a partir de casa dificultou de certa forma a resolução de problemas de código e esclarecimento de dúvidas.

A equipa reconhece que inicialmente teve dificuldades de divisão de tarefas e identificação de um objetivo de negócio, só começando a parte de Modeling mais tarde do que pretendido.

2.1. Divisão de Tarefas e Autoavaliação

André Macedo

- Autoavaliação:16
- Na fase inicial do projeto, fiquei encarregue da exploração dos datasets e criação de KPI's, tal como cada elemento do grupo. Realizei a transformação dos dados necessárias para a utilização no modelo, e realizei, juntamente com o Nelson Almeida, a modelação dos modelos identificados pelo AutoML. Juntamente com os restantes elementos, através das análises das métricas, identificamos os resultados atingidos.

Marta Andrade

- Autoavaliação:16
- Inicialmente, realizei a criação de KPI's como forma de explorar os datasets. Fiquei encarregue do estudo das regras de Associação e do Clustering sobre os dados do dataset, além de fornecer apoio à realização da modelação. Em conjunto com os outros elementos, realizei a avaliação dos resultados da modelação.

Nelson Almeida

- Autoavaliação:16
- Realizei um estudo dos datasets que resultou no desenvolvimento de KPI's, e fiquei encarregue do estudo da tecnologia de AutoML de forma a que esta forneça os melhores modelos para cada cenário. Continuei o estudo desta com a modelação dos modelos escolhidos, e posteriormente realizei o estudo dos resultados obtidos.

O grupo considera que todos os elementos trabalharam por igual, não sendo justo a realização de uma diferenciação de notas.

2.2. Autoavaliação do Projeto

O grupo considera que o projeto merece 16 valores porque conseguiu realizar a modelação de uma previsão, assim como estudar Clustering e Association dos dados, mas não consegui realizar mais objetivos de negócio com outros cenários possíveis de aplicação. Em termos de execução do projeto, teve um desempenho menos bom numa fase inicial porque ainda não se encontrava familiarizado com a linguagem de programação R, inicializando a fase de Modeling mais tarde do que previsto.

3. Estudo CRISP-DM

3.1. Business Understanding

3.1.1. Determine Business Objectives

3.1.1.1. Background

Em Outubro de 2020 a equipa foi contratada pelas escolas "Gabriel Pereira" e "Mousinho da Silveira" para identificar que alunos estariam mais inclinados a reprovar a uma disciplina, de forma a puder assim entrar em contato com estes e desenvolver um plano de ajuda personalizado para que estes consigam ter um desempenho positivo na disciplina em questão.

3.1.1.2. Business Objectives

Indo de encontro ao Background, a equipa tem como objetivo de negócio "Conhecer que alunos estão em risco de reprovar".

3.1.1.3. Business Success Criteria

A solução final criada deve ser classificada num destes três critérios:

- Resultado Final Ideal: Previsão dos casos acima de 80%;
  Neste caso a solução criada é considerada de qualidade e não requer qualquer revisão ou melhoramento, podendo ser implementada imediatamente.
- Resultado Final Aceitável: Previsão dos casos entre 55% e 80%;
  Caso a solução esteja inserida neste critério, continua a ter qualidade, mas requer uma revisão de forma a ser melhorada para posteriormente ser implementada.
- Resultado Final Insatisfatório: Previsão dos casos abaixo de 55%.
  Se a solução estiver inserida neste critério, a solução produzida não é benéfica e deve ser recomeçada.

3.1.2. Assess Situation

3.1.2.1. Inventory of Resources

- Recursos Humanos: A equipa é composta por 3 alunos, sem experiência previa em Data Mining;

- Dados: Os dados utilizados foram disponibilizados pelo docente, através dum ficheiro CSV;
- Hardware: 3 computadores, cada um deles pertencente a um membro da equipa de trabalho;
- Software: Como um dos requisitos era a utilização da linguagem R, a equipa utilizou RStudio, GitHub, OneDrive e Microsoft Office durante a realização do projeto.
-

3.1.2.2. Requirements, Assumptions, and Constraints

Em termos de requisitos, temos:

- Realização do projeto na linguagem R;
- Entrega do projeto até dia 12/01/2021;
- Utilizar a metodologia CRISP-DM;
- Compromisso de cumprir o contrato.

Em termos de Pressupostos, temos:

- A variável target é "Transition";

Em termos de Restrições, temos:

- Restrição temporal de realização do projeto;
- Realização do projeto à distância, sem reuniões presenciais;

3.1.2.3. Risks and Contingencies

A tabela a seguir exposta apresenta possíveis riscos e as suas contingências.

Tabela 1 - Tabela de Riscos e Contingências

Risco	Contingência
Inexperiência da Equipa	Visualização de tutoriais online como forma de ajuda; Reuniões com o docente como forma de tirar dúvidas;
Objetivos de Negócio e Objetivos de Data Mining mal definidos	Rever a fase de Business Understanding; Rever dataset de modo a identificar que informação pode ser retirada deste;
Mudança de tecnologias usadas	Rever requisites das tecnologias e efetuar uma pesquisa previa antes de a instalar;
Falta de comunicação entre elementos do grupo	Agendar reuniões de rotina de forma a identificar possíveis problemas;

3.1.3. Determine Data Mining Goals

O objetivo de Data Mining estabelecido pela equipa para este projeto é "Prever os alunos que estão em risco de reprovar". Esta previsão tem por base dois cenários:

- Cenário 1: Neste cenário pretendemos avaliar o impacto que certos comportamentos de risco como as saídas, consumo de álcool e uso excessivo da internet podem ter no sucesso académico;
- Cenário 2: O papel deste cenário é demonstrar a importância de fatores externos ao "estudo" individual, como por exemplo, o apoio escolar, o relacionamento familiar, o estatuto parental, o apoio da família e as "explicações" ou apoio individual pago, no sucesso académico.

3.1.3.1. Data Mining Success Criteria

De forma a assegurar a maior precisão possível das nossas previsões, recorremos a alguns mecanismos de avaliação de modelos como por exemplo:

- Matriz de confusão;
- Curva de ROC;
- Acuidade;
- Especificidade;
- Precisão.

Estas métricas devem encontrar-se nos seguintes intervalos:

- Resultado Final Ideal: Previsão dos casos acima de 80%;
- Resultado Final Aceitável: Previsão dos casos entre 55% e 80%;
- Resultado Final Insatisfatório: Previsão dos casos abaixo de 55%.

3.2. Data Understanding

3.2.1. Collect Initial Data

Os Datasets escolhidos para a realização deste projeto foram os datasets disponibilizados pelo regente sobre a abordagem do aproveitamento dos alunos do ensino secundário de duas escolas Portuguesas, criado através da realização de questionários e através do uso de relatórios das escolas. Estes Datasets são referentes ao desempenho em duas disciplinas distintas: Matemática (*Student-mat.csv*), com 395 registos e Língua Portuguesa (*Student-por.csv*), com 649 registos. Ambos datasets contêm 32 atributos.

3.2.2. Describe Data

De forma a perceber os dados adquiridos foi construída a tabela abaixo apresentada de forma a obter uma clara leitura das propriedades de cada atributo. Como os atributos são iguais para os dois Datasets, não se verificou a necessidade de construir duas tabelas descritivas, uma para cada Dataset. Além disso, nenhum dos atributos apresenta Valores Nulos ou Valores Únicos.

Tabela 2 - Análise Descritiva dos Atributos

Atributos	Descrição	Tipo do Atributo	Valores Possíveis
School	Escola do Estudante	Binário	"GP" - Gabriel Pereira ; "MS" - Mousinho da Silveira.
Sex	Género do Estudante	Binário	"F" – Mulher; "M" – Homem.
Age	Idade do Estudante	Numérico	15 a 22
Address	Tipo de local de residência do Estudante	Binário	"U" – Urbano; "R" – Rural.
Famsize	Tamanho da Familia	Binário	"LE3" – menor ou igual a 3 indivíduos; "GT3" – maior que 3 indivíduos.
Pstatus	Situação de coabitação dos pais	Binário	"T" - Moram juntos; "A" – Moram Separados.
Medu	Educação da mãe	Numérico	0 – Nenhuma; 1 – Educação Primária (até 4º ano); 2 – Educação Básica (5º ao 9º ano); 3 – Educação Secundária (12º ano); 4 – Ensino Superior.
Fedu	Educação do pai	Numérico	0 – Nenhuma; 1 – Educação Primária (até 4º ano); 2 – Educação Básica (5º ao 9º ano); 3 – Educação Secundária (12º ano); 4 – Ensino Superior.
Mjob	Profissão da mãe	String	"teacher" – Professora; "health" – Cargo relacionado com Saúde; "services" – Cargo relacionado com prestação de serviços civis, como polícia ou administração; "at_home" – Emprego a partir de Casa; "other" – Outros.
Fjob	Profissão do pai	String	"teacher" – Professor; "health" – Cargo relacionado com Saúde; "services" – Cargo relacionado com prestação de serviços civis, como polícia ou administração; "at_home" – Emprego a partir de

			Casa; "other" – Outros.
Reason	Razão de Escolha desta Escola	String	"home" – Perto de casa; "reputation" – Reputação da Escola; "course" – Preferência de curso; "other" – Outros.
Guardian	Encarregado de Educação do Aluno	String	"mother" – Mãe; "father" – Pai; "other" – Outro.
traveltime	Tempo de viagem de casa a escola	Numérico	1 – Menos de 15 minutos; 2 - 15 a 30 minutos; 3 - 30 minutos a 1 hora; 4 – Mais de 1 hora.
Studytime	Tempo de estudo semanalmente	Numérico	1 – Menos de 2 horas; 2 - 2 a 5 horas; 3 - 5 a 10 horas; 4 – Mais de 10 horas;
Failures	Número de chumbos passados	Numérico	0 a 3
Schoolsup	Tem apoio extra nos estudos	Binário	"yes" – Sim; "no" – Não.
Famsup	Tem apoio familiar extra nos estudos	Binário	"yes" – Sim; "no" – Não.
Paid	Frequenta aulas pagas extra para os estudos da disciplina (Português ou Matemática	Binário	"yes" – Sim; "no" – Não.
Activities	Frequenta atividades extracurriculares	Binário	"yes" – Sim; "no" – Não.
Nursery	Frequentou Jardim de Infância	Binário	"yes" – Sim; "no" – Não.
higher	Quer ingressar no Ensino Superior	Binário	"yes" – Sim; "no" – Não.
Internet	Possui acesso à Internet em casa	Binário	"yes" – Sim; "no" – Não.
Romantic	Encontra-se numa relação	Binário	"yes" – Sim; "no" – Não.
Famrel	Qualidade das relações familiares	Numérico	Classificação de 1 a 5, com 1 – Muito Mau a 5 - Excelente
Freetime	Tempo livre depois da escola	Numérico	Classificação de 1 a 5, com 1 – Muito Pouco a 5 - Bastante
Goout	Saidas com amigos	Numérico	Classificação de 1 a 5, com 1 – Muito Poucas a 5 - Bastantes
Dalc	Consumo de álcool durante os dias de escola	Numérico	Classificação de 1 a 5, com 1 – Muito Pouco a 5 - Bastante
Walc	Consumo de álcool durante os fins-de-semana	Numérico	Classificação de 1 a 5, com 1 – Muito Pouco a 5 - Bastante
Health	Condição de Saúde Atual	Numérico	Classificação de 1 a 5, com 1 – Muito

			Mau a 5 – Muito Bom
Absences	Número de faltas escolares	Numérico	0 a 93
Os Atributos abaixo representam as notas a Português no dataset de Português como as notas de Matemática no dataset de Matemática			
G1	Nota do 1º Período	Numérico	0 a 20
G2	Nota do 2º Período	Numérico	0 a 20
G3	Nota final	Numérico	0 a 20

3.2.3. Explore Data

Neste ponto foram realizadas vários estudos sobre os dados de forma a obter uma melhor compreensão destes. Estes estudos são descritos em Anexos.

- Análise Descritiva de Dados: encontram-se as análises feitas a cada atributo, onde é apresentado no caso dos atributos numéricos, o Min, Max, Mean, Median, 1st Quadran e 3rd Quadran. É verificado também a existência de Missing Values e valores nulos nos dois datasets;
- KPI's: foram construídos vários gráficos de forma a melhor perceber as relações entre os atributos;
- Association: estudo em termos de Regras de Associação de modo a encontrar correlações frequentes entre conjuntos de itens no Dataset;
- Clustering: é realizado um agrupamento automático de dados sobre dois atributos de modo a perceber quão semelhantes ou diferentes eles são uns dos outros

3.2.4. Verify Data Quality

Neste ponto não foram identificadas alterações a fazer nos Atributos pois estes já se encontravam tratados, sem qualquer tipo de erros de formatação.

3.3. Data Preparation

3.3.1. Select Data

Depois da exploração dos dados, a equipa chegou a conclusão de que os atributos "School", "G1" e "G2" deveriam ser retirados do estudo pois este pretende utilizar uma visão global dos estudantes, não sendo necessário a identificação da escola a que este pertence. Além disso, este estudo seria realizado previamente ou durante o ano escolar, o que impediria a utilização de G1 e G2 pois estes valores ainda seriam desconhecidos.

3.3.2. Clean Data

Como referido anteriormente não foram identificadas alterações a fazer nos Atributos pois estes já se encontravam tratados, sem qualquer tipo de erros de formatação.

3.3.3. Construct Data

Como o dataset já se encontra organizado, neste passo não foram realizadas nenhumas ações.

3.3.4. Integrate Data

Depois de realizar a exploração dos dados e conseguir assim uma melhor compreensão destes, a equipa chegou à conclusão de que como o objetivo de negócio não se encontra relacionado a uma disciplina especifica, trata-se de uma análise global do aproveitamento dos estudantes, os datasets poderiam ser juntos num dataset apenas, que seria então utilizado para realizar as previsões, não sendo necessário assim a diferenciação entre dados referentes a disciplina de Português e Matemática.

Foi realizado então o *merge* do dataset de Matemática (*Student-mat.csv*) e do dataset de Língua Portuguesa (*Student-por.csv*) de acordo com a imagem abaixo apresentada.

```
# Merge both dataframes
grades <- rbind(student_mat, student_por)
```

Figura 1 - Merge dos datasets

3.3.5. Format Data

De forma a preparar o dataset para os modelos de previsão, os atributos categóricos sofreram alterações, onde similarmente aos atributos numéricos, foi-lhes atribuído um número como identificador, em vez de texto.

Foi criada a variável Target, de nome “Transition”, baseada no atributo “G3”. Se “G3” se encontrar entre 0 e 10, isso indica-nos que o aluno reprovou; se for maior ou igual a 10 significa que o aluno passou, ou seja, teve um bom desempenho escolar.

Tabela 3 – Alterações realizadas nos Atributos

Atributos	Tipo do Atributo	Valores Possíveis
Transition	Numérico	0-10=0 10-20=1
romantic	Numérico	0=no 1=yes
internet	Numérico	0=no

		1=yes
higher	Numérico	0=no 1=yes
nursery	Numérico	0=no 1=yes
activities	Numérico	0=no 1=yes
paid	Numérico	0=no 1=yes
famsup	Numérico	0=no 1=yes
schoolsup	Numérico	0=no 1=yes
guardian	Numérico	1= mother 2= father 3= other
reason	Numérico	1= home 2= reputation 3= course 4= other
Fjob	Numérico	1= teacher 2= health 3= services 4= at_home 5= other
Mjob	Numérico	1= teacher 2= health 3= services 4= at_home 5= other
Pstatus	Numérico	1= T 2= A
famsize	Numérico	1= LE3 2= GT3
address	Numérico	1= U 2= R
sex	Numérico	1= M 2= F

3.4. Modeling

3.4.1. Select Modeling Technique

De forma a prever que alunos estão em risco de reprovar, foi utilizada a técnica de "Automatic Machine Learning" da plataforma H2O, de modo a identificar que modelo é melhor consoante aos atributos selecionados.

Foram criados 2 cenários que a equipa achou pertinente para o objetivo em questão:

- Cenário 1: Neste cenário a equipa escolheu atributos que representam possíveis comportamentos de risco, de forma a verificar assim o seu impacto:

  Atributos: "goout","Dalc","Walc",”internet”

Depois de selecionados estes atributos, através do uso da função *automl* da plataforma H2O, foi identificado que o Modelo Ideal para este cenário seria o *Gradient Boosting Machine* (GBM).

	model_id	mean_residual_deviance	rmse	mse	mae	rmsle
1	GBM_2_AutoML_20210111_183825	0.4093921	0.6398376	0.4093921	0.5248134	0.3907515
2	GBM_3_AutoML_20210111_183825	0.4104511	0.6406646	0.4104511	0.5257448	0.3917336
3	GBM_4_AutoML_20210111_183825	0.4109455	0.6410503	0.4109455	0.5252904	0.3912904
4	GBM_grid__1_AutoML_20210111_183825_model_3	0.4112253	0.6412685	0.4112253	0.5288300	0.3914531
5	StackedEnsemble_BestOfFamily_AutoML_20210111_183825	0.4120196	0.6418875	0.4120196	0.5321952	0.3926549
6	GBM_grid__1_AutoML_20210111_183825_model_5	0.4133891	0.6429534	0.4133891	0.5363356	0.3930037

Figura 2 - AutoML do Cenário 1

- Cenário 2: Neste cenário a equipa escolheu atributos que representem o suporte que os alunos têm fora da sala de aula:

  Atributos: “famsup","schoolsup", "paid", "Pstatus", "famrel"

Mais uma vez, com a utilização da função automl foi identificado que para este cenário seria ideal usar o modelo *General Linear Model* (GLM).

	model_id	mean_residual_deviance	rmse	mse	mae	rmsle
1	GLM_1_AutoML_20210111_183451	0.4186774	0.6470529	0.4186774	0.5470707	0.3963791
2	StackedEnsemble_AllModels_AutoML_20210111_183451	0.4200631	0.6481228	0.4200631	0.5458945	0.3972402
3	GBM_grid__1_AutoML_20210111_183451_model_5	0.4202381	0.6482577	0.4202381	0.5496434	0.3969777
4	GBM_grid__1_AutoML_20210111_183451_model_1	0.4213011	0.6490771	0.4213011	0.5488231	0.3978002
5	GBM_5_AutoML_20210111_183451	0.4215693	0.6492837	0.4215693	0.5491688	0.3975143
6	StackedEnsemble_BestOfFamily_AutoML_20210111_183451	0.4216691	0.6493606	0.4216691	0.5494348	0.3978959

Figura 3 - AutoML do Cenário 2

Para cada Cenário, deve ser então construído o modelo identificado pelo *automl* como sendo o modelo ideal para os atributos selecionados, e deve ser realizada a predição da nossa variável dependente, "Transition", de modo a identificar se o aluno Reprova (Transition=0) ou passa a disciplina (Transition=1).

3.4.2. Generate Test Design

Os passos seguidos para a construção do modelo foram os seguintes:

- Realização do Split do dataset em dataset de treino (*training_set*) e dataset de teste (*test_set*);
- Selecionar os atributos do dataset a serem utilizados como variáveis independentes (de acordo com o Cenário em questão) e a variável Target (*Transition*) como variável dependente;
- Construção do modelo (GBM para o cenário 1 e GLM para o Cenário 2) usando o dataset de treino;
- Realizar a previsão do dataset de teste (*test_set*) com o modelo criado;
- Apresentar e avaliar resultados.

É de salientar que foi realizada *Cross-Validation*, com o parâmetro k=10, ou seja, o dataset será dividido em 10 grupos.

3.4.3. Build Model

Cenário 1 – Modelo GBM

Para a construção do nosso modelo, apoiamo-nos dados recolhidos após o uso de AutoML, escolhemos um algoritmo de GBM(Gradient Boosting Machines). Este algoritmo cria uma árvore, árvore esta que é "treinada" e ao fim de cada "iteração" adiciona peso às diversas observações realizadas pela mesma.

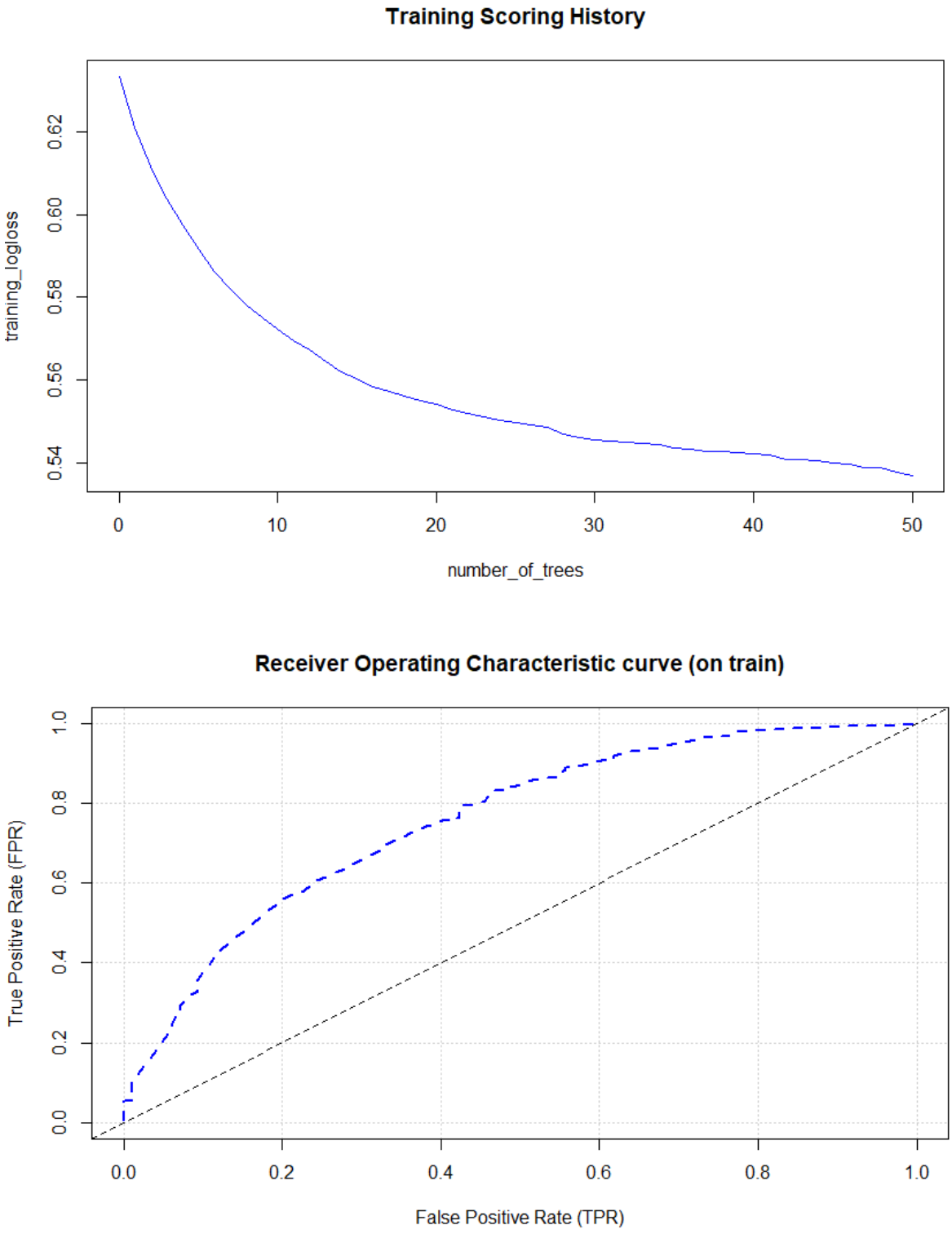

Gráfico 1- Modelo GBM: ROC Curve

Cenário 2 – Modelo GLM

Para a construção do nosso modelo, apoiamo-nos dados recolhidos após o uso de AutoML, escolhemos um algoritmo de GLM (Generalized Linear Model). Tal como o nome indica este algoritmo apoia-se em regressões, lineares, de Poisson ou binomiais. Com este algoritmo pretendíamos, usando os argumentos indicados anteriormente para prever o sucesso académico.

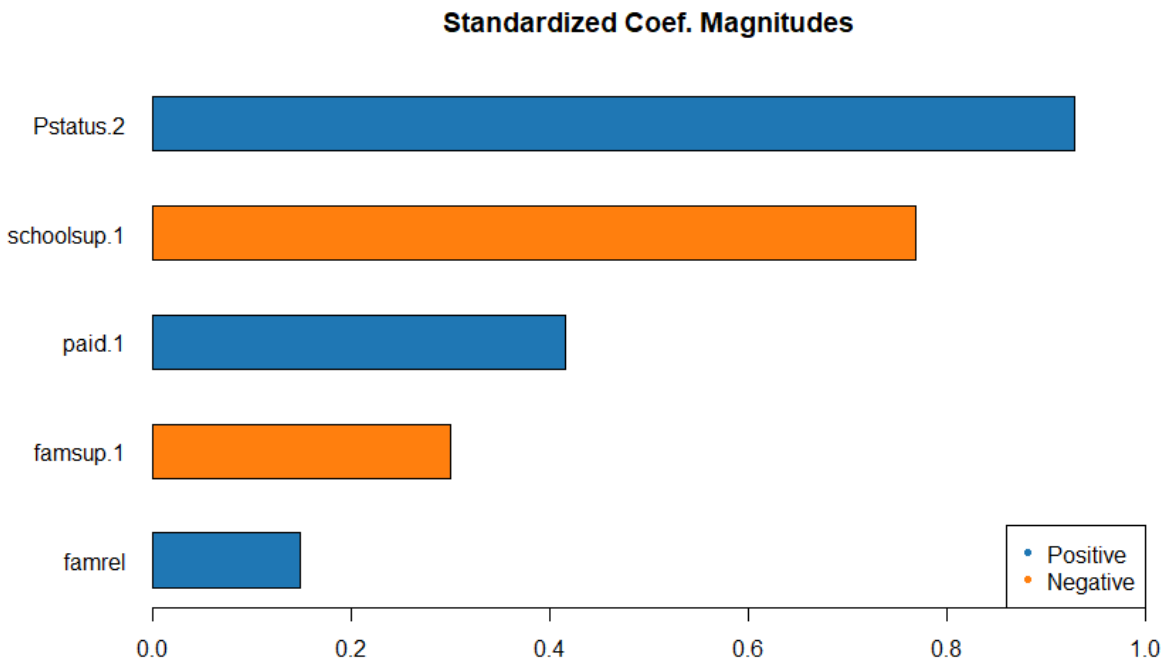

Gráfico 2 - Modelo GLM: Coefficient Magnitudes

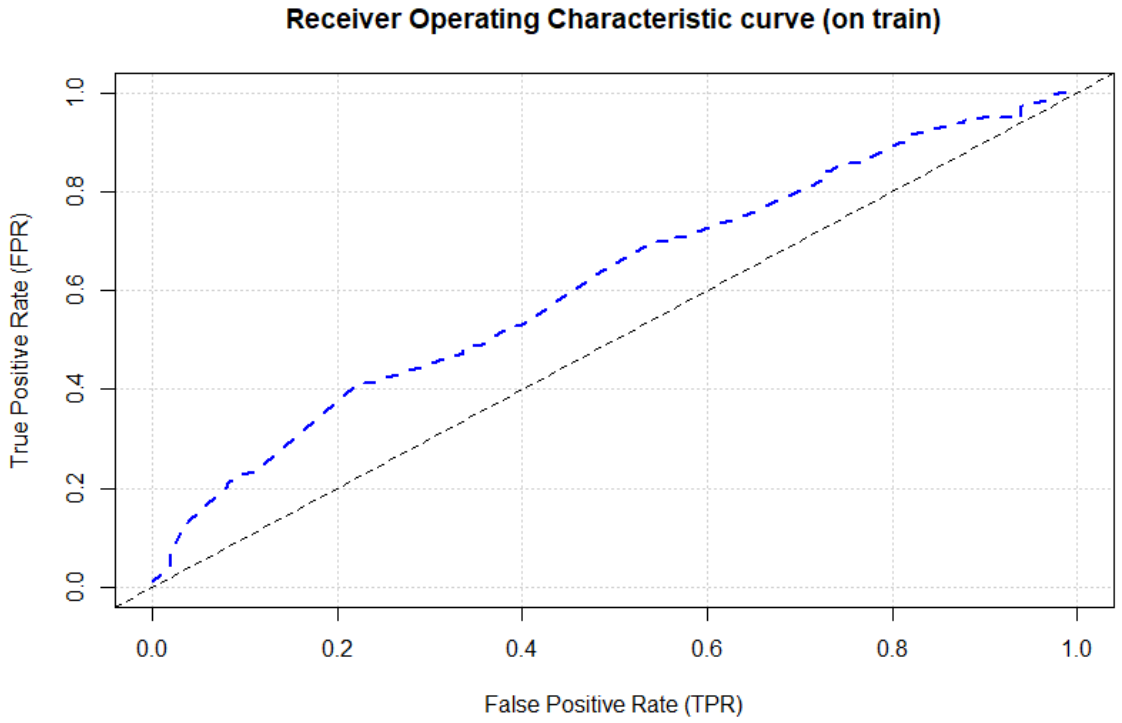

Gráfico 3 - Modelo GLM: ROC Curve

3.4.4. Assess Model

Cenário 1 – Modelo GBM

Após a construção do modelo foram verificadas as métricas obtidas que indicam a performance do mesmo. Na figura 1 estão demonstradas as principais métricas. De salientar que o modelo foi elaborado com uma cross-validation em 10 "folds".

A medida MSE que traduz a comparação entre os dados verificados e as previsões toma um valor bastante baixo. Foi possível então constatar através da medida MSE, que se trata de um modelo algo fiável porque quanto mais baixo for o valor da MSE mais previsões acertadas o modelo faz. Para fortalecer esta hipótese o modelo possui um valor AUC algo elevado também, que se traduz em valores preditivos verdadeiros positivos e verdadeiros negativos.

```
MSE:  0.178553
RMSE:  0.4225553
LogLoss:  0.537016
Mean Per-Class Error:  0.3862855
AUC:  0.7574196
AUCPR:  0.8512052
Gini:  0.5148391
R^2:  0.1909522
```

Figura 4 - Modelo GBM: Métricas nos dados de treino

```
Confusion Matrix (vertical: actual; across: predicted)
       fail pass    Error     Rate
fail     25   72 0.742268   =72/97
pass      6  192 0.030303   =6/198
Totals   31  264 0.264407  =78/295
```

Figura 5 - Modelo GBM: Matriz de Confusão

```
Maximum Metrics: Maximum metrics at their respective thresholds
                        metric threshold      value idx
1                        max f1  0.466458   0.831169  63
2                        max f2  0.293384   0.915033  69
3                  max f0point5  0.632666   0.794798  45
4                  max accuracy  0.578740   0.742373  56
5                 max precision  0.889880   1.000000   0
6                    max recall  0.220573   1.000000  74
7                max specificity  0.889880   1.000000   0
8               max absolute_mcc  0.632666   0.383191  45
9    max min_per_class_accuracy  0.697197   0.670103  31
10 max mean_per_class_accuracy  0.646477   0.685124  42
11                      max tns  0.889880  97.000000   0
12                      max fns  0.889880 197.000000   0
13                      max fps  0.220573  97.000000  74
14                      max tps  0.220573 198.000000  74
15                      max tnr  0.889880   1.000000   0
16                      max fnr  0.889880   0.994949   0
17                      max fpr  0.220573   1.000000  74
18                      max tpr  0.220573   1.000000  74
```

Figura 6 – Modelo GBM: Performance

Cenário 2 – Modelo GLM

Após a construção do modelo foram verificadas as métricas obtidas que indicam a performance do mesmo. Na figura 1 estão demonstradas as principais métricas

A medida MSE que traduz a comparação entre os dados verificados e as previsões toma um valor não tão baixo quanto o anterior. Foi possível verificar através da matriz de confusão (Figura 8) que apesar te o modelo obter uma boa taxa de acerto nos verdadeiros positivos, este fica aquém do esperado no número de verdadeiros negativos, obtendo assim uma taxa demasiado elevada de erro para a previsão dos alunos que reprovarão.

```
MSE:  0.2111566
RMSE:  0.4595178
LogLoss:  0.6116838
Mean Per-Class Error:  0.4663386
AUC:  0.6085859
AUCPR:  0.7485313
Gini:  0.2171717
R^2:  0.04322077
Residual Deviance:  360.8934
AIC:  372.8934
```

Figura 7 - - Modelo GLM: Métricas nos dados de treino

```
Confusion Matrix (vertical: actual; across: predicted)
       fail pass    Error     Rate
fail      8   89 0.917526   =89/97
pass      3  195 0.015152   =3/198
Totals   11  284 0.311864  =92/295
```

Figura 8 - - Modelo GLM: Matriz de Confusão

```
Maximum Metrics: Maximum metrics at their respective thresholds
                        metric threshold      value idx
1                        max f1  0.467287   0.809129  41
2                        max f2  0.382834   0.912442  43
3                  max f0point5  0.642066   0.734463  29
4                  max accuracy  0.467287   0.688136  41
5                 max precision  0.907266   1.000000   0
6                    max recall  0.382834   1.000000  43
7                max specificity  0.907266   1.000000   0
8               max absolute_mcc  0.642066   0.179635  29
9    max min_per_class_accuracy  0.682302   0.536082  23
10 max mean_per_class_accuracy  0.659224   0.590857  25
11                      max tns  0.907266  97.000000   0
12                      max fns  0.907266 196.000000   0
13                      max fps  0.342813  97.000000  44
14                      max tps  0.382834 198.000000  43
15                      max tnr  0.907266   1.000000   0
16                      max fnr  0.907266   0.989899   0
17                      max fpr  0.342813   1.000000  44
18                      max tpr  0.382834   1.000000  43
```

Figura 9 – Modelo GLM: Performance

3.5. Evaluation

3.5.1. Evaluate Results

No decorrer da aplicação de ambos algoritmos, deparamo-nos com alguns problemas quanto ao dataset "*Student-por.csv*", pois tendo em conta que não balanceamos os dados, e segundo pudemos aferir, de todo o dataset, apenas 100 alunos reprovaram a disciplina e supomos assim que é por isto que o erro obtido é substancialmente maior do que o pretendido.

Podemos concluir então, que apesar da acuidade (em ambos os cenários) não ser a pretendida, a especificidade e a precisão foram bastante altas, tal como as áreas de ROC.

Tabela 4 - Resultados Finais

Modelo	**Métricas**	**Resultado Final**
GBM	Accuracy:57.9% Precision:88.9% Specificity:88.9% Área de ROC: 75%	Aceitável
GLM	Accuracy:46.17% Precision:90.7% Specificity:90.7% Área de ROC: 60%	Insatisfatório

3.5.2. Review Process

Como referido anterior, acreditamos que a não realização do balanceamento dos dados (ou por *UnderSampling* ou por *OverSampling*) afetou drasticamente os resultados.

Por esta razão, deverá ser realizada uma revisão do projeto onde deverá ser realizado um balanceamento dos dados.

4. Anexos

Contrato

Universidade do Minho

Contrato Grupo 6 DMCD 2020/2021

- Garante-se que não se irão falsificar resultados, nem copiar/plagiar projetos (de outros grupos) ou conteúdos da Internet sem que estes sejam devidamente identificados e referenciados (quem é o autor, onde foi publicado) e esta utilização não seja exagerada (face ao restante trabalho desenvolvido).
- Este grupo compromete-se a entregar materiais genuínos e com o maior rigor possível.

X A. Macedo Marta Isabel Marinho Andrade

André Macedo — Marta Isabel

X Nelson Almeida X Pedro Oliveira

Nelson Almeida — Pedro Oliveira

Análise Descritiva dos Dados

Na imagem abaixo esta apresentada a análise dos atributos do Dataset *student_por.csv* referente a disciplina de Português.

```
> summary(student_por)
    school              sex                 age           address             famsize            Pstatus              Medu            Fedu
 Length:649         Length:649         Min.   :15.00   Length:649         Length:649         Length:649         Min.   :0.000   Min.   :0.000
 Class :character   Class :character   1st Qu.:16.00   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:1.000
 Mode  :character   Mode  :character   Median :17.00   Mode  :character   Mode  :character   Mode  :character   Median :2.000   Median :2.000
                                       Mean   :16.74                                                         Mean   :2.515   Mean   :2.307
                                       3rd Qu.:18.00                                                         3rd Qu.:4.000   3rd Qu.:3.000
                                       Max.   :22.00                                                         Max.   :4.000   Max.   :4.000
     Mjob              Fjob               reason            guardian          traveltime      studytime        failures       schoolsup
 Length:649         Length:649         Length:649         Length:649         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:649
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character
                                                                            Mean   :1.569   Mean   :1.931   Mean   :0.2219
                                                                            3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                                                                            Max.   :4.000   Max.   :4.000   Max.   :3.0000
   famsup              paid            activities          nursery            higher            internet          romantic            famrel
 Length:649         Length:649         Length:649         Length:649         Length:649         Length:649         Length:649         Min.   :1.000
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.000
                                                                                                                                     Mean   :3.931
                                                                                                                                     3rd Qu.:5.000
                                                                                                                                     Max.   :5.000
    freetime        goout            Dalc            Walc           health         absences            G1              G2              G3
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
 1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
 Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median :4.000   Median : 2.000   Median :11.0   Median :11.00   Median :12.00
 Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536   Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
>
```

Figura 6 - Análise dos atributos do dataset student_por.csv

Na imagem a seguir esta apresentada a análise dos atributos do Dataset *student_mat.csv* referente a disciplina de Matemática.

```
> summary(student_mat)
    school              sex                 age           address             famsize            Pstatus              Medu            Fedu
 Length:395         Length:395         Min.   :15.0   Length:395         Length:395         Length:395         Min.   :0.000   Min.   :0.000
 Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
 Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character   Mode  :character   Median :3.000   Median :2.000
                                       Mean   :16.7                                                         Mean   :2.749   Mean   :2.522
                                       3rd Qu.:18.0                                                         3rd Qu.:4.000   3rd Qu.:3.000
                                       Max.   :22.0                                                         Max.   :4.000   Max.   :4.000
     Mjob              Fjob               reason            guardian          traveltime      studytime        failures       schoolsup
 Length:395         Length:395         Length:395         Length:395         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character
                                                                            Mean   :1.448   Mean   :2.035   Mean   :0.3342
                                                                            3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                                                                            Max.   :4.000   Max.   :4.000   Max.   :3.0000
   famsup              paid            activities          nursery            higher            internet          romantic            famrel
 Length:395         Length:395         Length:395         Length:395         Length:395         Length:395         Length:395         Min.   :1.000
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.000
                                                                                                                                     Mean   :3.944
                                                                                                                                     3rd Qu.:5.000
                                                                                                                                     Max.   :5.000
    freetime        goout            Dalc            Walc           health         absences            G1              G2              G3
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
 Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median : 4.000   Median :11.00   Median :11.00   Median :11.00
 Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
>
```

Figura 7 - Análise dos atributos do dataset student_mat.csv

Verificamos que não existem “missing values” nem valores nulos em qualquer um dos datasets.

```
> sum(is.na(student_por))
[1] 0
> sum(is.na(student_mat))
[1] 0
>
```

Figura 9 - Análise de "Missing Values"

```
> sum(is.null(student_por))
[1] 0
> sum(is.null(student_mat))
[1] 0
>
```

Figura 8 - Análise de valores nulos

KPI'S

Nesta seção são apresentadas as análises preliminares feitas aos dados de forma a permitir uma melhor compreensão destes. É de salientar que os gráficos abaixo podem não estar diretamente relacionados com o objetivo de negócio escolhido, sendo estes usados para orientação ao dataset.

- Relação do género dos estudantes com as ausências e horas dedicadas ao estudo: verificamos uma diferenciação nas horas dedicadas ao estudo, pois enquanto que os rapazes estudam geralmente menos de 2 horas, raparigas estudam geralmente 2 a 10 horas. Não verificamos diferenciação de género no número de faltas, sendo o número destas geralmente menores que 20.

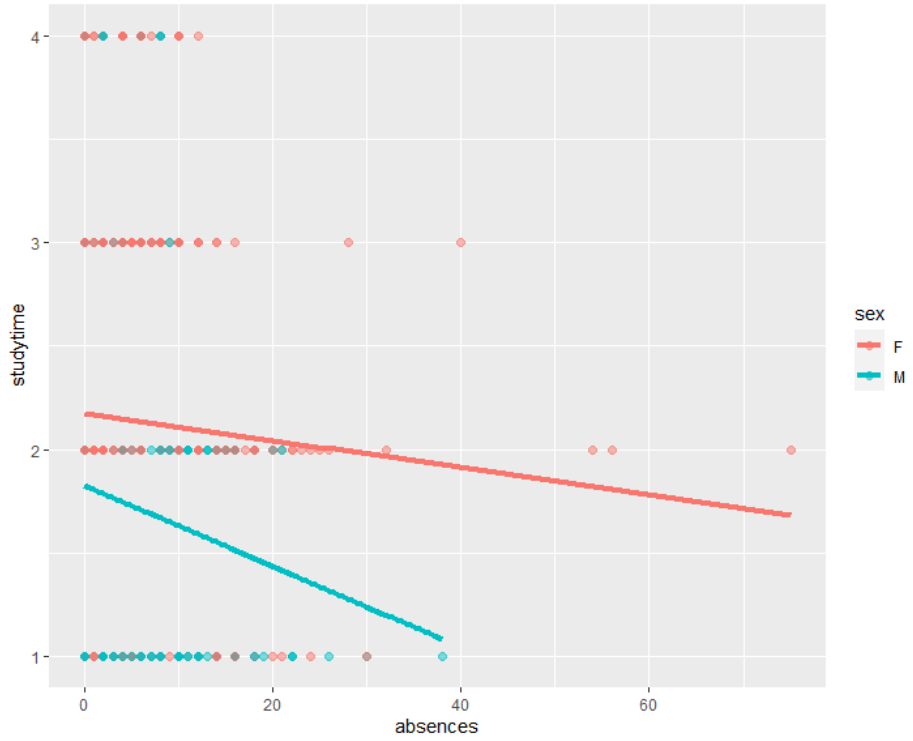

Gráfico 4 - Gráfico de relação do genero dos estudantes com as ausências e horas dedicadas ao estudo

- Relação entre o tempo de estudo com a nota final e o estado romântico: verificamos que apesar do número de horas de estudo serem diferentes, existe na mesma uma grande variedade de notas, desde as mais baixas as mais altas, ou seja, o número de horas de estudo não influencia drasticamente a nota final. Verificamos também que no caso dos alunos que estudam mais de 10 horas, quase todos eles não se encontram numa relação.

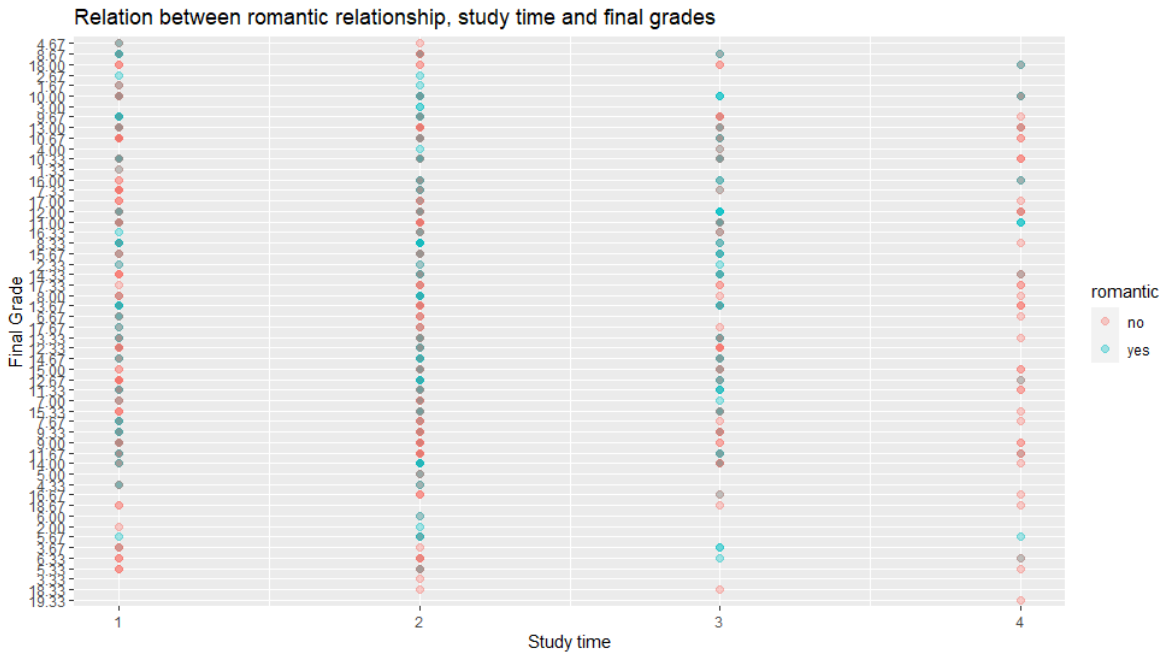

Gráfico 6 - Relação entre o tempo de estudo com a nota final e o estado romântico

- Correlação entre a Nota Final e o Numero de Faltas e Chumbos: verificamos que a maior parte dos alunos não possui qualquer chumbo anterior e que o número de faltas é maioritariamente menor que 20 e concentram-se no tempo de estudo de menos de 2 horas. Já a nota final concentra-se, tal como as faltas, no tempo de estudo de menos de 2 horas, e as notas finais concentram-se maioritariamente entre 9 e 17 valores.

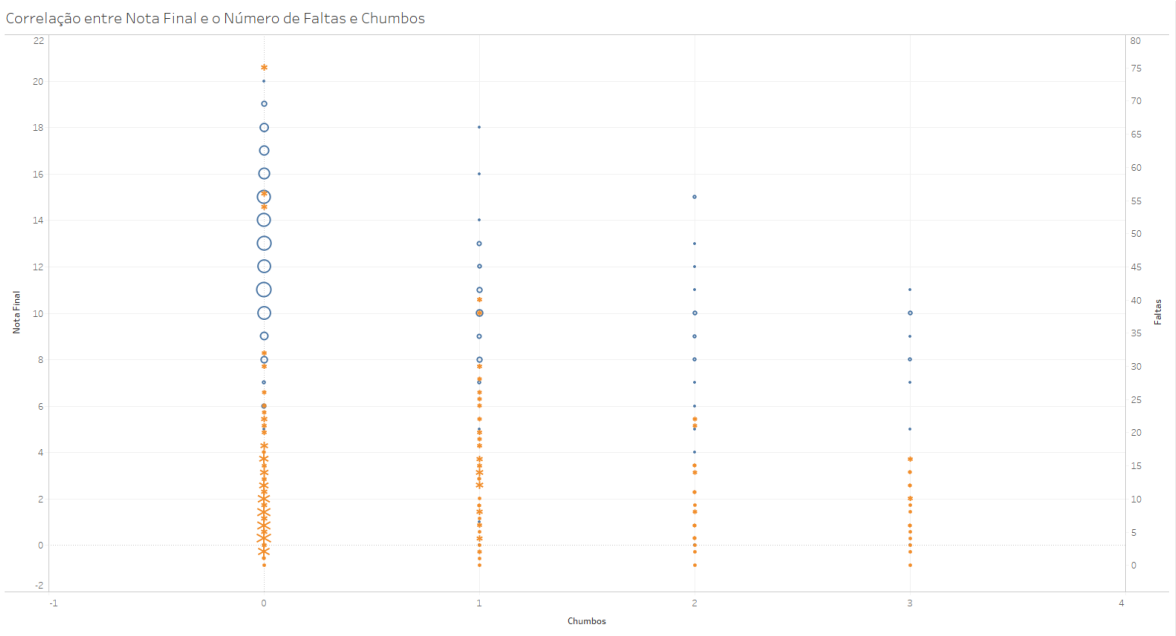

Gráfico 5 - Correlação entre a Nota Final e o Numero de Faltas e Chumbos

- Influência do Suporte da Escola e de explicações na nota final: verificamos que a maior parte dos alunos não recebe apoio da escola nem têm explicações extra, seguidos pelos alunos que apesar de não receberem apoio da escola, têm explicações. Chegamos então a conclusão que a frequência de explicações ou a existência de suporte por parte da escola não influência drasticamente a nota final.

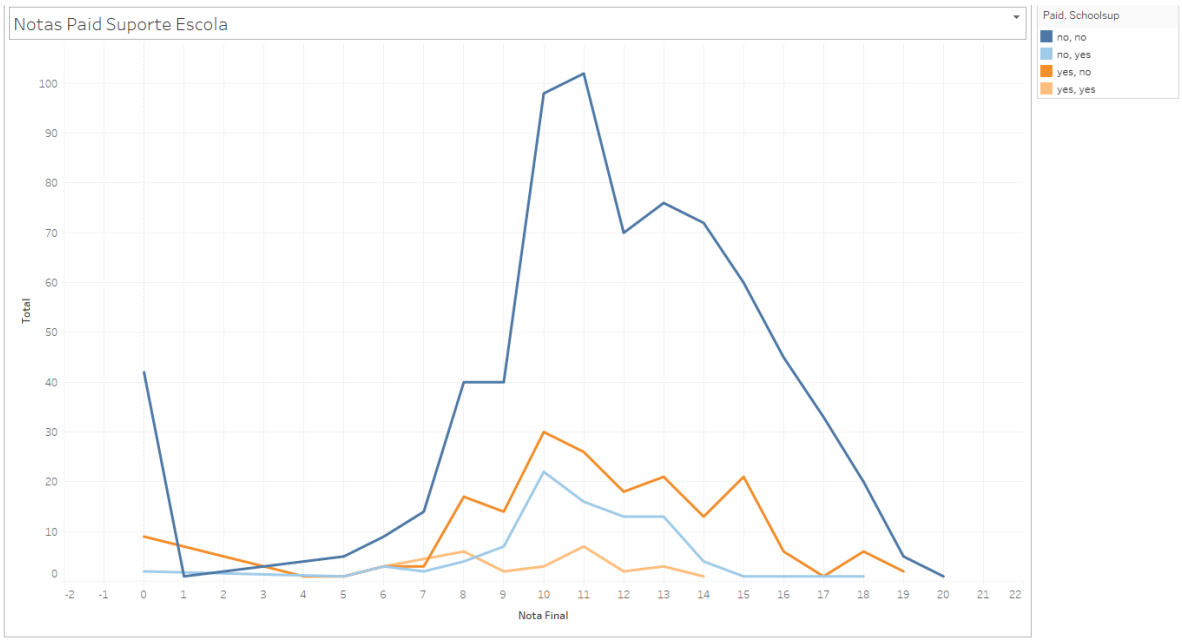

Gráfico 7 - Influência do Suporte da Escola e de explicações na nota final

- Influência das Profissões dos Pais na nota final: verificamos que, no caso da profissão da mãe, a maior parte dos alunos têm "other", seguido por "services". Verificamos também que apesar de existirem alunos cuja mãe tem a profissão de "teacher", isto não significa que irão ter as melhores notas.

  No caso da profissão do pai, verificamos novamente que a maior parte dos alunos têm o pai com a profissão "other", seguido por "services", mas no caso do pai, a descrepância entre estas duas profissões e as restantes é maior. Tal como aconteceu com a profissão da mãe, também com a profissão do pai, este ser "teacher" não significa que irão ter as melhores notas.

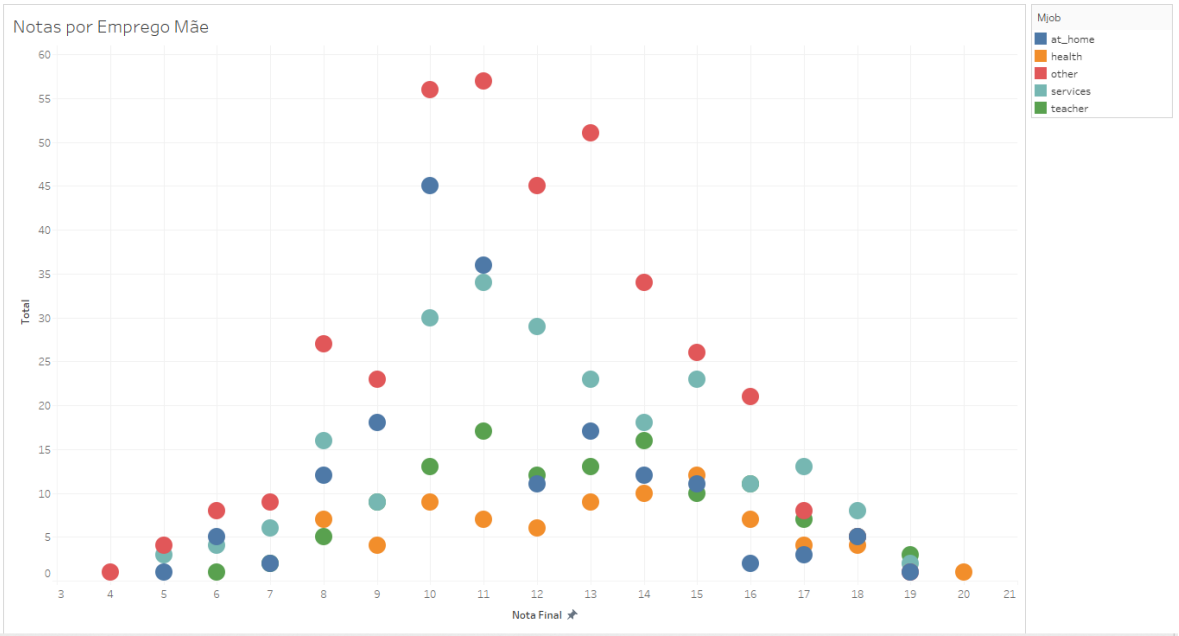

Gráfico 8 - Influência da Profissão da Mãe na nota final

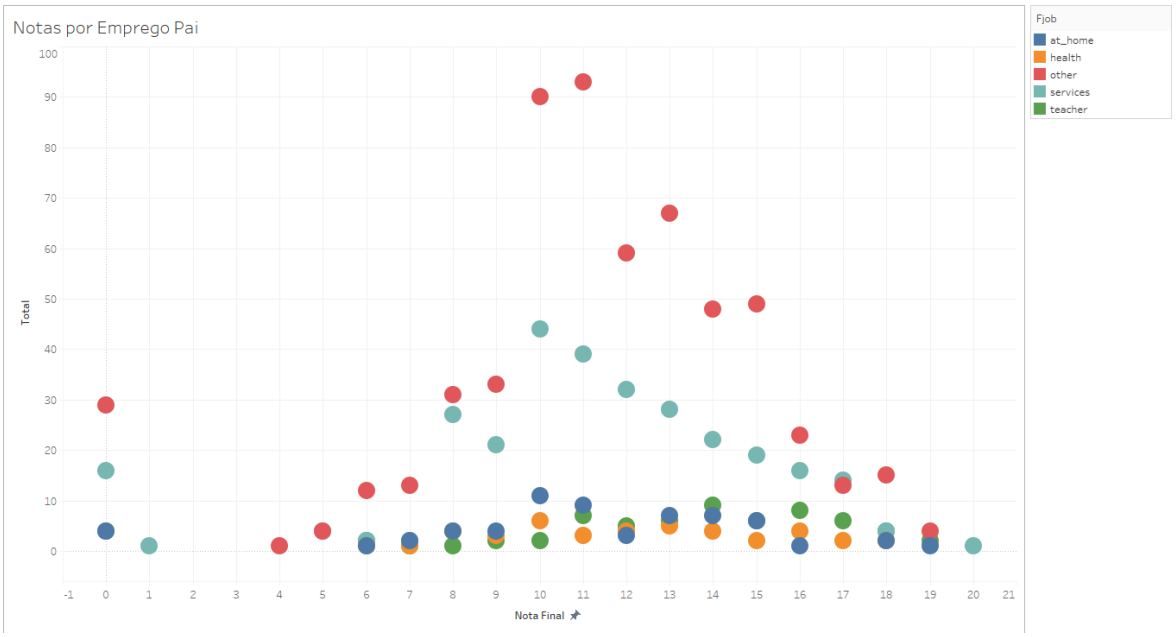

Gráfico 9 - Influência da Profissão do Pai na nota final

- Correlação entre Saídas e Consumo de Álcool: verificamos que tanto para o consumo de álcool a semana com ao fim-se-semana, os alunos tendem a ter a mesma frequência de saídas na classificação 3. Verificamos também que ao fim-de-semana existe uma maior distribuição dos valores de consumo de álcool comparativamente a semana, que estão mais concentrados na classificação 2 e 3 da frequência de saídas e no consumo "Muito pouco" de álcool. Ou seja, os alunos bebem mais e saem mais ao fim de semana. Não foi identificado que com o aumento do número de saídas, aumenta o consumo de álcool.

Gráfico 10 - Correlação entre Saídas e Consumo de Álcool

Association

Foi usado o modelo **Apriori**, e foi estabelecido um **Support** de 0.6 e **Confidence** de 0.8, o que gerou a identificação de 265 regras.

Na figura abaixo apresentada encontram-se as 10 regras com os melhores valores de Lift. Podemos verificar então, por exemplo, que 80% dos registos com idade até 17 anos e que pretendem seguir estudos para a universidade não apresentam chumbos.

```
     lhs                    rhs                 support confidence  coverage     lift count
[1]  {age=-17,
      higher=yes,
      absences=-20}      => {failures=none} 0.6034483  0.8987161 0.6714559 1.089732   630
[2]  {age=-17,
      higher=yes}        => {failures=none} 0.6130268  0.8938547 0.6858238 1.083838   640
[3]  {age=-17,
      absences=-20}      => {failures=none} 0.6273946  0.8887381 0.7059387 1.077634   655
[4]  {age=-17}           => {failures=none} 0.6369732  0.8843085 0.7203065 1.072263   665
[5]  {higher=yes,
      internet=yes,
      absences=-20}      => {failures=none} 0.6264368  0.8778523 0.7136015 1.064434   654
[6]  {address=U}         => {internet=yes}  0.6111111  0.8405797 0.7270115 1.061143   638
[7]  {higher=yes,
      internet=yes}      => {failures=none} 0.6369732  0.8704188 0.7318008 1.055421   665
[8]  {Pstatus=T,
      schoolsup=no,
      higher=yes,
      absences=-20}      => {failures=none} 0.6091954  0.8700410 0.7001916 1.054963   636
[9]  {age=-17,
      failures=none}     => {higher=yes}    0.6130268  0.9624060 0.6369732 1.052096   640
[10] {age=-17,
      failures=none,
      absences=-20}      => {higher=yes}    0.6034483  0.9618321 0.6273946 1.051469   630
```

Figura 10 - Regras de Associação

Podemos depois relacionar estas regras encontradas com as avaliações, como representado na tabela abaixo que demonstra 3 exemplos.

Tabela 5 - Relação entre as Regras de Associação e a Avaliação

Regra	Taxa de Reprovações	Taxa de Avaliações Satisfatórias	Taxa de Avaliações Excelentes
$\{\boldsymbol{Idade\ até\ 17\ anos}, \boldsymbol{Quer\ prosseguir\ estudos}, \boldsymbol{Até\ 20\ faltas}\} \rightarrow [\boldsymbol{Sem\ Chumbos}\}$	$\frac{153}{630}$*100=24.3%	$\frac{390}{630}$*100=61.9%	$\frac{87}{630}$*100=13.8%
$\{\boldsymbol{Morada\ Urbana}\} \rightarrow \{\boldsymbol{Tem\ Internet}\}$	$\frac{195}{638}$*100=30.6%	$\frac{352}{638}$*100=55.2%	$\frac{91}{638}$*100=14.3%
$\{\boldsymbol{Idade\ até\ 17\ anos}, \boldsymbol{Sem\ chumbos}\} \rightarrow \{\boldsymbol{Pretende\ seguir\ estudos}\}$	$\frac{159}{640}$*100=24.8%	$\frac{394}{640}$*100=61.6%	$\frac{87}{640}$*100=13.6%

Clustering

Foi realizado um estudo de forma a agrupar as idades dos alunos e o número de Chumbos destes, de modo a identificar possíveis agrupamentos.

Primeiramente foi utilizado o *Elbow Method* de forma a identificar o número ideal de clusters para este cenário.

Podemos ver pela figura abaixo que apesar de haver uma certa discrepância, o número ideal de clusters seria 4.

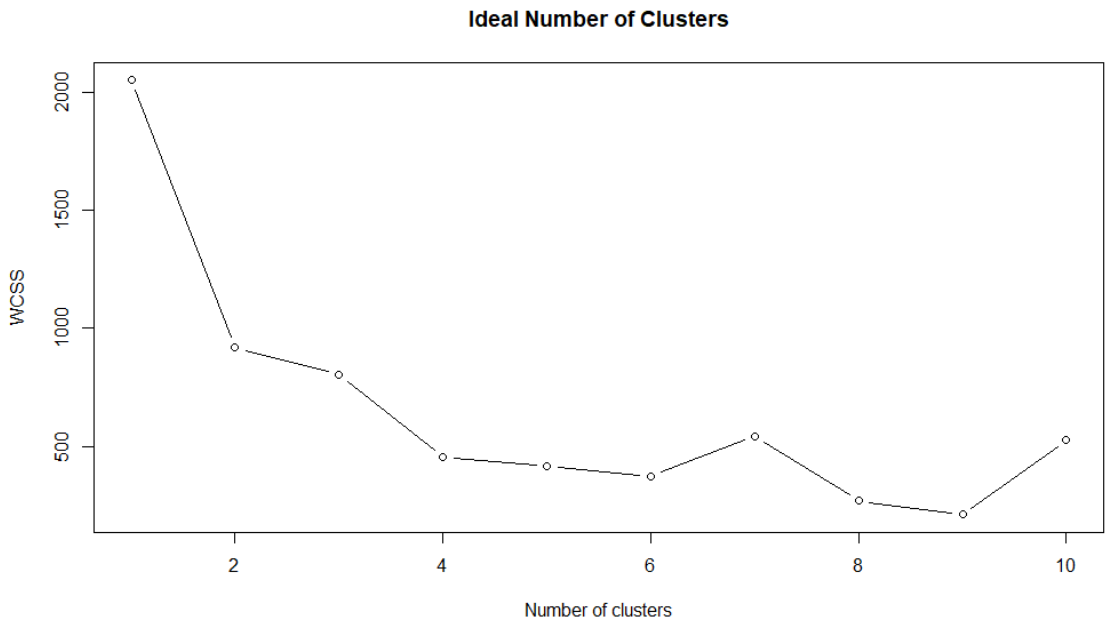

Gráfico 11 - Determinação do Número de Cluster usando o Elbow Method

Com a identificação feita do número ideal de clusters, utilizando kmeans foi criado então o gráfico abaixo apresentado com a representação dos clusters. A tabela a seguir representada realiza uma caracterização dos Clusters.

Tabela 6 - Caracterização dos Clusters

	Cluster 1	Alunos pertencentes a este cluster são alunos novos e sem um número de chumbos reduzido.
Cluster 2	Este cluster apresenta a maior variação de valores de Idade e Chumbos, contendo os alunos que são mais velhos com um grande número de chumbos.	
Cluster 3	Alunos pertencentes ao Cluster 3 apesar de terem aproximadamente a mesma idade que os alunos do Cluster 1, estes apresentam um maior nivel de Chumbos.	
Cluster 4	Semelhante ao Cluster 1 em termos de quantidade de chumbos, mas alunos pertencentes a este Cluster são mais velhos que os alunos do Cluster 1	

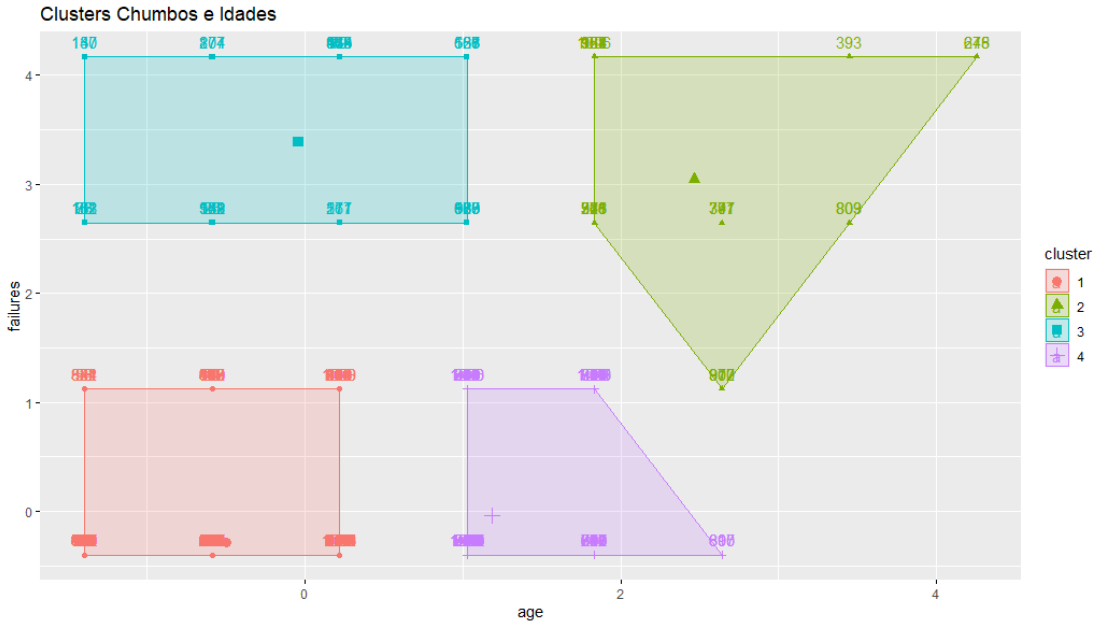

Figura 11 - Cluster Chumbos e Idades

Uma das conclusões que podemos tirar é que alunos pertencentes aos Clusters 1 e 4 não necessitam de apoio escolar pois até a data, apresentam um bom desempenho escolar apesar de serem novos. Já tanto o Cluster 3 como o Cluster 4 necessitam de apoio pois apresentam um número de Chumbos maior que nos outros dois clusters, independentemente da idade.