Universidade do Minho

Mestrado Integrado em Engenharia Informática

Aprendizagem e Extração de Conhecimento

EXTRAÇÃO DE CONHECIMENTO

Tomada de decisão do aluno

Luís Gomes (A78701)

Joel Rodrigues (A79068)

Elisa Valente (A79093)

5 de Dezembro de 2018

Resumo

O stress é um característica intrínseca ao ser humano que tanto pode ter efeitos positivos como negativos nos seus comportamentos. A resposta ao stress é determinada pela capacidade de perceção de um evento ou problema sendo que, dependendo do nível de stress em que se encontra, poderá ter respostas diferentes. Como tal, analisar o comportamento humano e tentar prever o estado de stress em que uma pessoa se encontra é fundamental para ter noção da capacidade de resposta de um indivíduo. Deste modo, foram recolhidas informações acerca do comportamento de estudantes universitários durante a realização de um exame e do nível de stress que eles apresentaram no decorrer da prova. Através desses dados é possível criar um modelo que, por meio de determinadas características, seja capaz de prever o nível de stress que os estudantes apresentam.

Conteúdo

Figuras

1 Introdução

O presente documento, elaborado no âmbito da unidade curricular de Aprendizagem e Extração de Conhecimento, pretende apresentar as decisões tomadas na realização do trabalho prático. Este trabalho tem como objetivo o desenvolvimento de um modelo para prever o estado de stress, associado à tomada de decisão, de alguns estudantes universitários durante a realização de um exame.

Primeiramente, na secção 2, de forma a explicar o *dataset* em análise, é efetuada uma descrição do conjunto de dados onde são expostos os seus atributos e apresentadas as primeiras impressões que dele foram recolhidas. Com o intuito de perceber o comportamento dos dados e da gama de valores que estes apresentam, na secção 3 é feita a sua visualização com o apoio de gráficos. Após isto, parte-se para a pré-preparação dos dados onde são realizadas todas as transformações essenciais e necessárias para o aprimoramento do modelo, como apresentado na secção 4. Depois dos dados estarem devidamente processados são treinados e validados em modelos diferentes. Tendo em conta os *scores* obtidos em cada modelo, é feita a seleção do modelo a utilizar. Estes tópicos encontram-se detalhados na secção 5.

Como finalização do trabalho, na secção 6, é realizada a procura da melhor combinação de parâmetros para o modelo encontrado.

2 Descrição e análise do conjunto de dados

O conjunto de dados utilizado refere-se às tomadas de decisão dos alunos na realização de exames universitários. Este apresenta o nível de stress percecionado por cada estudante durante o exame, um conjunto de características relativas às decisões tomadas durante a sua realização, bem como atributos do exame:

- **ExamID** (Var. Categórica)

  Identificador do tipo de exame;

- **FinalGrade** (Var. Quantitativa Contínua)

  Nota final do aluno no exame;

- **StudyID** (Var. Categórica)

  Identificador único para cada instância de recolha de dados;

- **TotalQuestions** (Var. Quantitativa Discreta)

  Número de questões do exame;

- **ATBQ - Average Time Between Questions** (Var. Quantitativa Contínua)

  Tempo médio que o aluno gastou entre cada questão. (Medido em milissegundos);

- **ATBD - Average Time Between Decision** (Var. Quantitativa Contínua)

  Tempo médio que cada aluno apresenta para tomar uma decisão (inserção, alteração ou remoção de uma resposta). (Medido em milissegundos);

- **MTBD - Median Time Between Decision** (Var. Quantitativa Contínua)

  Mediana do tempo que cada aluno demora a tomar uma decisão. (Medida em milissegundos);

- **DTE - Decision Time Efficiency** (Var. Quantitativa Contínua)

Tempo médio que um aluno demora a tomar uma boa decisão. (Medido em milissegundos);

- **GDTE - Good Decision Time Efficiency** (Var. Quantitativa Contínua)

  Tempo médio que um aluno demora a tomar uma boa decisão. (Medido em milissegundos);

- **MaxDuration** (Var. Quantitativa Contínua)

  Tempo máximo que um aluno demorou para tomar uma decisão. (Medido em milissegundos);

- **MinDuration** (Var. Quantitativa Contínua)

  Tempo mínimo que um aluno demorou para tomar uma decisão. (Medido em milissegundos);

- **NumDecisions** Var. Quantitativa Discreta)

  Número de decisões tomadas pelo aluno durante o exame, sendo elas inserção, remoção ou alteração de uma resposta;

- **QuestionsEnter** (Var. Quantitativa Discreta)

  Número de vezes que o aluno abriu qualquer questão do exame;

- **DMR - Decision Making Ratio** (Var. Quantitativa Contínua)

  Rácio entre o número de respostas inseridas/alteradas/removidas e o número total de ações. (Medida percentual);

- **CDMR - Correct Decision Making Ratio** (Var. Quantitativa Contínua)

  Rácio entre o número de decisões consideradas corretas e o número de respostas inseridas/alteradas/removidas. (Medida percentual);

- **TotalDuration** (Var. Quantitativa Contínua)

  Tempo total de duração do exame. (Medido em milissegundos);

- **VTBD - Variance Time Between Decision** (Var. Quantitativa Contínua)

  Variância do tempo entre as decisões. (Medido em milissegundos);

- **PSS_Stress** (Var. Quantitativa Discreta)

  Estado de stress do estudante.

Numa primeira análise ao conjunto de dados, visualizando apenas o seu conteúdo, sem qualquer tipo de ferramenta de análise, começa-se a ter a noção da ordem de grandeza dos valores que cada variável pode tomar. Assim, podemos concluir de imediato que a variável *StudyID* não será interessante para o modelo uma vez que representa apenas um identificador, não tendo nenhum conhecimento relevante a si associado. Da mesma forma, a variável *ExamID* representa um identificador do tipo de exame realizado, no entanto esta variável poderá ter influência no nível de stress percecionado, uma vez que os exames terão características diferentes, como por exemplo o seu nível de dificuldade. Assim, representou-se esta variável na forma de categorias com inteiros de 0 a 8.

3 Visualização dos dados

Para obter uma melhor perceção do comportamento dos dados e da gama de valores que estes apresentam é necessária a construção de modelos gráficos que facilitem a sua observação. Assim, para cada uma das variáveis em análise, é realizado um diagrama de extremos e quartis (*box plot*), um histograma e um diagrama de dispersão (*scatter plot*). O primeiro é útil para identificar a variação associada ao valor de cada variável com a separação em quartis e fornecendo informações de interesse como por exemplo a mediana. O segundo apresenta a frequência das ocorrências, permitindo assim visualizar, a quantidade de dados que apresentam o mesmo valor e de que forma estes se distribuem. Este tipo de gráfico permite obter indicadores como a média e a moda. Por último, no terceiro é evidenciada a relação de cada variável com a variável de interesse *PSS_Stress*. Para conseguir alcançar este objetivo é necessário substituir os campos com falta de informação por valores que não enviesem os dados e, sendo assim, todas as entradas cujos valores são desconhecidos foram substituídos pela média dos restantes dados conhecidos dessa variável. Os gráficos obtidos podem ser observados nas figuras 1 a 17.

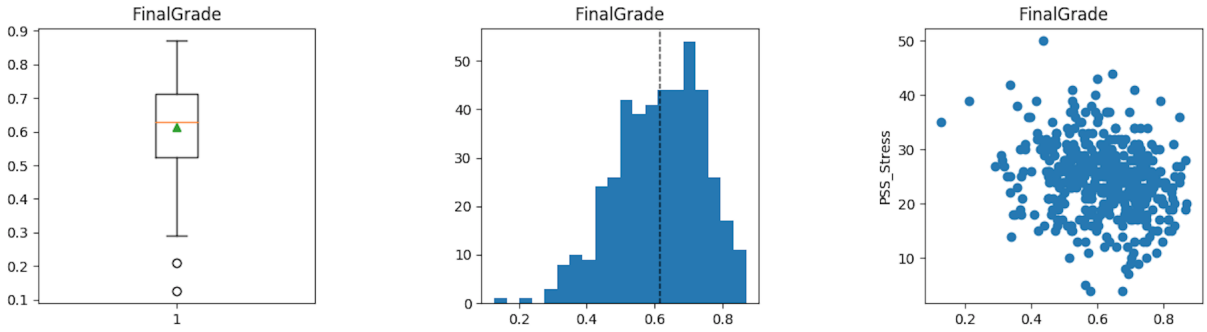

Figura 1: Modelos gráficos: **FinalGrade**.

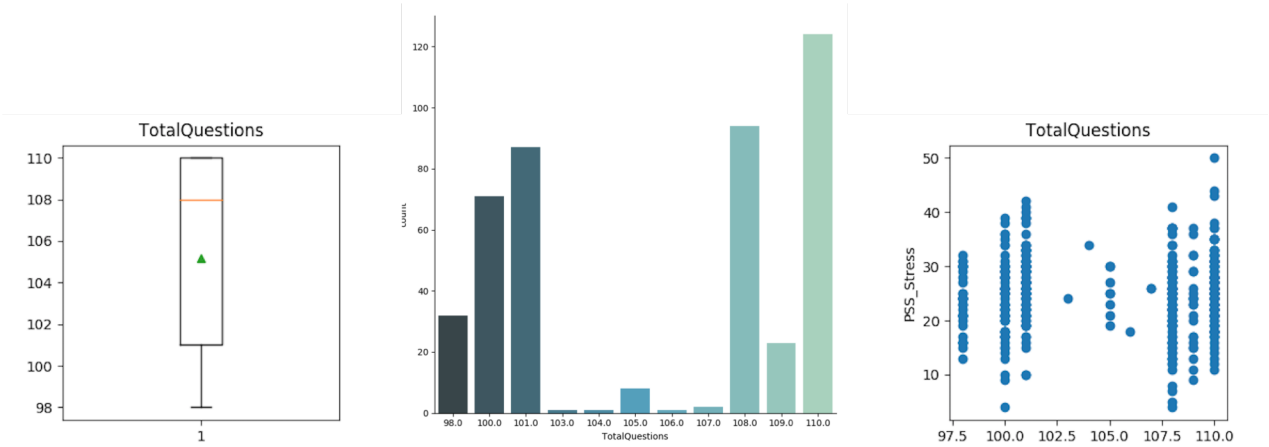

Figura 2: Modelos gráficos: **TotalQuestions**.

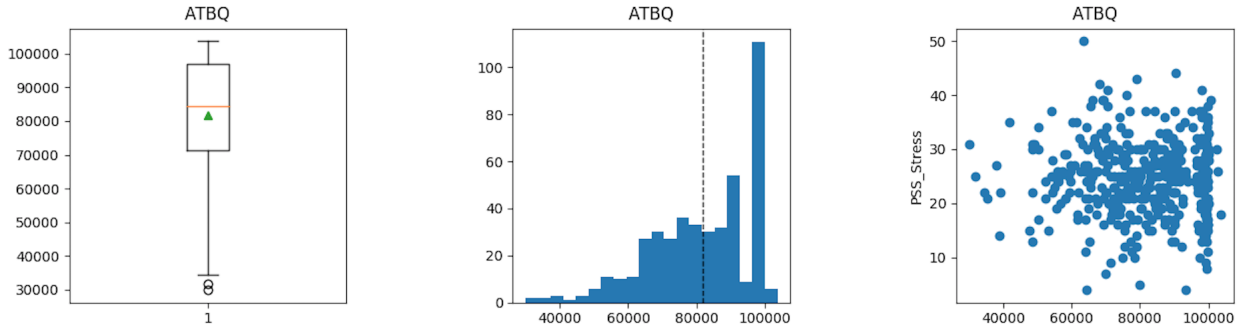

Figura 3: Modelos gráficos: **ATBQ**.

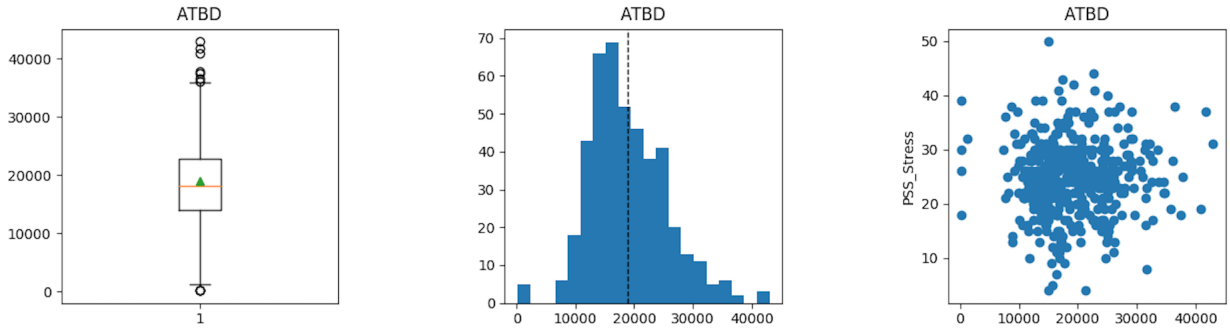

Figura 4: Modelos gráficos: **ATBD**.

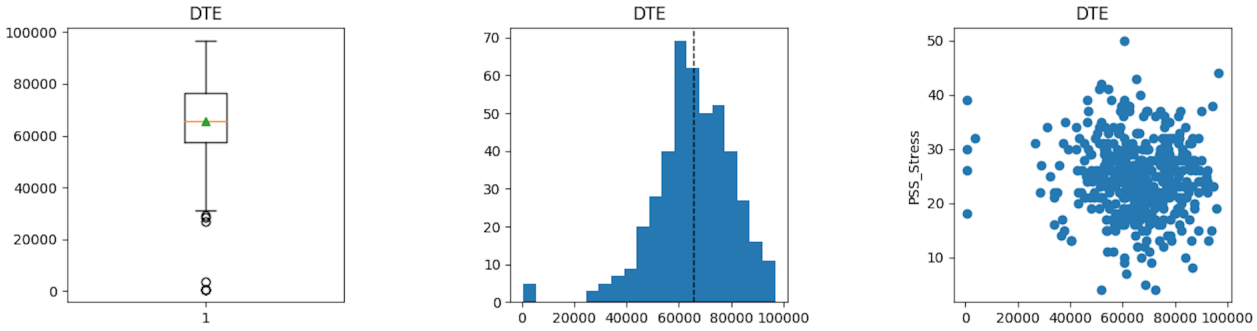

Figura 5: Modelos gráficos: **DTE**.

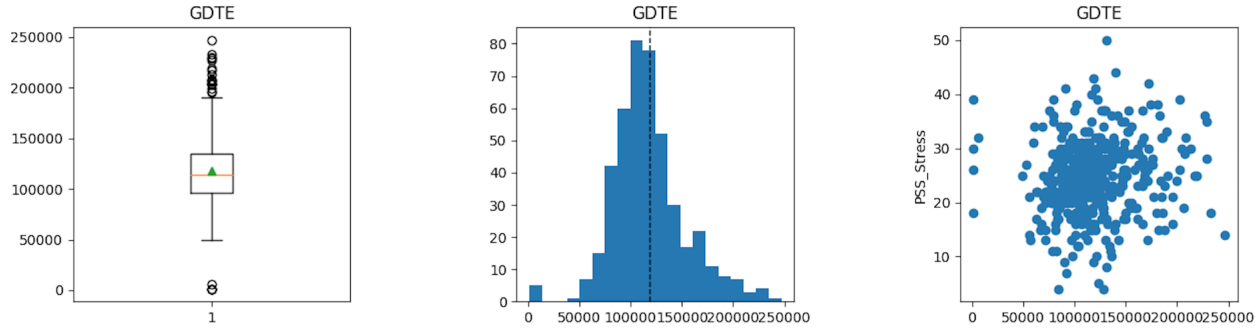

Figura 6: Modelos gráficos: **GDTE**.

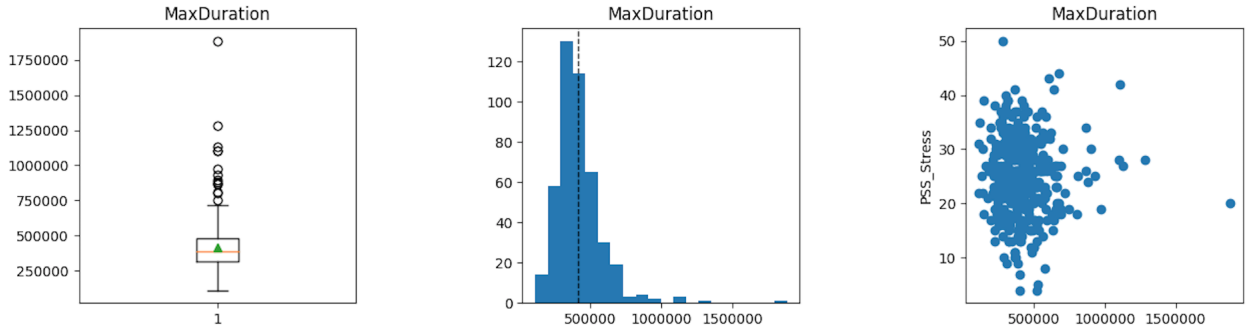

Figura 7: Modelos gráficos: **MaxDuration**.

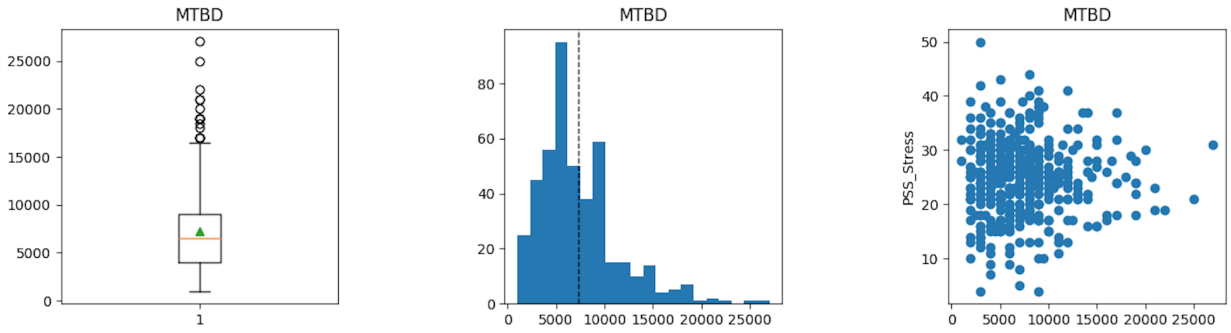

Figura 8: Modelos gráficos: **MTBD**.

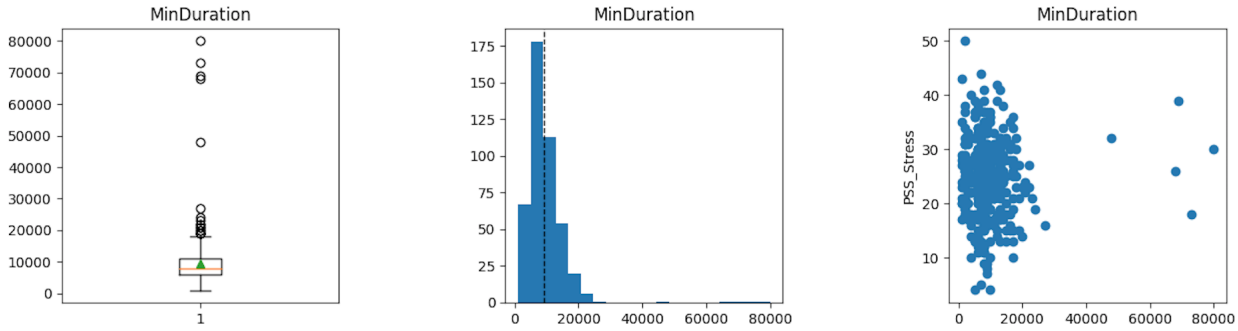

Figura 9: Modelos gráficos: **MinDuration**.

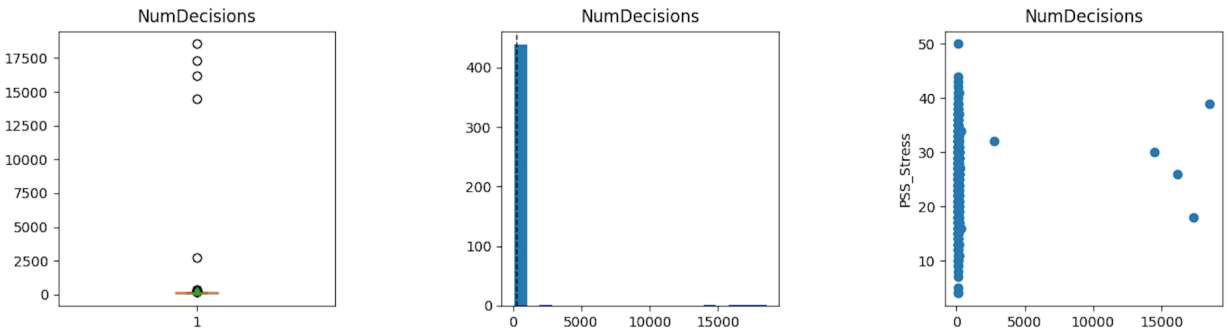

Figura 10: Modelos gráficos: **NumDecisions**.

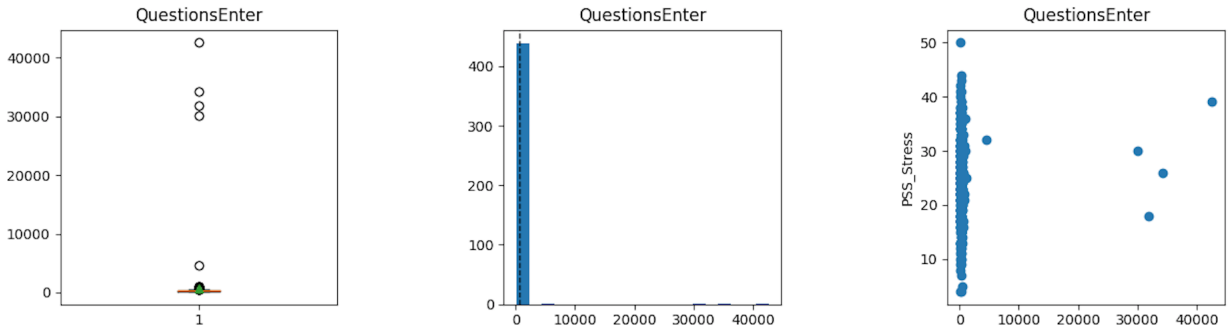

Figura 11: Modelos gráficos: **QuestionsEnter**.

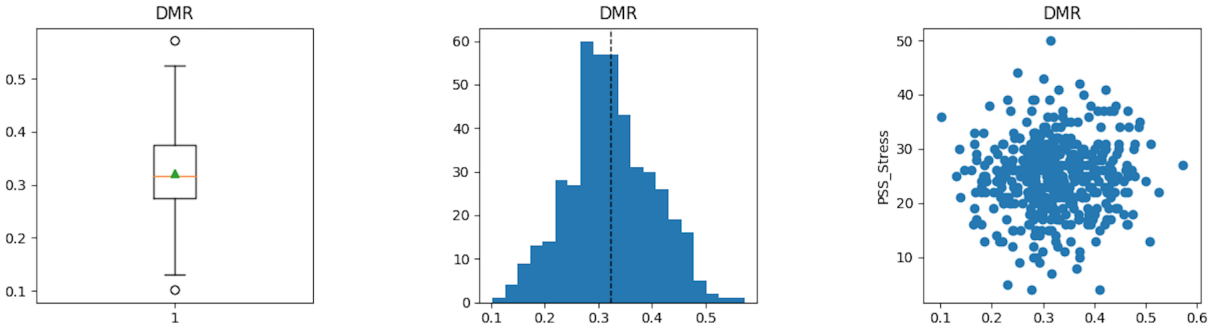

Figura 12: Modelos gráficos: **DMR**.

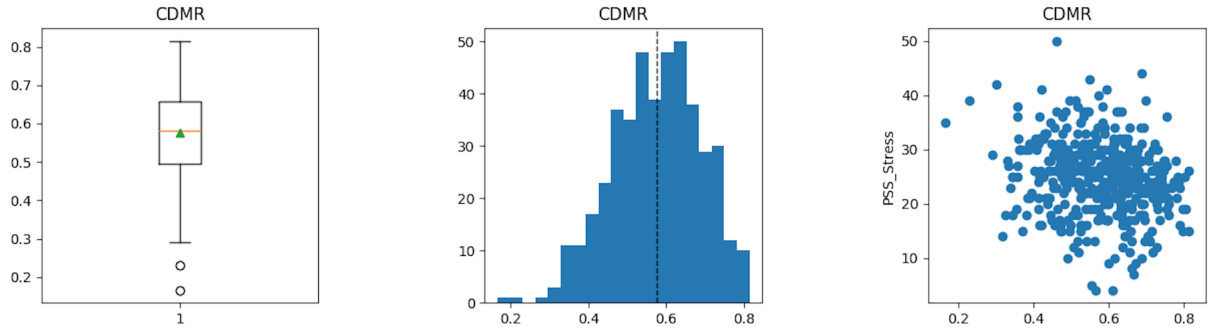

Figura 13: Modelos gráficos: **CDMR**.

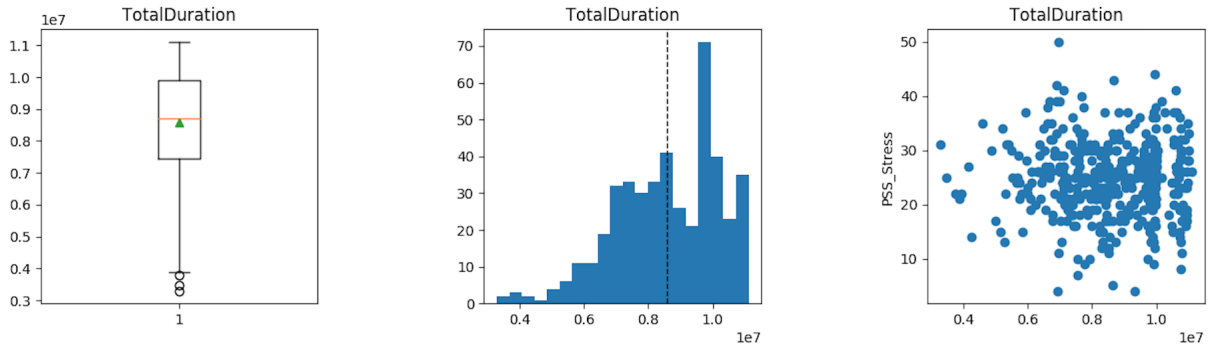

Figura 14: Modelos gráficos: **TotalDuration**.

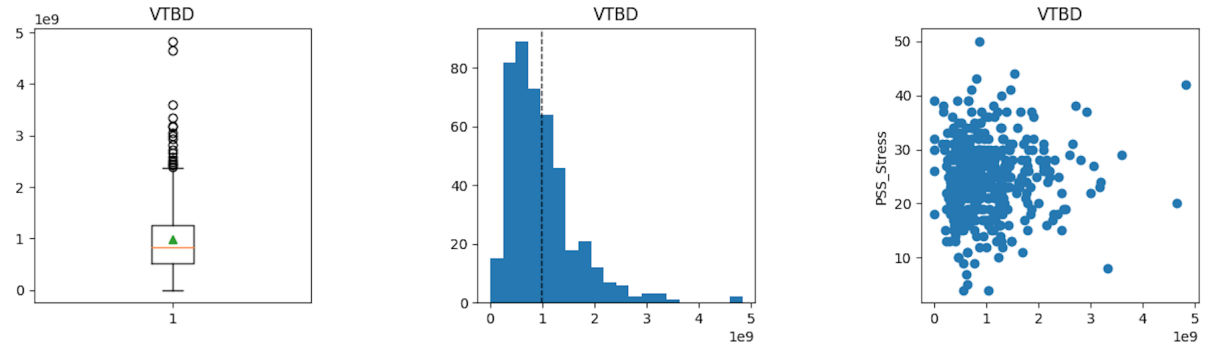

Figura 15: Modelos gráficos: **VTBD**.

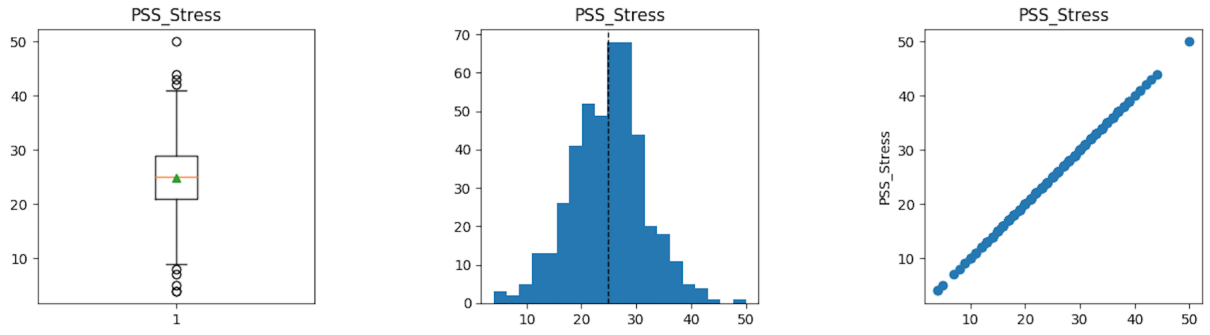

Figura 16: Modelos gráficos: **PSS_Stress**.

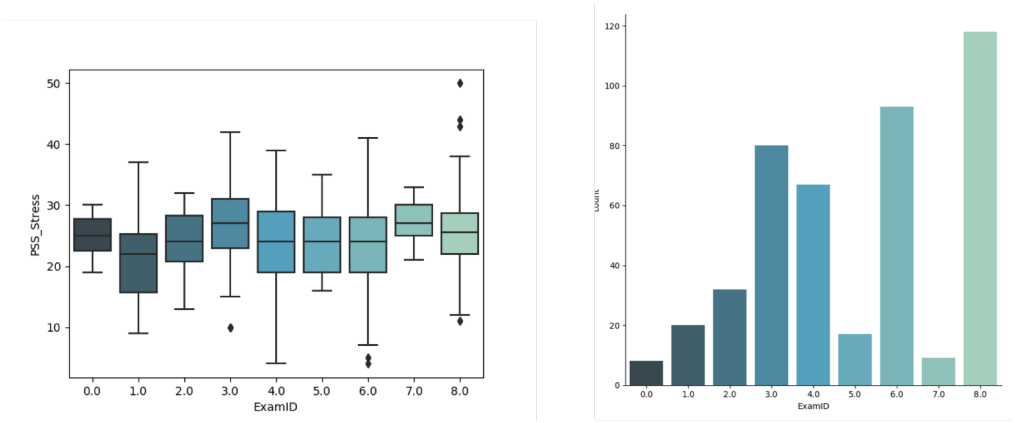

Figura 17: Modelos gráficos: **Exame_Id**.

Com a análise dos gráficos é de fácil perceção a existência de vários *outliers* agressivos, estando em alguns casos situados numa ordem de grandeza diferente das restantes observações. Mais curioso ainda é que existe um padrão na ocorrência destes *outliers*, uma vez que existem 5 valores em muitas das variáveis que estão fora da gama padrão, mais acentuadamente nos gráficos:

- 4 - valores próximos de 0 (abaixo de 5000);
- 5 - valores próximos de 0 (abaixo de 20 000);
- 6 - valores próximos de 0 (abaixo de 20 000);
- 9 - valores acima de 40 000;
- 10 - valores acima de 2500;
- 11 - valores acima de 2000.

Através desta análise surge a questão *Será que estes outliers correspondem à mesma fonte de recolha de dados?*, ou seja, *Será que, numa linha do dataset, para uma dada variável de valor incomum, as outras apresentam um valor desviado também?*. A resposta a esta questão, respondida afirmativamente, revela que de facto os dados com estas características vêm da mesma fonte, o que revela que por algum motivo, estes dados estarão errados. As linhas correspondentes a estes dados são as linhas 8 a 12 (inclusive).

Olhando para o *dataset* original verifica-se também que nestas 5 linhas, 4 delas apresentam incompletude dos dados na variável **MTBD** o que pode ser mais um indicador para a ocorrência de erros na obtenção dos dados destas instâncias.

Após retirar estas instâncias a visualização dos gráficos torna-se mais útil, uma vez que já é possível visualizar os gráficos numa escala apropriada aos dados. A título de exemplo podemos verificar na figura 18 a diferença de escala quando comparada com a figura 9 que continha *outliers* agressivos e o mesmo se verifica na comparação da figura 19 com a figura 10.

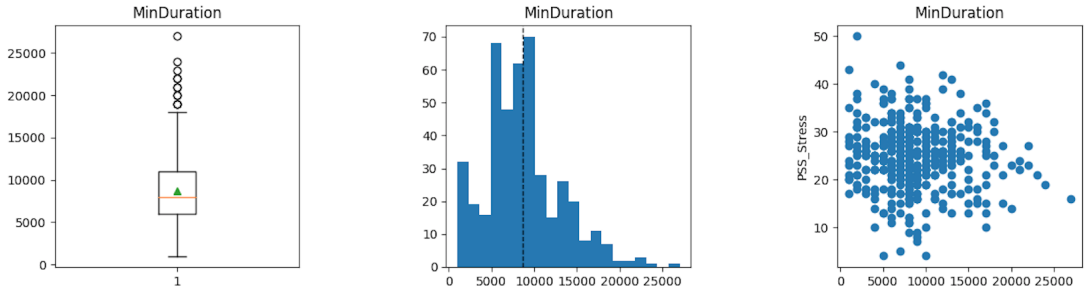

Figura 18: Modelos gráficos sem *outliers* agressivos: **MinDuration**.

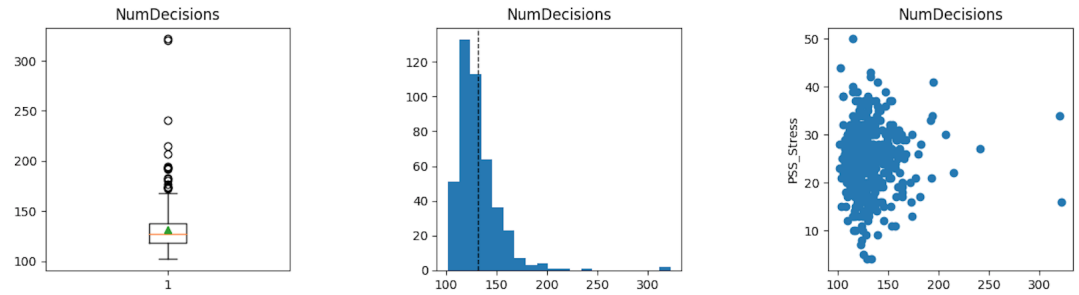

Figura 19: Modelos gráficos sem *outliers* agressivos: **NumDecisions**.

4 Pré-preparação dos dados

Após visualizados os dados na sua forma bruta, há a consciência de que o conjunto de dados deve sofrer alterações para se poder criar um modelo ótimo. Como tal, primeiramente é feita uma seleção dos preditores a utilizar, de seguida é realizada a filtragem dos dados em falta e finaliza-se o processo com a transformação dos dados.

4.1 Seleção de Dados

A seleção dos dados a utilizar para a construção do modelo foi feita tendo em conta as observações da fase de visualização. Como tal, selecionaram-se os preditores que apresentam maior relação com a variável de resposta.

4.1.1 Preditores

Como já referido, a primeira variável a ser excluída do grupo de preditores é o *StudyID* uma vez que não tem informação relevante associada. Como o objetivo principal é definir um modelo que seja capaz de prever o nível de stress de um estudante de acordo com as suas decisões, devem ser excluídas as variáveis que estejam relacionadas com outras características que não a tomada de decisão. Nestas enquadram-se assim o *ExamID* e *TotalQuestions*, uma vez que são atributos do exame realizado, independentes das decisões tomadas.

Assim, as variáveis a utilizar como preditores são:

- FinalGrade
- ATBQ
- ATBD
- DTE

- GDTE
- MaxDuration
- MTBD
- MinDuration
- NumDecisions
- QuestionsEnter
- DMR
- CDMR
- TotalDuration
- VTBD

4.1.2 Variável de resposta

Como o que se pretende é prever o estado de stress dos estudantes, a variável de resposta é a *PSS_Stress*.

4.2 Transformações

Para a criação de um modelo ótimo é necessário admitir a homogeneidade das variâncias dos dados. Como verificado através da análise da gama de valores que estes apresentam, tal não acontece. Neste caso, as transformações serão usadas para resolver o problema dos dados em falta, alterar a escala de medida dos valores, tornando a variância dos dados homogénea, e tratar os *outliers*.

4.2.1 Filtragem de dados em falta

Para a filtrar os dados em falta, poderiam ser utilizadas várias estratégias. Como visto na secção 3, a primeira filtragem utilizada foi a substituição dos valores inexistentes pela média das restantes entradas como forma de visualizar os dados.

Como podemos ver na figura 20, sabe-se que existe apenas uma coluna (MTBD) onde existem dados em falta, sendo a sua percentagem muito reduzida (menos de 1% de valores em falta nesta coluna).

ExamID	0.000000
FinalGrade	0.000000
PSS_Stress	0.000000
StudyID	0.000000
TotalQuestions	0.000000
ATBQ	0.000000
ATBD	0.000000
DTE	0.000000
GDTE	0.000000
MaxDuration	0.000000
MTBD	0.009009
MinDuration	0.000000
NumDecisions	0.000000
QuestionsEnter	0.000000
DMR	0.000000
CDMR	0.000000
TotalDuration	0.000000
VTBD	0.000000

Figura 20: Percentagem de NaN nas variáveis do dataset.

Aliando esta informação ao facto de as variáveis correspondentes apresentarem valores muito desviados do que seria informação verdadeira, pode-se inferir que a melhor decisão será a eliminação destas linhas. Para além destas linhas, como visto na fase de visualização dos dados, tem sentido também a remoção de uma quinta entrada, que apesar de não ter informação em falta apresenta da mesma forma dados que se desviam da realidade das restantes ocorrências.

4.2.2 Normalização

A normalização dos dados é feita para acelerar a otimização. Se os dados estiverem em diferentes escalas, o tempo exigido para que a função de otimização encontre os pontos ideais é muito superior. No entanto, se o *dataset* contiver *outliers* no seu conjunto de dados, a normalização dos dados aumentará os dados "normais"para um intervalo muito pequeno. Neste estudo os dados, depois de retirados os *outliers* agressivos, foram normalizados para o intervalo [0,1].

4.2.3 Padronização

Padronizar os recursos de forma a que eles tenham como média o valor 0 e como desvio padrão o valor 1 é importante quando se comparam medidas que têm unidades diferentes. Para além disso, é um requisito geral para muitos algoritmos de *Machine Learning*. Comparativamente à normalização, a padronização apresenta melhor desempenho quando o *dataset* contém *outliers* uma vez que esta não limita o conjunto de valores dos dados. Como alternativa à normalização, e tendo este *dataset* alguns *outliers*, esta será uma abordagem que será tida em conta também.

4.2.4 Codificação de etiquetas e binarização

A codificação do etiquetas foi utilizada na fase de visualização dos dados para mapear cada tipo de exame (*ExamID*) numa classe específica, atribuindo-lhe um inteiro correspondente, no entanto esta variável foi excluída do grupo de preditores a utilizar.

4.2.5 Discretização

A discretização de atributos transforma dados numéricos em nominais e tem como objetivo encontrar representações de dados concisas, como categorias que sejam adequadas para a tarefa de aprendizagem, mantendo o máximo possível de informações do atributo

original. Neste caso, a discretização foi apenas aplicada à variável de resposta uma vez que se pretende prever o estado de stress em que um aluno se encontra e não o valor exato de stress que o aluno apresenta.

Muito baixo nível de stress	Baixo nível de stress	Médio nível de stress	Alto nível de stress	Muito alto nível de stress
]0,10]	]10,20]	]20,30]	]30,40]	]40,52]

Figura 21: Distribuição das categorias pelos intervalos de valores definidos.

Para esta variável decidiu-se fazer uma discretização de igual largura, figura 22, em vez de uma discretização de igual altura, figura 23, pois definiu-se a amplitude das categorias em função do conhecimento que se pretende adquirir com a variável de resposta, como se verifica na figura 21.

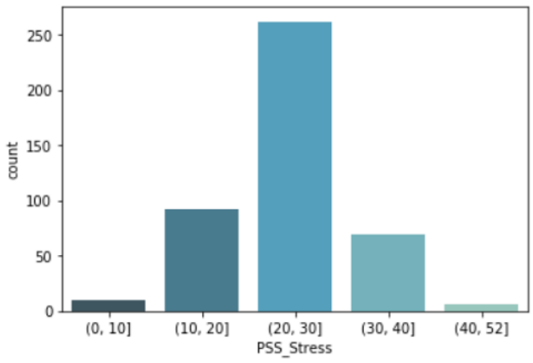

Figura 22: Discretização em largura.

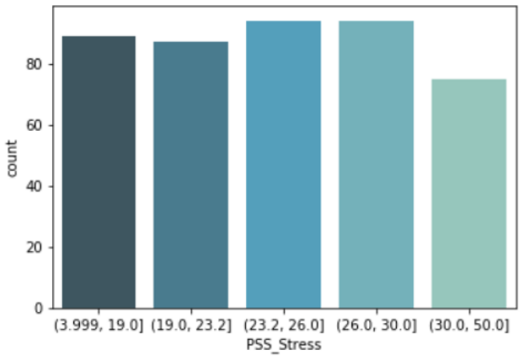

Figura 23: Discretização em altura.

5 Treino, validação e seleção do modelo

No seguimento do processo detalhado na secção 4.2, foram guardadas várias instâncias dos dados com as diferentes transformações efetuadas:

- **Preditores:** normalizados; padronizados.
- **Variável de resposta:** discretizada; padronizada e discretizada; normalizada e discretizada.

De forma a encontrar o modelo que melhor se ajusta aos dados fornecidos, foram treinados e validados diversos modelos de classificação pois a variável de resposta encontra-se discretizada. Deste modo foram usados os seguintes modelos: *Support Vector Machine* (SVM), *Gaussian Naive Bayes* (GaussianNB), *Logistic Regression* (LR), *Artificial neural networks* (ANR) e *K-nearest neighbors* (KNN).

Para a avaliação dos modelos foram utilizados dois métodos: a validação cruzada (*KFold*) e a divisão dos dados de treino e teste (80-20). De forma a obter os melhores resultados possíveis, foram testados vários valores para o número de preditores a utilizar (1 a 14) e o número de divisões na validação cruzada. Não existindo melhorias significativas nos resultados obtidos, foram utilizados todos os preditores e um k=10. Paralelamente a estas variações, foram conjugadas todas as combinações dos restantes fatores (preditores, variável de resposta e modelo).

De todos os testes efetuados, o modelo onde se obteve melhor *score* foi o **Logistic Regression**, com os preditores normalizados e a variável de resposta discretizada. Os *scores* obtidos com *Kfold* foram de 0.599 e com a divisão dos dados de 80% para treino e 20% para teste foram de 0,693.

6 Hiperparameterização

Como forma de melhorar os resultados obtidos, foi realizada uma hiperparemeterização do modelo selecionado. Deste modo, utilizou-se o algoritmo *Grid Search* que faz uma procura exaustiva de subconjuntos de parâmetros do modelo que apresentem melhores resultados.

Após a implementação deste algoritmo concluiu-se que os parâmetros que o modelo estava a usar já eram os ótimos uma vez que não houve um melhoramento significativo dos resultados (*score* de 0,69 para 0.7).

```
gridSearch.best_params_ {'C': 1e-06, 'penalty': 'l2'}
lr.C=0.001
lr.penalty='l2'
x_train,x_test,y_train,y_test = train_test_split(features_normalized,target_discretized,test_size=0.2)
lr.fit(X=x_train,y=y_train)
LogisticRegression(C=0.001, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
lr.score(x_test,y_test) 0.7045454545454546
```

Figura 24: Extrato de código representativo da implementação do algoritmo *Grid Search*.

7 Conclusão

Tendo em conta o tempo necessário para a execução de cada uma das etapas associadas a um projeto desta natureza, é de destacar uma maior importância na fase de análise, visualização e pré-preparação dos dados uma vez que é nestas fases que são tomadas as decisões mais importantes. De facto, os dados podem ser tratados e alterados nestas fases o que irá ter uma grande influência no resultado final, independentemente do modelo escolhido.

Relativamente ao modelo com melhor ajuste aos dados encontrado, pode concluir-se que apresenta resultados bastante positivos, visto que possibilita a previsão correta de cerca de 70% dos casos de teste. Admite-se, no entanto, a possibilidade de este resultado estar um pouco influenciado pelo facto de a maioria dos casos de teste e treino se encontrarem desbalanceados, sendo que mais de 50% dos dados se encontram no intervalo [20,30] (Fig. 16).

De forma a melhorar o modelo para o poder utilizar de uma maneira mais precisa, seria importante um novo levantamento de dados ou a utilização de técnicas que permitissem o seu balanceamento.