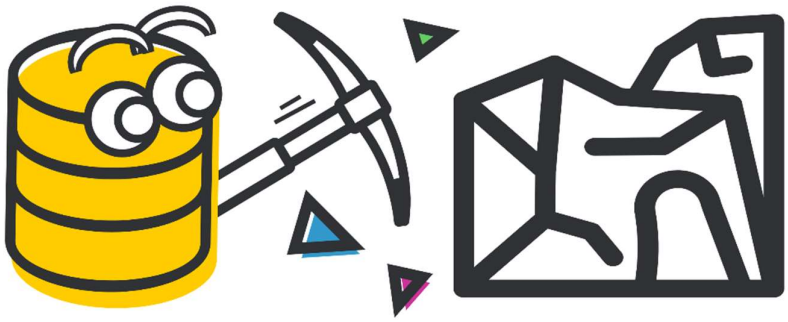
# Introdução à Ciência de Dados

Análise circunstancial das notas de matemática e português




Trabalho realizado por:

Fabian Gobet nº97885

Carlos Esteves nº98004

João Correia nº94576

David Rosa nº98359


November 20, 2022

# Conteúdo



# Índice de Figuras

# Introdução

No âmbito da unidade curricular Introdução à Ciência de Dados foi-nos proposto este projeto com base num conjunto de dados referentes a avaliações de estudantes no ensino secundário nas disciplinas de matemática e português.

Este projeto serve como forma de aplicar os conhecimentos adquiridos ao longo da unidade curricular, nomeadamente a utilização de ferramentas como o Orange para limpar, processar e analisar os dados fornecidos. Cada vez mais recorremos à análise de dados para retirar informação que tem como objetivo motivar uma escolha ou um plano ótimo.

Posto isto, com base na metodologia CRISP-DM, criámos um modelo que tem como alvo a nota final do aluno, que visa responder a perguntas sobre como as circunstâncias do aluno podem afetar a mesma.

Algumas das perguntas a que tentámos dar resposta foram, nomeadamente:

- relação entre a dinâmica familiar e o desempenho geral do estudante
    - estudo da performance do aluno quando os pais estão juntos vs. quando os pais separados, e quando separados, sobre a guarda de quem.
    - influência da escolaridade do pai e da mãe nos géneros
    - consumo de substâncias e saúde
- relação entre desempenho do estudante e géneros nas disciplinas diferentes
- relação entre desempenho do estudante, mobilidade e urbanização de residência

## Descrição do conjunto de dados

| Nome da Variável | Descrição da Variável |
|---|---|
| school | Escola do estudante ('GP' - Gabriel Pereira ou 'MS' - Mouzinho da Silveira) |
| sex | Sexo do estudante ('F' - feminino ou 'M' - masculino) |
| age | Idade do estudante (de 15 a 22) |
| address | Tipo de casa do estudante ('U' - urbano ou 'R' - rural) |
| famsize | Tamanho da família ('LE3' – menor ou igual a 3 ou 'GT3' – maior que 3) |
| Pstatus | Estado de coabitação dos pais ('T' – a viver juntos ou 'A' - separados) |
| Medu | Educação da mãe (0 - nenhuma, 1 – educação primária (4ºano), 2 - 5º ao 9ºano, 3 educação secundária (9ºano) ou 4 educação superior) |
| Fedu | Educação do pai 0 - nenhuma, 1 – educação primária (4ºano), 2 - 5º ao 9ºano, 3 educação secundária (9ºano) ou 4 educação superior) |
| Mjob | Trabalho da mãe ('teacher', 'health', 'services', 'at_home' ou 'other') |
| Fjob | Trabalho do pai ('teacher', 'health', 'services', 'at_home' ou 'other') |
| reason | Razão para escolher esta escola ('home' (perto de casa), 'reputation' (reputação da escola), 'course' (preferência de curso) ou 'other') |
| guardian | Guardião do estudante ('mother', 'father' ou 'other') |
| traveltime | Tempo de deslocação (1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - >1 hora) |
| studytime | Tempo de estudo semanal (1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, ou 4 - >10 horas) |
| failures | Quantidade de chumbos (n se 1<=n<3, caso contrário 4) |
| schoolsup | Suporte educacional extra (sim ou não) |
| famsup | Suporte educacional da família (sim ou não) |
| paid | Aulas extra (Português ou Matemática) (sim ou não) |
| activities | Atividades extra-curriculares (sim ou não) |
| nursery | Frequentou infantário (sim ou não) |
| higher | Quer frequentar ensino superior (sim ou não) |
| internet | Acesso à Internet em casa (sim ou não) |
| romantic | Com relação amorosa (sim ou não) |
| famrel | Qualidade de relações familiares (de 1 – muito mau a 5 – excelente) |
| freetime | Tempo livre pós escola (de 1 – muito pouco a 5 – muito alto) |
| goout | Sai com os amigos (de 1 – poucas vezes a 5 – muitas vezes) |
| Dalc | Consumo de álcool em dias de semana (de 1 – muito pouco a 5 – muito alto) |
| Walc | Consumo de álcool em fim de semana (de 1 – muito pouco a 5 – muito alto) |
| health | Estado de saúde atual (de 1 – muito mau a 5 – excelente) |
| absences | Número de faltas à escola (de 0 a 93) |
| G1 | Nota do primeiro período (de 0 a 20) |
| G2 | Nota do segundo período (de 0 a 20) |
| G3 | Nota final (de 0 a 20) |

Não foram encontradas faltas de dados no conjunto de dados.

## Preparação dos dados

Para grande parte dos nossos estudos foi necessário concatenar as tabelas. Como estas eram referentes a disciplinas diferentes, português e Matemática, não achámos necessário ter em conta a possível repetição de estudantes, ainda que esta análise pudesse ser feita tendo em conta todos os atributos utilizados exceto G1, G2 e G3.

Para além da concatenação, para efeitos de teste e previsão foi necessário excluir os outliers relativos à variável alvo G3. Para tal calculámos a amplitude interquartil (Q1 = 10, Q2 = 14, IQA = 4) e filtrámos o dataset considerando apenas os estudantes com nota G3 igual ou superior 4.

Para podermos fazer a previsão, e o uso de árvores de decisão e regressão linear, foi ainda necessário categorizar as variáveis G3, Fedu, Medu e ainda Famrel, que antes eram numéricas. Para a variável G3, foi decidido dividir em apenas dois valores "Passou" e "Não Passou", para valores de G3 menores que 10, e caso contrário, respetivamente, constante com o nome G3Cat. As variáveis Fedu e Medu foram divididas de forma semelhante, sendo apenas necessário criar a variável do tipo categórico, visto que estas eram numéricas, mas apenas tinham valores inteiros de 1 a 5. Por fim, a variável Famrel teve um processamento igual ao descrito anteriormente para as variáveis Fedu e Medu.

## Análise exploratória dos dados, Modelação e Conclusões

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | C | Mjob | 5.0 | nan | 0.36887913680999496 | 17 | C | paid | 2.0 | nan | 0.16400825024316437 |
| 2 | C | reason | 4.0 | nan | 0.3590657682553364 | 18 | N | G1 | | nan | 0.16138300477133852 |
| 3 | C | Fjob | 5.0 | nan | 0.2788122963350607 | 19 | C | nursery | 2.0 | nan | 0.160796834627274 |
| 4 | N | goout | | nan | 0.24129216041236753 | 20 | N | freetime | | nan | 0.1596412902366237 |
| 5 | N | health | | nan | 0.22536849249676377 | 21 | C | sex | 2.0 | nan | 0.15482838819540096 |
| 6 | N | G2 | | nan | 0.2226829903112624 | 22 | C | famsize | 2.0 | nan | 0.15201165014375512 |
| 7 | C | Dalc | 5.0 | nan | 0.21061991759163134 | 23 | N | failures | | nan | 0.1377426886291632 |
| 8 | C | address | 2.0 | nan | 0.20827352357838422 | 24 | N | famrel | | nan | 0.137549477957199 |
| 9 | C | romantic | 2.0 | nan | 0.20587473617637717 | 25 | N | traveltime | | nan | 0.12965705398248442 |
| 10 | C | activities | 2.0 | nan | 0.20369762201808558 | 26 | N | age | | nan | 0.1288215092320948 |
| 11 | C | guardian | 3.0 | nan | 0.19428624487457397 | 27 | C | school | 2.0 | nan | 0.12688096342299027 |
| 12 | N | studytime | | nan | 0.19353505426118778 | 28 | C | internet | 2.0 | nan | 0.12262314666082581 |
| 13 | N | Medu | | nan | 0.18221933825081768 | 29 | C | Pstatus | 2.0 | nan | 0.08005240867257987 |
| 14 | C | famsup | 2.0 | nan | 0.17643356913620098 | 30 | C | schoolsup | 2.0 | nan | 0.0711017910702348 |
| 15 | N | Walc | | nan | 0.1715521504368583 | 31 | N | absences | | nan | 0.06442305645323887 |
| 16 | N | Fedu | | nan | 0.1699522609002611 | 32 | C | higher | 2.0 | nan | 0.04596472960821026 |

*Figura 1. Ranks das varáveis c/ G3*

## Pergunta 1

### Subtema 1

A primeira questão a analisar prende-se com o desempenho do estudante mediante a sua dinâmica familiar, mais precisamente o estado matrimonial dos pais e a guarda relativa ao aluno aquando da separação. Para tal foi necessário fazer uso das variáveis PStatus, Guardian e G3Cat, aplicando uma distribuição com alvo G3Cat e divisão por PStatus (Figura 2)

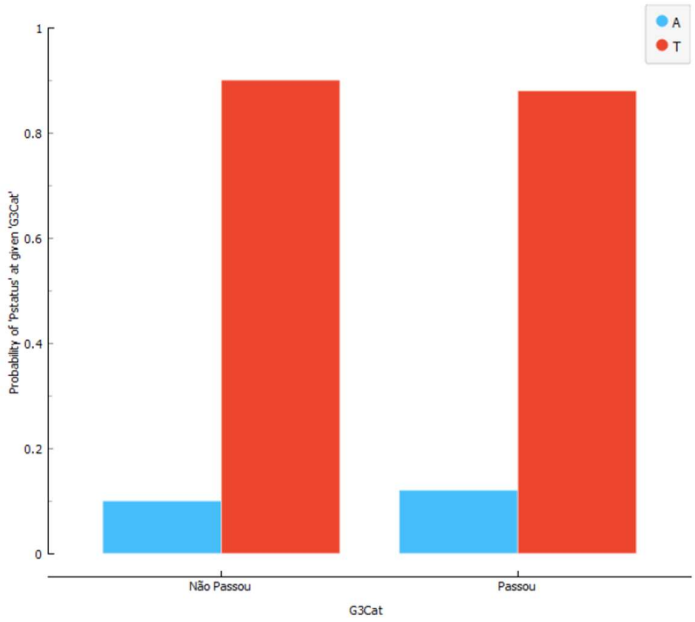



*Figura 2. Pais separados vs. juntos*

Com estes resultados concluímos que não existe evidente correlação entre o estado matrimonial dos pais e o desempenho do aluno. Isto é, os alunos de pais separados que não passaram encontram-se em número semelhante ao número de alunos de pais separados que passaram. Analogamente, a mesma conclusão podemos aferir aos alunos cujos pais estejam juntos.

No entanto, podemos verificar que existe uma diferença notória entre o número de alunos com pais separados e alunos com pais juntos.

Atendendo ao conjunto de alunos com pais separados achámos pertinente fazer a mesma avaliação tendo em conta o respetivo guardião. (Figura 3)

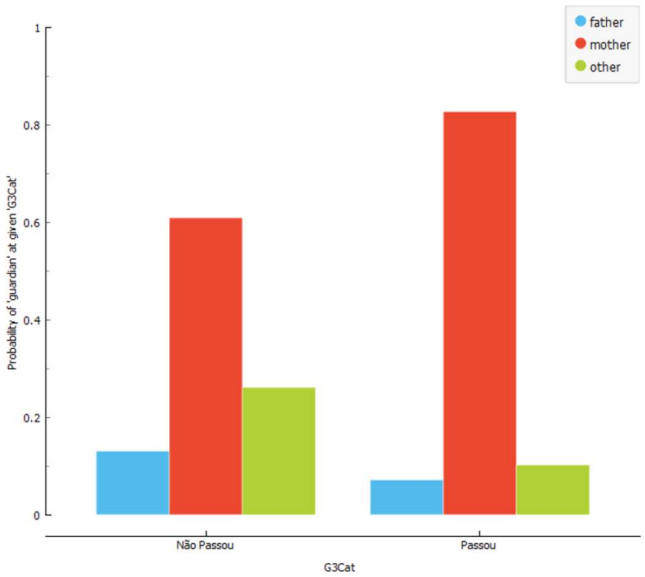



*Figura 3. Pais separados - guardiões*

Quando o guardião é o pai existe uma pequena diferença na frequência relativa entre os alunos que não passaram e os que passaram, sendo esta menor no último caso. Estes dados levantam a suspeita que o aluno sob guarda do pai tem mais tendência a não passar.

Por outro lado, sendo o guardião a mãe, existe uma maior diferença na frequência relativa entre os alunos que não passaram e os que passaram, sendo esta maior no último caso. Estes dados são indicativos de que os alunos sob guarda da mãe têm maior tendência a passar.

Por fim, o caso em que a guarda são outros, tem uma análise análoga à da guarda do pai.

Podemos também verificar que na separação dos pais o maior detentor da guarda é a mãe, depois outros, e por fim o pai.

Vale a pena notar que a análise dos pais separados se fundamenta em apenas 121 instancias das 1044 originais.

## Subtema 2

A segunda questão a analisar prende-se com o desempenho do estudante mediante o seu género e o nível de educação dos pais. Para tal foi necessário fazer uso das variáveis Medu, Fedu e Sex, aplicando uma distribuição com alvo G3Cat e divisão por Medu/Fedu (Figura 4 e 5).

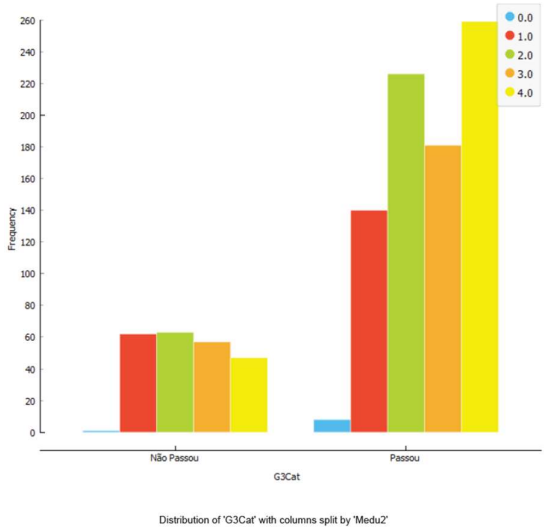



*Figura 4. Desempenho c/ Nivel educação Mãe*

Analisando o gráfico obtido (Figura 4) podemos observar que à medida que a educação da mãe cresce a frequência de alunos que passa tende a aumentar, e consequentemente os que não passam tende a diminuir.

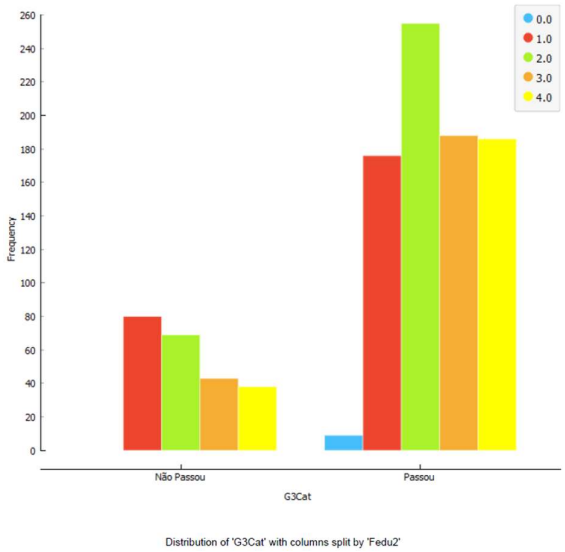



*Figura 5.Desempenho c/ Nivel educação Pai*

Analisando o gráfico obtido (Figura 5) podemos observar que à medida que a educação da pai cresce a frequência de alunos que não passa tende a diminuir. Por outro lado, Á medida que a

educação do pai aumenta, a frequência de alunos que passa segue uma distribuição normal com média no nível 3 da educação do pai.

Em ambos os gráficos denotamos uma maior quantidade de alunos que passaram e uma quantidade reduzida (negligenciável perante o nosso problema) de pais com nível de educação 0.

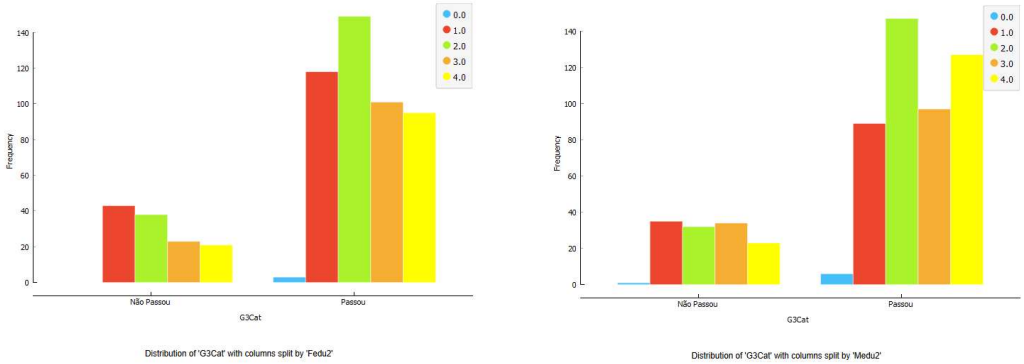



*Figura 6. Desempenho aluno sexo feminino nas educações dos pais*

Para alunos do sexo feminino que não passaram podemos verificar um decréscimo continuo à medida que a educação do pai cresce. No entanto, quando estes alunos passam podemos verificar que a distribuição do nível de educação do pai segue uma distribuição normal com média em 2. Estes dados são relevantes e não carecem de amostras, comprometendo um volume total de 466 alunos do sexo feminino que passaram. Este pode ser um indicativo de que um Pai com maior nível de educação pode não ser benéfico na perspetiva do desempenho de uma estudante. Esta razão pode ser possivelmente justificada com os hábitos do quotidiano e a dinâmica familiar geral envolvida, saindo esse estudo fora do âmbito do corrente.

O nível da educação da mãe não apresenta aparente influência sobre o desempenho do estudante do sexo feminino que não passou. Para a aluna que passou, a distribuição sofre grandes variâncias à medida que a educação da mãe cresce, não sendo possível aferir uma conclusão.

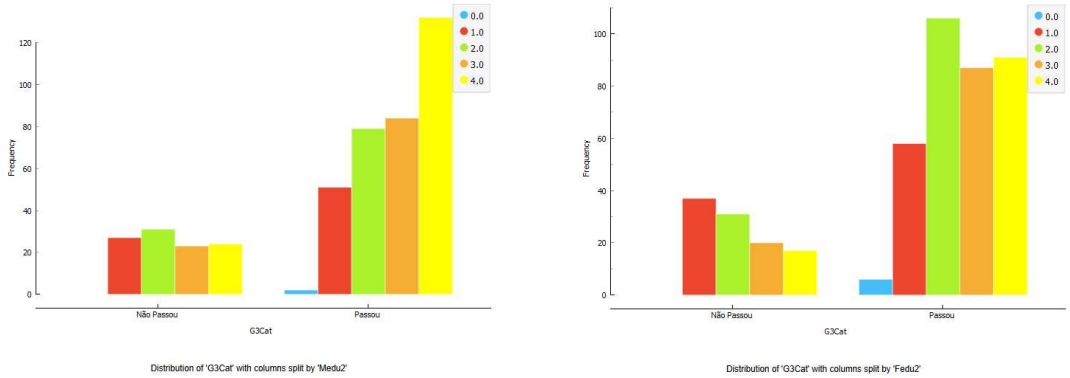



*Figura 7.Desempenho aluno sexo masculino nas educações dos pais*

Atendendo à Figura 7 podemos verificar que nos alunos do sexo masculino que não passaram, a educação da mãe não tem influência notória, sendo as frequências similares em número em todos os níveis. Por outro lado, quando o aluno do sexo masculino passa, existe uma forte correlação entre a educação da Mãe e a quantidade de alunos que passam, sendo estes igualmente crescentes.

Ainda na mesma figura verificamos que à medida que a educação do pai cresce, a quantidade de alunos do sexo masculino que não passaram decresce, isto é, há uma forte correlação negativa. Os dados relativos aos alunos do sexo masculino que passam, apresentam grande variação de frequência à medida que a educação do pai cresce, não sendo possível aferir qualquer conclusão.

## Subtema 3

A terceira questão a analisar prende-se com o desempenho do estudante mediante o seu consumo de álcool semanal e o seu nível de saúde geral.

Na figura 8 podemos notar um notório decréscimo entre a quantidade de consumo de álcool em concorrência com a frequência de alunos que passaram e não passaram. No entanto também podemos verificar uma frequência substancialmente maior nos alunos que consomem menos álcool e passaram relativamente aos que não passaram. Estes dados estão em concordância com os prejuízos inerentes ao consumo de álcool na cognição.

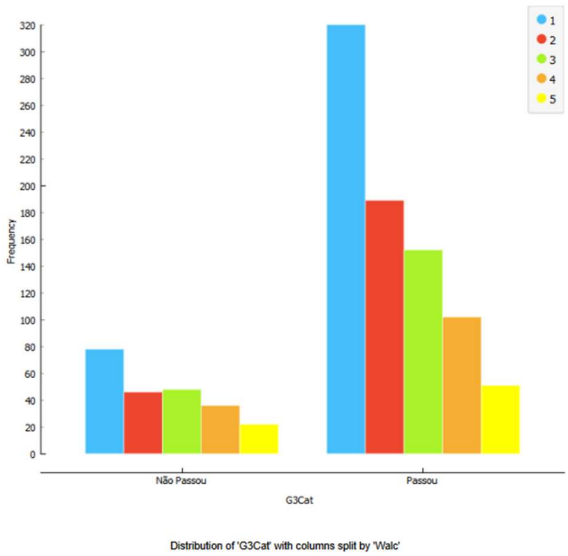
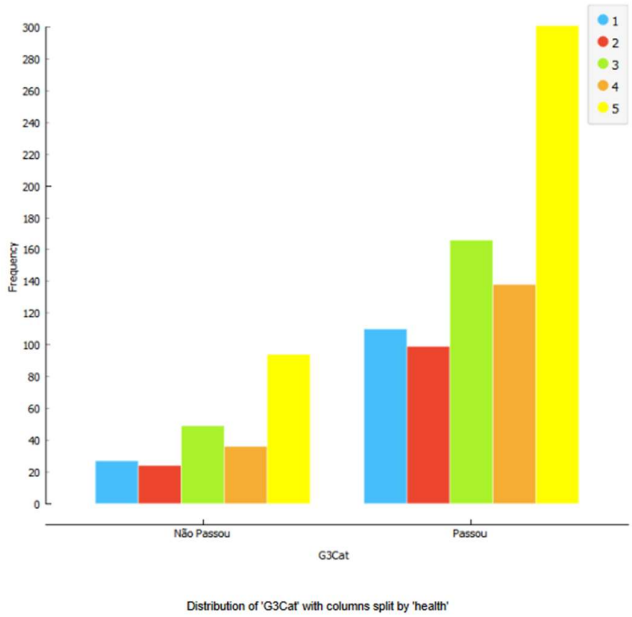



*Figura 8. Consumo alcool semanal e desempenho*

Na figura 9 podemos verificar que existe uma maior frequência no histograma relativo aos alunos que passaram do que os que não passaram. Mais ainda, aqueles que estão de melhor saude tendem substancialmente a passar mais. Curiosamente, podemos verificar o mesmo em relação aos que não passam, mas este facto pode ser justificado pela reduzida amostra de alunos do espaço amostral dos que não passaram.



## Pergunta 2

A segunda grande análise que pretendemos fazer prende-se com o estudo do desempenho dos diferentes géneros nas diferentes disciplinas. Para tal, categorizamos, conforme a pergunta 1, o atributo G3 para os parâmetros 'Passou' e 'Não Passou'.

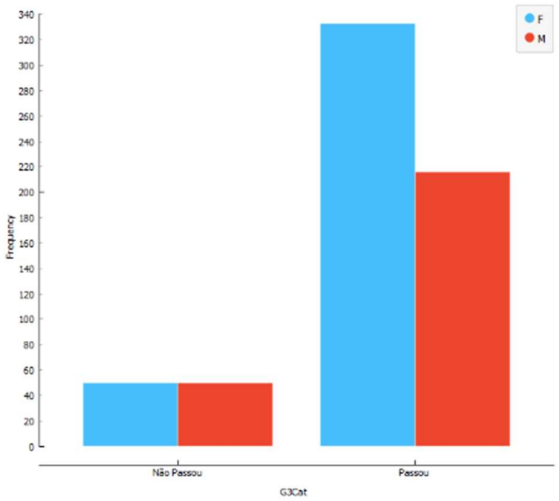



*Figura 9. Disciplina português e desempenho dos géneros*

Olhado para a distribuição da figura 9 podemos verificar de imediato que existe diferença substancial entre os alunos que passaram e os alunos que não passaram à disciplina de português.

Nos alunos que não passaram a distribuição foi relativamente semelhante para ambos os géneros, no entanto, para os alunos que passaram podemos verificar que estes se são em maior número femininos. Esta observação pode-se se justificar com a maior afluência de estudantes do sexo feminino (521, aposto a 453 do sexo masculino).

Em pé de igualdade, podemos então alegar que o percentil relativo masculino tende a ser maior nos alunos que não passaram a português.

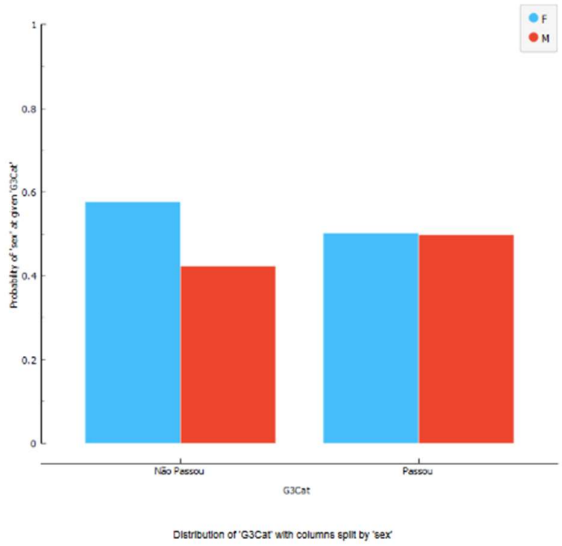



*Figura 10. Disciplina matemática e desempenho dos géneros*

Atendendo à distribuição do gráfico 10 podemos verificar uma maior frequência de alunos do sexo feminino que não passaram à disciplina de matemática. Esta observação, consoante a análise da figura 9, pode ser justificada pela maior afluência de estudantes do sexo feminino.

No entanto, se olharmos para os alunos que passaram à disciplina de matemática, podemos ver que estes se encontram em pé de igualdade. Tendo em conta o percentil relativo dos géneros, podemos afirmar que os estudantes do sexo masculino tendem a ter um melhor desempenho nesta disciplina.

Em suma, tudo quanto os dados indicam e as informações que daí podemos inferir, o sexo masculino tende a ter um melhor desempenho na disciplina de matemática e um menor desempenho na disciplina de português, em comparação com o aluno do sexo feminino. Esta observação pode ser futuramente estudada de forma vinculada à neurofisiologia de ambos os sexos, sendo esse estudo fora do âmbito deste projeto.

## Pergunta 3

Recorrendo à mesma categorização da variável G3, conforme as perguntas 1 e 2, fizemos um estudo das distribuições em relação ao desempenho do estudante consoante o tempo de deslocamento para os espaços académicos e a sua área de residência (urbana ou rural).

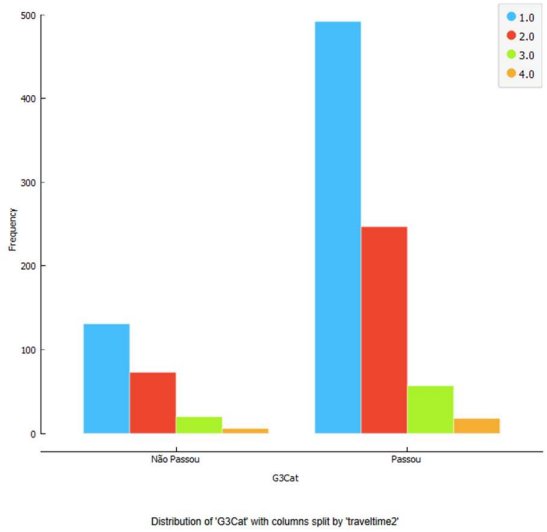



*Figura 11. Desempenho e tempo de deslocamento*

Nos histogramas da figura 11 podemos verificar que apesar da distância, as frequências relativas dos estudantes que passaram e não passaram seguem uma distribuição muito similar, sendo a diferença entre totais estudantes que passaram e estudantes que não passaram irrelevante face aos resultados.

Por outras palavras, podemos aferir que a distância de deslocamento do estudante não apresenta ser algo que influencie de forma determinativa o desempenho do aluno.

Podemos ainda inferir deste histograma que a grande maioria dos alunos vive até 1 km de distância da sua escola.

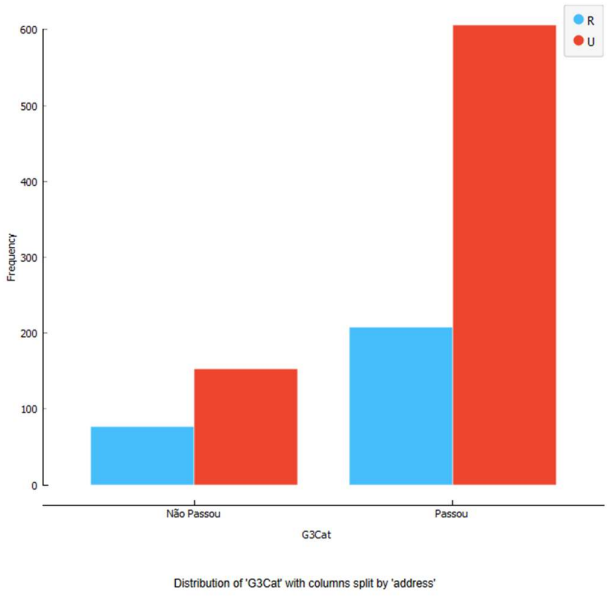



*Figura 12. Desempenho e área de residência*

De forma análoga à análise da figura 11, podemos aferir que a área de residência não tem influência no desempenho do aluno, sendo as distribuições para ambas as partes 'Passou' e 'Não Passou' similares.

Podemos aferir deste histograma que a maior parte dos estudantes vive numa área urbana, sendo este facto concordante com o de a grande maioria dos alunos viver até 1 km de distância, da análise anterior.

## Previsão

| Model | MSE | RMSE | MAE | R2 |
|---|---|---|---|---|
| kNN | 1.244 | 1.115 | 0.872 | 0.850 |
| Tree (6) | 1.314 | 1.147 | 0.859 | 0.841 |
| Random Forest (5) | 0.882 | 0.939 | 0.718 | 0.893 |
| Neural Network | 1.855 | 1.362 | 1.075 | 0.776 |

*Figura 13. Modelos de previsão*

Para este projeto tentamos criar vários modelos de previsão limitando o conjunto de variáveis a ser utilizado para aquelas que constavam nas questões que tratamos. No entanto, todos esses modelos apresentavam um baixo factor R2. Posto isto, o melhor modelo de previsões que conseguimos construir agrega todas as variáveis do dataset com exceção das variáveis G1 e G2, sendo dentro destes o melhor resultado apresentado por uma Ramdom Forest.

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Random Forest (2) | 0.535 | 0.747 | 0.702 | 0.680 | 0.747 |
| Neural Network | 0.511 | 0.726 | 0.702 | 0.686 | 0.726 |
| Logistic Regression | 0.567 | 0.777 | 0.696 | 0.685 | 0.777 |

*Figura 14. Previsão à pergunta 1*

Para o problema tratado na pergunta 1 foi possível construir um modelo de previsões com uma precisão máxima de 0.686 com base numa rede neuronal.