Departamento de Informática

P.PORTO

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO

Machine Learning

Student Performance

Trabalho realizador por:

João Silva (8220024),

Luís Silva (8220025)

Estudantes do mestrado em Engenharia Informática

Supervisores/Orientadores

Responsável pelo projeto: João Ramos

ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Resumo

A educação em Portugal está em constante evolução, contudo ainda existe uma taxa de insucesso significativa, nomeadamente nas disciplinas de Português e Matemática.

Existe, portanto a necessidade de utilização de diferentes técnicas e estudos para promoção do aumento de resultados escolares. Com o proposto trabalho, pretende-se a demonstração de técnicas de *Machine Learning* (DM) que analisam informações escolares, juntamente com aspectos pessoais dos alunos e fazer previsões a partir daí para que mais ajuda eficiente pode ser dada, de acordo com o desempenho previsto de cada aluno. As duas disciplinas referidas (Matemática e Português) foram primeiro modeladas com cenários de classificação e regressão binários e de cinco níveis. Dentro de cada tipo de cenário, foram testados vários modelos de *Machine Learning*.

Adicionalmente, foi aplicada uma técnica de aprendizado não supervisionada, *Clustering*, para dividir os dados em duas turmas diferentes de alunos (aprovado ou reprovado na respectiva disciplina).

Com este trabalho, métodos de predição mais eficientes podem ser construídos como forma de melhorar a qualidade do ensino, centrada principalmente nos alunos com maior risco de insucesso.

Índice

Lista de Figuras

1 Introdução

Este capítulo tem como objetivo apresentar o enquadramento e a motivação do tema do trabalho, os objetivos e resultados esperados no seu desenvolvimento e a abordagem de investigação selecionada.

1.1 Contextualização

No âmbito da Unidade Curricular de *Machine Learning* (ML), do mestrado em Engenharia Informática, foi-nos proposto o estudo de um *Dataset* à escolha do Grupo. Neste sentido, os *datasets* escolhidos pelo grupo, devido ao interesse do grupo pela evolução no Sistema Educacional, foram de *Student Performance*, do qual o artigo *"Using data mining to predict secondary school student performance"* retrata[1].

1.2 Motivação

A nível de motivação, este trabalho apresenta como tais, a aplicação dos diversos conhecimentos adquiridos ao longo do semestre. Sejam estas motivações a aplicação de diversas técnicas de Análise de Dados, de promoção de limpeza e alteração dos dados, através de diferentes técnicas de *Feature Engineering*, bem como a aplicação de diversos algoritmos de *Machine Learning*, sejam estes de aprendizagem supervisionada ou de aprendizagem não-supervisionada.

Por último, apresenta-se ainda como objetivo, o desenvolvimentos de uma pequena interface, para a visualização dos algoritmos em prática.

1.3 Objetivos

O principal objetivo do trabalho, é através da Análise dos dados existentes, conseguir fazer a previsão do desempenho dos estudante com base em 3 esquemáticas diferentes.

- Classificação binátia (Passar/Reprovar).
- Classifcação em 5 níveis (sendo o nível I, Muito bom, e o nível V, Insuficiente).
- Regressão, previsão da nota final (entre 0 e 20).

1.4 Metodologia de Trabalho

Durante o processo de desenvolvimento do presente trabalho, optou-se por uma estrutura de trabalho denominada por *CRISP-DM Methodology (Cross Industry Standard Process for Data Mining)*.

Este tipo de metodologia, é usualmente utilizada em contexto profissional, sendo vista como a abordagem mais comum aos diversos problemas, que podem ser resolvidos através de *Data Mining*.

Neste sentido, o *CRISP-DM* assenta em 6 fases, das quais se destacam:

- Business Understanding;
- Data Understanding;
- Data Preparation;
- Modeling;
- Evaluation;
- Deployment.

As diferentes fases do *CRISP-DM*, serão descritas na secção de Desenvolvimentos, com os diferentes processos executados em cada uma das fases.

2 Desenvolvimento

Considerando que todos os dados nos conjuntos de dados são factuais, há algumas análises que precisam de ser feitas antes de escolher ou aplicar um algoritmo de aprendizado de máquina. Estes conjuntos de dados foram construídos extraindo dados de duas escolas portuguesas (região do Alentejo) durante o ano lectivo de 2005-2006.

De uma forma bastante resumida, o seguinte Fluxograma elabora o processo idealizado para o projeto.

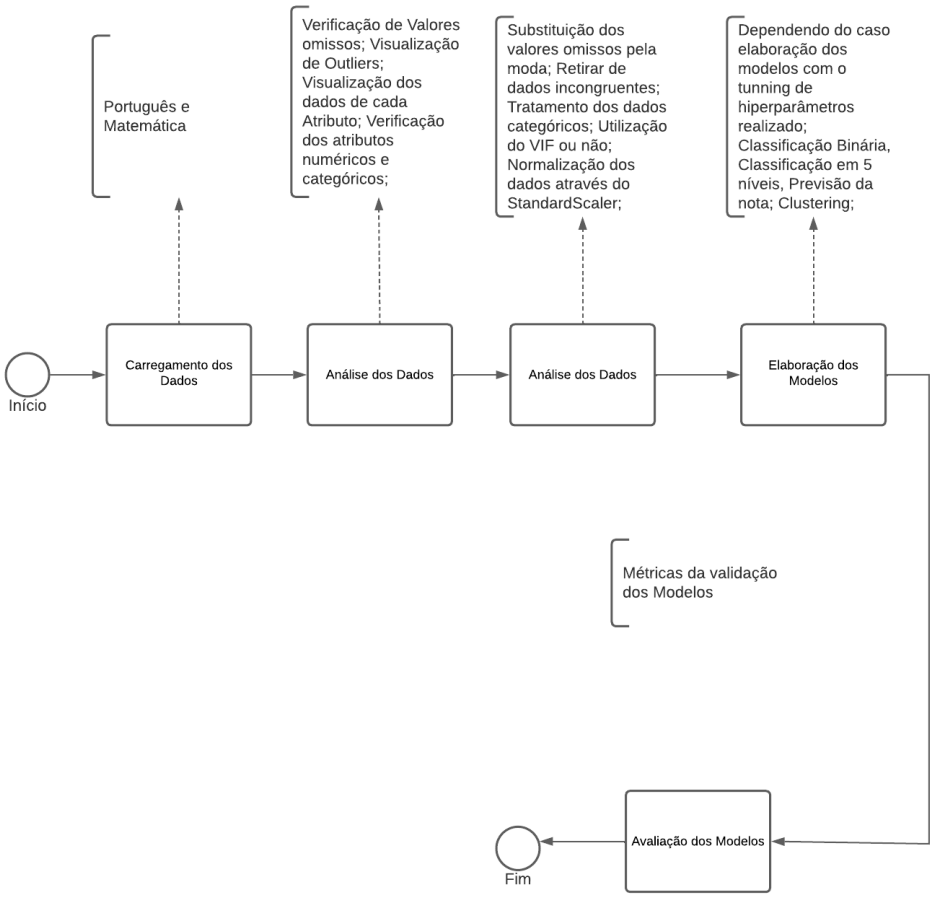

Figura 1: Fluxograma de resumo de todo o Processo.

2.1 *Business Understanding*

No atual contexto, o Ministério da Educação português não tem como saber exatamente quais são os principais fatores que afetam o desempenho dos alunos, o que não permite saber ao certo, quais as medidas a tomar, de forma a promover a melhoria da

situação académica dos estudantes.

O foco principal deste trabalho é determinar, se um estudante tem probabilidade de passar/reprovar às disciplinas, ou mesmo saber, a que nível de alunos pertence o mesmo.

Tendo em conta esta divisão, a identifcação dos possíveis atributos de cada grupo de alunos torna-se essencial, de forma a verificar a raiz da diferença entre os alunos e promover a melhoria futura dos alunos com maiores dificuldades.

2.2 *Data Understanding*

No contexto do problema identificado no seguinte projeto, existem dois *Datasets* que necessitam de ser comprendidos.

Um dos dois conjuntos de dados se refere-se às notas obtidas em Matemática com 395 exemplos, enquanto a outra se refere ás notas obtidas na disciplina de Português contendo 649 registros. Existem 32 atributos para cada instância e um atributo-alvo (G3) conforme Tabela 1 (obtida por meio de pesquisas) e Tabela 2 (obtidos pelos boletins escolares).

Tabela 1: Attributes - Surveys

Attribute	**Type**	**Description**
school	Binary	School 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
sex	Binary	Gender 'F'-female or 'M'-male
age	Numeric	Age from 15 to 22
address	Binary	Home address type 'U' - urban or 'R' - rural
famsize	Binary	Family size 'LE3'-less or equal to 3 or 'GT3'-greater than 3
Pstatus	Binary	Parent's cohabitation status 'T' - living together or 'A' - apart
Medu	Numeric	Mother's education 0-none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education or 4-higher education
Fedu	Numeric	Father's education 0-none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education or 4-higher education
Mjob	Nominal	Mother's job 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Fjob	Nominal	Father's job 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
reason	Nominal	Reason to choose this school close to 'home', school 'reputation', 'course' preference or 'other'
guardian	Nominal	Student's guardian 'mother', 'father' or 'other'
traveltime	Numeric	Home to school travel time 1-<15 min., 2-15 to 30 min., 3-30 min. to 1 hour, or 4->1 hour
studytime	Numeric	Weekly study time 1-<2 hours, 2-2 to 5 hours, 3-5 to 10 hours, or 4 - >10 hours
failures	Numeric	Number of past class failures n if 1<=n<3, else 4
schoolsup	Binary	Extra educational support yes or no
famsup	Binary	Family educational support yes or no
paid	Binary	Extra paid classes within the course subject (Math or Portuguese) yes or no
activities	Binary	Extra-curricular activities yes or no
nursery	Binary	Attended nursery school yes or no
higher	Binary	Wants to take higher education yes or no
internet	Binary	Internet access at home yes or no
romantic	Binary	with a romantic relationship yes or no
famrel	Numeric	Quality of family relationships from 1 - very bad to 5 - excellent
freetime	Numeric	Free time after school from 1 - very low to 5 - very high
goout	Numeric	Going out with friends from 1 - very low to 5 - very high
Dalc	Numeric	Workday alcohol consumption from 1 - very low to 5 - very high
Walc	Numeric	Weekend alcohol consumption from 1 - very low to 5 - very high
health	Numeric	Current health status from 1 - very bad to 5 - very good

Tabela 2: Attributes - School reports

Attribute	Type	Description
absences	Numeric	Number of school absences from 0 to 93
G1	Numeric	First period grade from 0 to 20 of Math or Portuguese
G2	Numeric	Second period grade from 0 to 20 of Math or Portuguese
G3	Numeric	Final grade from 0 to 20 of Math or Portuguese

Ainda nesta etapa, extrairam-se diversos gráficos dos atributos dos *Datasets*, tanto de Matemática como de Português. O objetivo destes gráficos passa pela verificação da possível existência de valores estranhos nos diversos atributos, bem como pela visualização dos diferentes valores em cada atributos, bem como pela sua contagem. Outro fator importante desta visualização, passa pela obtenção da informação acerca do balanceamento dos atributos.

No *Dataset* de Matemática, os *plot bar* dos diferentes atributos, encontram-se nas Figuras 2 e 3.

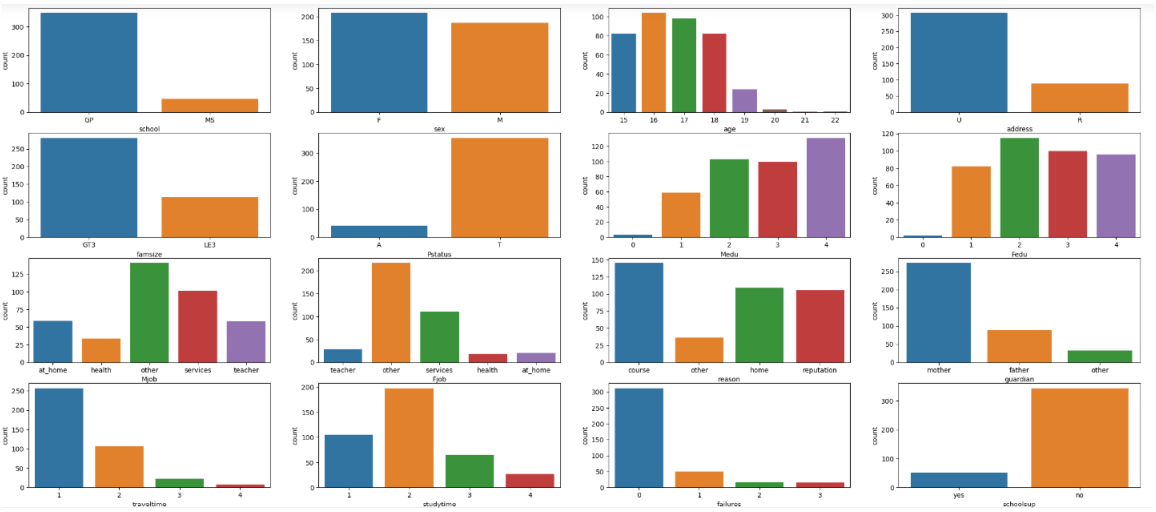

Figura 2: Primeiro Gráfico dos atributos do *Dataset* de Matemática.

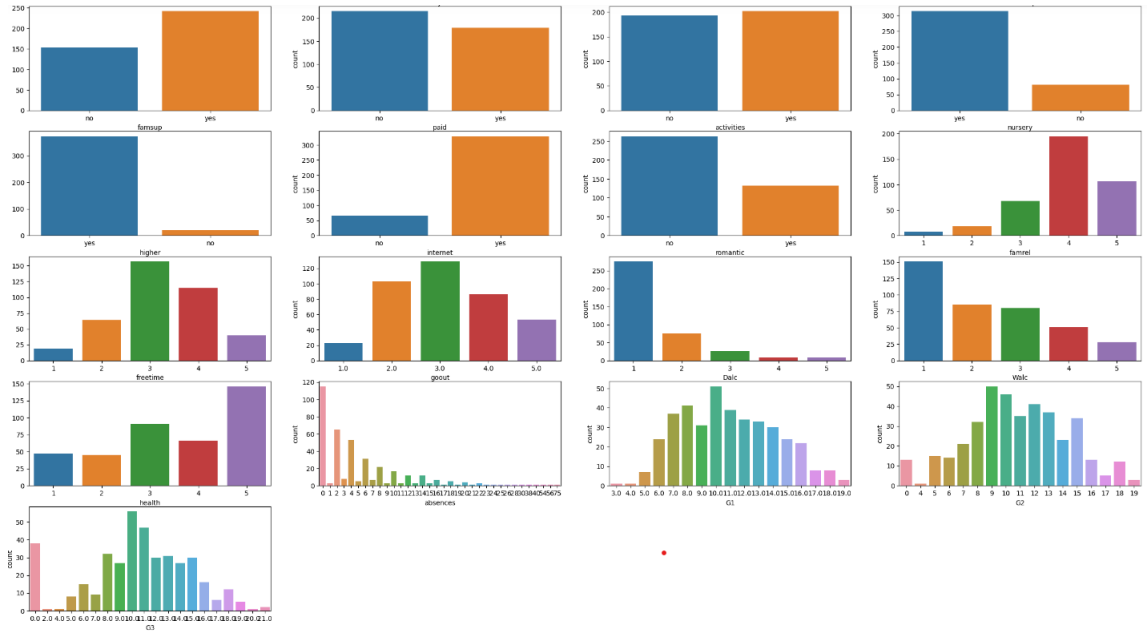

Figura 3: Segundo Gráfico dos atributos do *Dataset* de Matemática.

No *Dataset* de Português, os *plot bar* dos diferentes atributos, encontram-se nas Figuras 4 e 5.

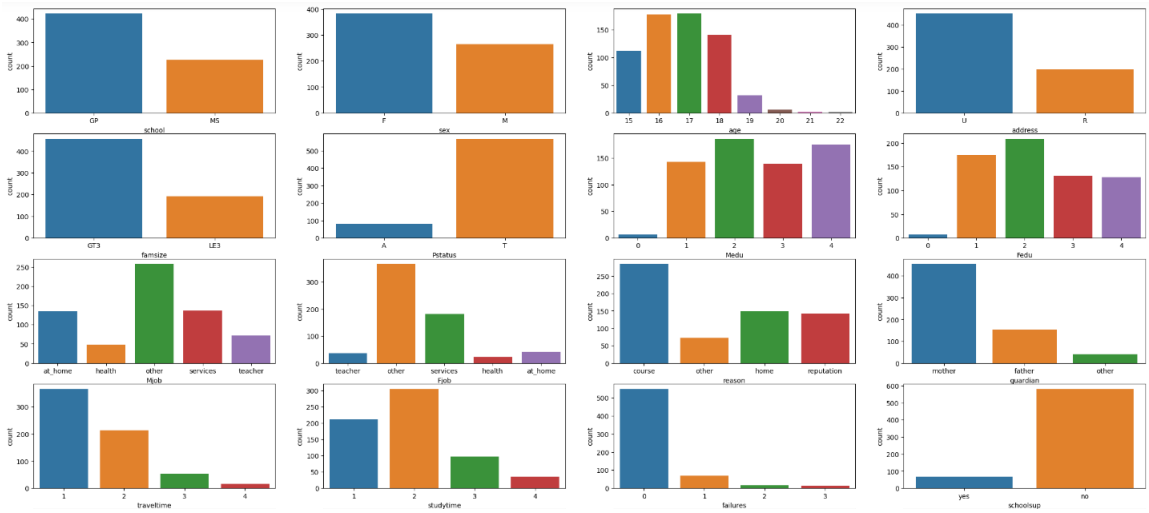

Figura 4: Primeiro Gráfico dos atributos do *Dataset* de Português.

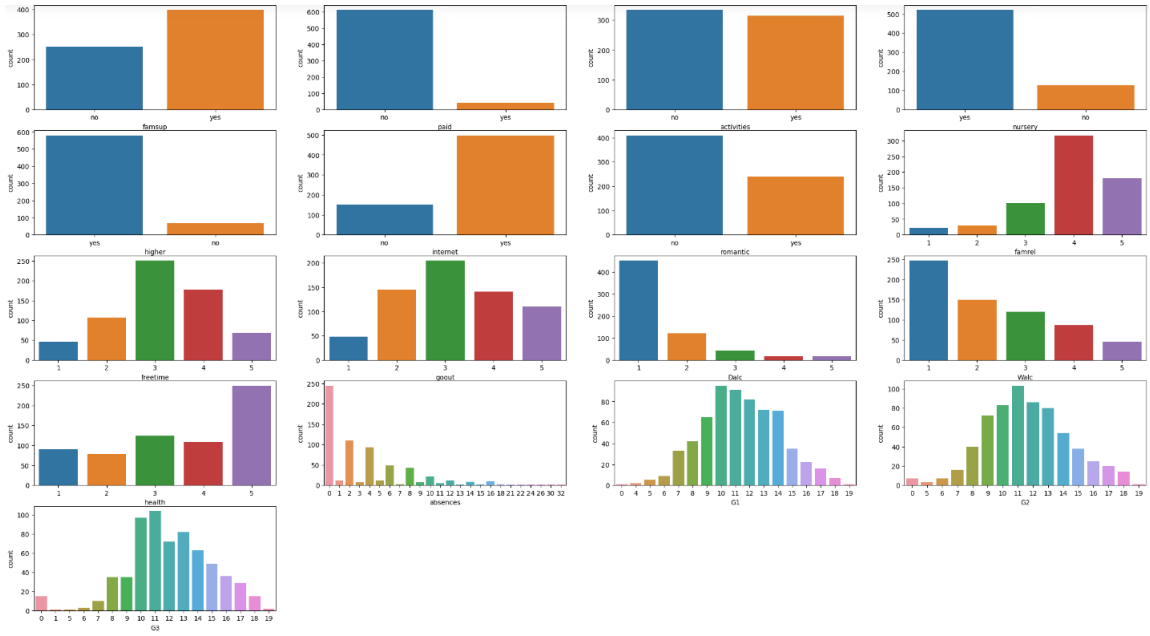

Figura 5: Segundo Gráfico dos atributos do *Dataset* de Português.

O atributo *"sex"* é aproximadamente equilibrado (cerca de 200 instâncias de "F"e cerca de 180 de "M"), enquanto os atributos *"address"*, *"famsize"*, *"Pstatus"*, possuem um número discrepante de instâncias para cada valor. Os atributos *"Medu"* e *"Fedu"* são significativamente equilibrados, mas, por outro lado, *traveltime*, *"studytime* e *"failures"* são mal distribuídos com uma série de entradas para um valor discrepante dos demais. Os seguintes atributos binários *"higher"*, *"school"*, *"schoolsup"*, *"internet"*, *"nursery"*, *"romantic"*, *"famsup"*, *"paid"* e *"activities"* são ordenados do menos equilibrado ao mais equilibrado (*"higher"* tem uma diferença de 355 e *"ativities"* apenas uma diferença de 7). Os atributos restantes são consideravelmente equilibrado, sendo importante destacar o atributo *"absences"* e *"Dalc"* que apresentam poucos valores altos.

Por outro lado, no *Dataset* de Português, todos os atributos seguem uma distribuição semelhante aos do *Dataset* de Matemática. Observa-se que o atributo "sexo" é aproximadamente balanceado (aproximadamente 400 instâncias de "F" e aproximadamente 300 de "M"), o atributo "escola" tem menos discrepância entre os valores "MS" e "GP" e para a *label* "G3", como no outro conjunto de dados, há uma alta porcentagem de valores ao lado da nota "10".

2.3 *Data Preparation*

2.3.1 *Missing Values*

Para o conjunto de dados de Matemática e de Português, foi possível verificar que existem *Missing Values*.

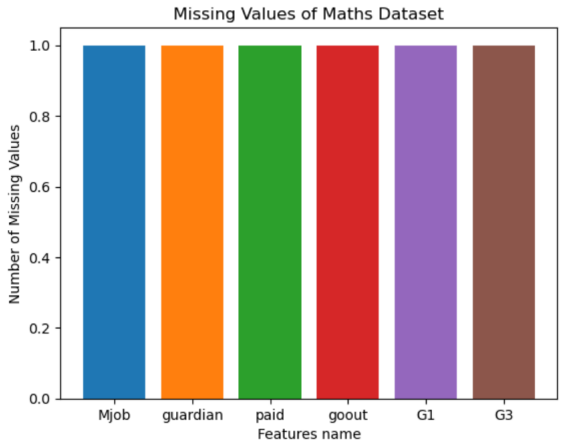

Figura 6: Quantidade de valores nulos por atributo no *Dataset* de Matemática.

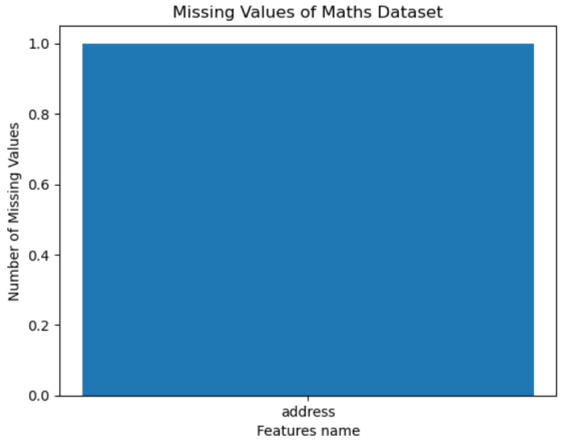

Figura 7: Quantidade de valores nulos por atributo no *Dataset* de Português.

Verificando-se esta existência, torna-se necessário a realização da sua eliminação ou substituição. No caso do presente trabalho, fez-se a sua substituição pela moda do atributo em questão.

De seguida, procedeu-se a uma análise da quantidade de atributos e registos por *Dataset*, dos tipos dos atributos. No *Dataset* de Matemática, verificou-se a seguinte análise.

```
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        394 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    394 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        394 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       394 non-null    float64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          394 non-null    float64
 31  G2          395 non-null    int64
 32  G3          394 non-null    float64
dtypes: float64(3), int64(13), object(17)
```

Figura 8: Análise dos Atributos do Dataset de Matemática.

No *Dataset* de Português, verificou-se que este apresenta o mesmo número de atributos, contudo, este apresenta mais registos por atributo.

```
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64
 3   address     648 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64
 7   Fedu        649 non-null    int64
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64
 13  studytime   649 non-null    int64
 14  failures    649 non-null    int64
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64
 24  freetime    649 non-null    int64
 25  goout       649 non-null    int64
 26  Dalc        649 non-null    int64
 27  Walc        649 non-null    int64
 28  health      649 non-null    int64
 29  absences    649 non-null    int64
 30  G1          649 non-null    int64
 31  G2          649 non-null    int64
 32  G3          649 non-null    int64
dtypes: int64(16), object(17)
```

Figura 9: Análise dos Atributos do Dataset de Português.

2.3.2 Atributos Categóricos

Tendo em conta a Análise dos dois *Datasets*, verificou-se a existência de várias features com atributos Categóricos. Os modelos de *Machine Learning*, não apresentam uma capacidade facilitada de trabalhos com este tipo de atributos, como apresenta com atributos numéricos.

Desta forma, torna-se necessário a alteração destes atributos Categóricos, para atributos numéricos. Existem dois tipos de Algoritmos capazes de promover a alteração destes atributos, o *Label Encoder* e o *OneHotEncoder*.

No presente trabalho, optou-se pela utilização do *OneHotEncoder* para este processo. O *Label Encoder* atribui um número a cada coluna, o que não é relevante,porque certos algoritmos podem dar mais importância a números maiores. Por outro lado, o O algoritmo de codificação One-Hot cria várias colunas com os nomes dos atributos de uma coluna que tem dados categóricos. Essas colunas exibem apenas os valores de 1 e 0, dependendo se a coluna tem o valor categórico ou não, respetivamente.

2.3.3 *Outliers*

De forma a se verificar a presença de *Outliers*, bem como a necessidade ou não necessidade de se efetuar a sua remoção, procedeu-se à visualizção dos *Boxplot* dos atributos numéricos relevantes dos *Dataset* de Português e de Matemática.

Pela análise dos *Boxplot*, facilmente se verifica o que é considerado como sendo um *Outlier*, uma vez que estes aparecem como pontos individuais e separados dos restantes, no *plot*.

Posto isto, nas seguintes figuras podem ser visualizados os *Boxplot* das *Features* numéricas do *Dataset* de Matemática e de Português, respetivamente.

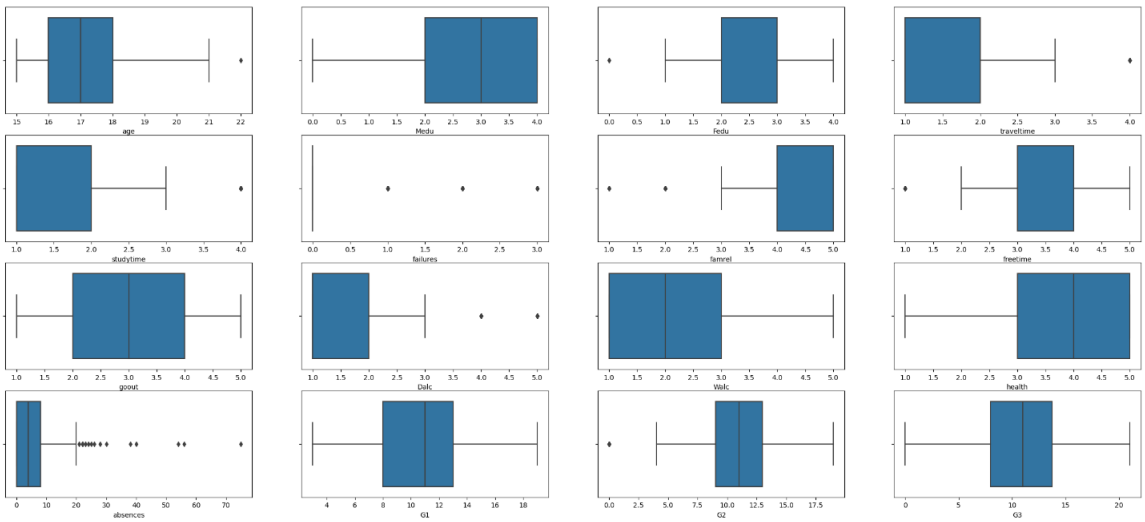

Figura 10: Diagramas de caixa e bigote dos atributos numéricos do *Dataset* de Matemática.

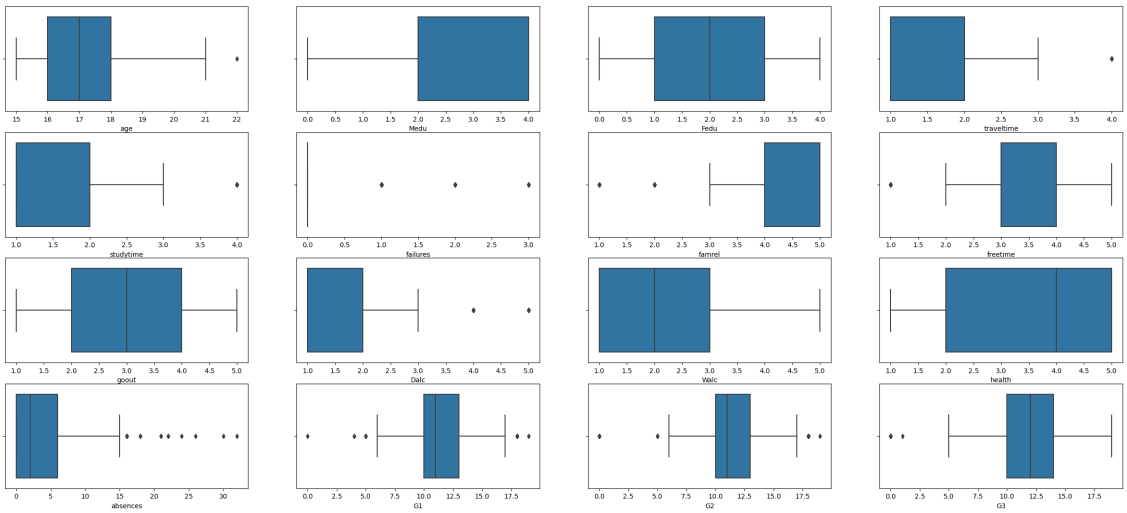

Figura 11: Diagramas de caixa e bigote dos atributos numéricos do *Dataset* de Português.

Tal como se pode verificar pelas Figuras 10 e 11, existem muitos outliers ao longo das diferentes *Features*. A partir desta verificação, é necessário o estudo dos valores considerados como outliers, em cada uma das features.

Apesar dos *Outliers* puderem levar a uma baixa *accuracy* dos Modelos de *Machine*

Learning, nem sempre estes devem ser removidos, pois, apesar de representarem um valor anormal, não quer dizer que não represente algo real e que o modelo deva ter em conta.

Neste sentido, em ambos os *Dataset*, se pode verificar que o atributo *absences*, é o que representa o maior número de *Outliers*. De facto, o estado mais normal é que os alunos faltem poucos às aulas, contudo, optou-se por não se proceder à remoção dos *Outliers*, pois, apesar de representarem uma quantidade elevada de faltas às aulas, estas podem representar a realidade.

2.3.4 Correlação entre atributos

É importante a análise de Correlação entre atributos, de forma a promover a redução do número de atributos como *Input* dos Algoritmos de *Machine Learning*, bem como a promoção de redução de fenómenos de *Overfitting*.

Para análise de correlação entre os diferentes atributos, optou-se pela utilização do fator *VIF*, Variância Inflacionária de Fator.

Este Algoritmo, é um indicador de multicolinearidade usado para a avaliação de correlação entre os diferentes atributos, onde Multicolinearidade ocorre quando uma ou mais variáveis independentes têm uma forte correlação entre si.

O valor de fator *VIF* divide-se em 3 possíveis acontecimentos:

- Se *VIF* é igual a 1, não existe correlação;
- Se *VIF* é maior que 1 e menor que 5, existe correlação intermédia;
- Se *VIF* é maior que 10, existe altra correlação;

2.3.5 Normalização

Os algoritmos de *Machine Learning* são sensíveis à escala apresentada nos dados. Para contornar esse problema, os dados são normalizados para que apresentem uma escala semelhante, entre 0 e 1 ou entre -1 e 1, através de Algoritmos de Normalização ou *Z-Score*, respetivamente. Essa normalização pode ser feita de várias formas, como *Z-Score* ou escala min-max.

No caso do presente trabalho, foi escolhido o algoritmo de *Z-Score* para realizar a normalização dos dados.

2.4 *Modelling*

2.4.1 Classificação e Regressão

Para cada conjunto de dados, três cenários diferentes foram testados, mais concretamente:

- Cenário 1 - A *label* é binária, se G3>=10, significa que passa(1), caso contrário, reprova (0);
- Cenário 2 - *label* é dividida em 5 níveis (desde Insuficiente (5) a muito bom (1));
- Cenário 3 - A verdadeira nota do G3 (deste 1 a 20, onte 1 é a pior 2 20 é a melhor nota).

O Cenário 2 é baseado num sistema de conversões de notas de Erasmus(Table 3) [1].

Nos cenários de Classificação, foram utilizados diversos algoritmos, a Regressão Logística, o *Support Vector Machine*, o *Random Forest*, o *Decision Tree*, *Nearest Neighbour* e o *Naive Bayes*.

No problema de Regressão foi utilizado apenas o algoritmo de *Decision Tree*.

Todos estes algoritmos, na sua documentação, é possível visualizar que têm diversos parâmetros de entrada, que influenciam a performance do algoritmo. Uma vez que é desejado que o Algoritmo tenha o melhor desempenho possível, espera-se que este apresente os melhores parâmetros possíveis.

Esta escolha dos melhores parâmertros possíveis é feita através de um processos demoninado por *Hyperparameter Tuning*, que podem ser feitos de duas formas distintas, através da função*Grid Search* ou da função *Random Search*, presentes na biblioteca *Scikit-learn*. No presente trabalho, para este processo utilizou-se a função *Grid Search*, onde a melhor combinação é selecionada de acordo com a validação na partição de *Cross-Validation*.

No que diz respeito ao Algoritmo de Aprendizagem Não-Supervisionada, escolheu-se o algoritmo de *K-Means*. O *k-Means* é um algoritmo com foco em dividir os dados em *k clusters*, onde a divisão dos dados é feita de uma a providenciar uma elevada similaridade dentro do mesmo *cluster* (intra-*cluster*), e uma baixa similaridade entre *clusters* diferentes (inter-*cluster*).

Este algoritmo baseia-se num processos iterativo de 2 passos consecutivos, onde o 1º

passo é atribuir um centróide a um *Cluster*, e o 2º passo é maximização deste centróide, promovendo o conhecimento que se tem acerca do *cluster*.

Tabela 3: Classificação em 5 níveis baseado num sistema de conversão de notas de erasmus.

Country	**1** (excellent/very good)	**2** (good)	**3** (satisfactory)	**4** (sufficient)	**5** (fail)
Portugal/France	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

2.4.2 Clustering

O objetivo da realização de um *Clustering*, foi para agrupar os estudantes que passam ou reprovam. Esta separação teve por base as colunas, "*absences*"(número de faltas do estudante) , "G1"(valor da nota no 1º periodo), "G2"(valor da nota no 2º periodo), sendo estes os dados mais relevantes para o sucesso do estudante.

2.5 *Evaluation*

2.5.1 Classificação e Regressão

A validação e avaliação de uma Modelo de Aprendizagem Supervisionada de Classificação e de Regressão, apresenta métricas de avaliação distintas.

No presente trabalho, para os problemas do Cenário 1 e Cenário 2, ou seja, de problemas de Classificação, optou-se pela Utilização de *Accuracy*, *Sensitivity* e de *F1 Score*. Estas diferentes métricas são obtidas através da Matriz de Confusão.

Para o problema do Cenário 3, ou seja, um problema de Regressão optou-se pela útiliza do *Root Mean Square Error* e do *Mean Absoluto Error*, *RMSE* e *MAE*, respetivamente.

Estas diferentes métricas de avaliação e validação de algoritmos, são obtidas através das seguintes equações.

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{1}$$

$$Sensitivity = TP/(TP + FN) \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \tag{4}$$

$$MAE = \sqrt{\sum_{i=1}^{N} |y_i - \hat{y}_j|} \tag{5}$$

Tabela 4: Binary classification results using accuracy (bold value is best for each input configuration)

Input	Matemática					Português				
Modelo	DT	SVM	RF	NN	NB	DT	SVM	RF	NN	LR
Binário	**91.59**	91.59	85.71	76.47	77.83	**91.79**	89.23	86.15	86.15	88.2
5 Níveis	**82.35**	73.10	36.97	63.03	62.18	72.82	**72.82**	33.84	53.85	60.00

De forma a visualizar a significância das previsões, foi elaborada a matriz de confusão do algoritmo *Decision Tree* no Cenário de Classificação Binária, utilizando-se o *Dataset* de Matemática.

Tabela 5: Matriz de Consuão da Classifcação Binária, através do *Dataset* de Matemática

		References	
		0	1
Predictions	0	**38**	2
	1	8	**71**

No Cenário de qualificação de 5 níveis, utilizou-se a Matriz de Confusão, também do Algoritmo de *Decision Tree* do *Dataset* de Matemática, uma vez que foram estes que obteram as melhores *Accuracy*.

Tabela 6: Matriz de confusão da Classificação de 5 níveis do *Dataset* de Matemática

		References				
		5	4	3	2	1
Predictions	5	**11**	1	0	0	0
	4	0	**18**	4	0	0
	3	0	0	**13**	2	0
	2	0	0	4	**18**	8
	1	0	1	0	1	**38**

No problema de Regressão, apenas se utilizou o Algoritmo de *Decision Tree*, tanto na tentativa de obtenção de nota final nos *Datasets* de Português e Matemática. De forma a ser facilitada a comparação entre o valor previsto e o valor real, fez-se um *plot* sobreposto da nota real e da nota obtida pelo Modelo de *Decision Tree*. Estes gráficos podem ser observados na Figura 12 e Figura 13, onde a azul se visualiza a nota real, enquanto que a laranja se apresenta a previsão.

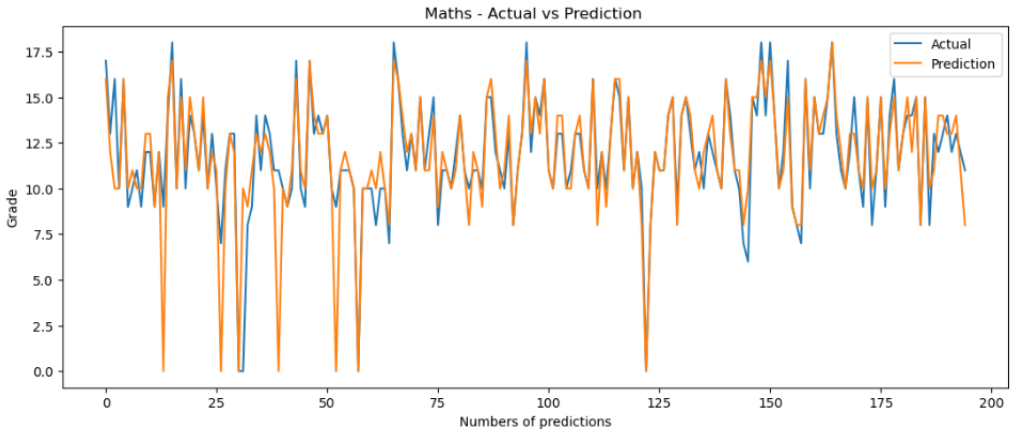

Figura 12: Comparação entre a nota verdadeira e a prevista de Matemática.

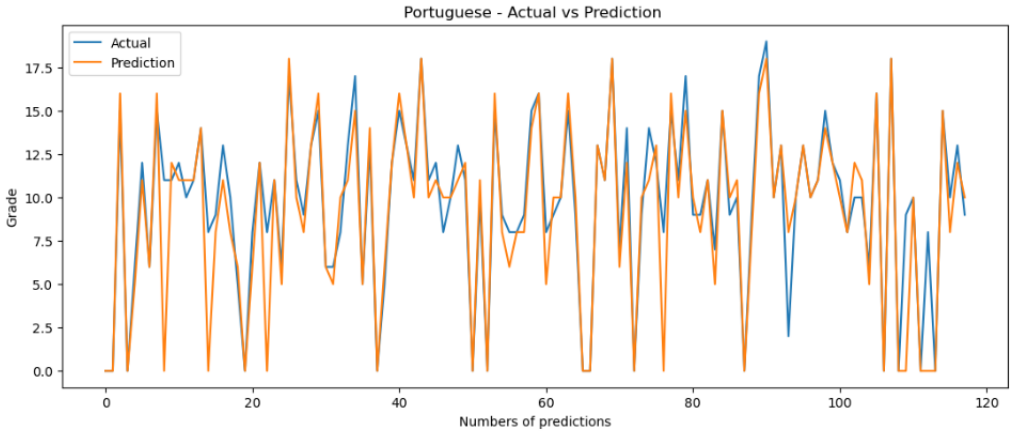

Figura 13: Comparação entre a nota verdadeira e a prevista de Português.

Ainda neste sentido, se verificou que a comparação apresentou melhores no *Dataset* de Português, dado que este apresentou valores *MAE* e *RMSE* inferiores aos do *Dataset* de Matemática. Os valores obtidos, encontram-se na seguinte tabela.

Tabela 7: Grade Prediction

Input	Matemática	Português
Modelo	DT	DT
MAE	1.186	**0.8974**
RMSE	2.285	**1.815**

2.5.2 *Clustering*

De forma a se avaliar o *clustering* realizado, foram criados dois gráficos de comparação das *labels* reais das *labels* obtidas através do *clustering*, no *dataset* de Matemática e Português, respetivamente. Estes gráficos podem ser visualizados nas seguintes figuras.

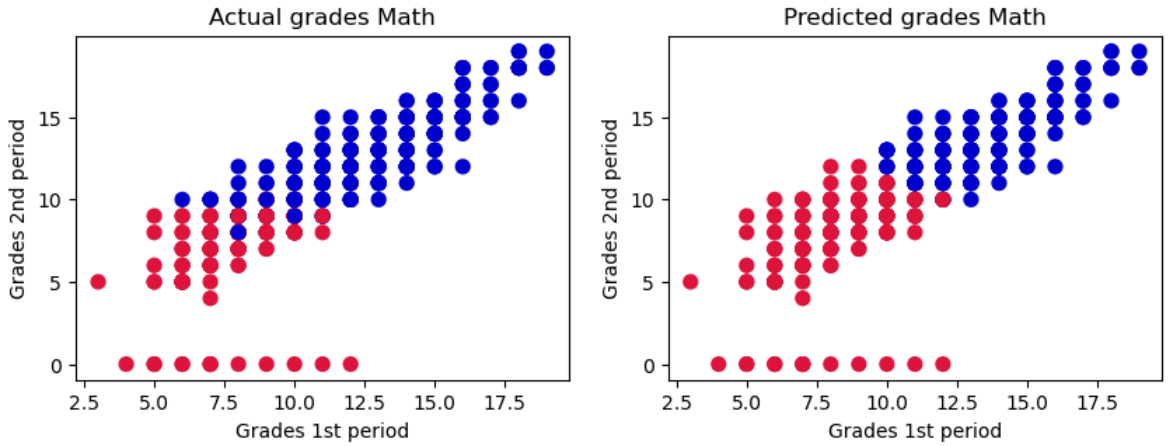

Figura 14: *Clustering* para comparação dos gráficos do *dataset* de Matemática

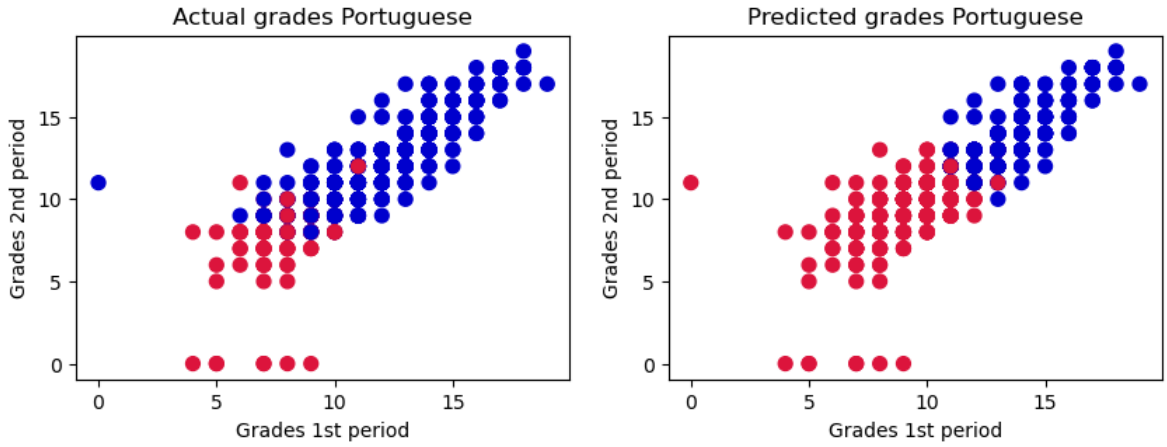

Figura 15: *Clustering* para comparação dos gráficos do *dataset* de Português

Como se pode verificar, os gráficos de previsão dos *datasets*, não estão perfeitamente separados pelos *clusters*. Seguem as figuras que apresentam a precisão da separação de *clusters*.

```
Accuracy score Math:  0.8126582278481013
Classification report:                precision    recall  f1-score   support

           0       0.64      0.99      0.78       131
           1       0.99      0.72      0.84       264

    accuracy                           0.81       395
   macro avg       0.82      0.86      0.81       395
weighted avg       0.88      0.81      0.82       395
```

Figura 16: Figura da precisão do *cluster* para o *dataset* de Matemática

```
Accuracy score Portuguese:  0.6194144838212635
Classification report:                precision    recall  f1-score   support

           0       0.29      1.00      0.45       100
           1       1.00      0.55      0.71       549

    accuracy                           0.62       649
   macro avg       0.64      0.78      0.58       649
weighted avg       0.89      0.62      0.67       649
```

Figura 17: Figura da precisão do *cluster* para o *dataset* de Português

2.6 *Deployment*

Na última fase do projeto, desenvolveu-se uma aplicação *Web*, com intuito de demonstrar a utilização dos diversos algoritmos em prática.

Esta Interface tem uma *Navbar*, que nos permite fazer a navegação entre os diferentes Cenários, bem como, fazer a escolha do *Dataset* dentro dos cenários, isto é, se se pretender realizar uma previsão do *Dataset* de Português, ou do *Dataset* de Matemática. Também permite a visualização das característica extraídas numa fase Inicial da parte referente à análise dos dados dos *Datasets*, onde se pode visualizar um resumo dos Dados, bem como os valores omissos e o comportamentos dos diferentes atributos.

Uma vez escolhido o Cenário que se pretende testar, bem como o *Dataset* que se pretende testar, a aplicação permite escolher o valor dos diversos atributos, e com este *input*, proceder à avaliação do mesmo, isto é, chegar ao *Output* esperado, com base nos valores escolhidos.

A aplicação *Web* foi desenvolvida com recurso à framework de desenvolvimento *Web* da Linguagem de Programação *Python* denominada por *Flask*.

Uma demonstração do uso da Aplicação *Web*, está presente na Figura 18, onde se pode verificar o resultado do Estudo do Cenário 1, onde se verifa de forma binária, se o estudante Pass ou não, consoantes os parâmetros colocados no formulário *Web*.

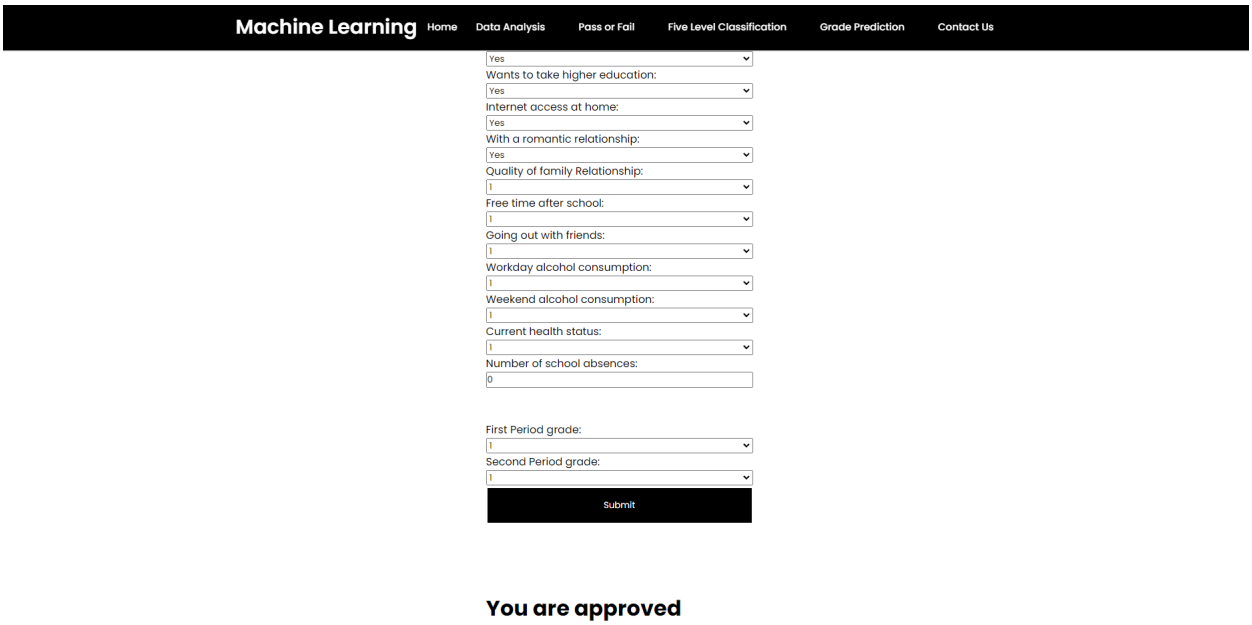

Figura 18: Aplicação Web na prática.

3 Conclusão

Em restropetiva, consegue-se afirmar que o seguinte projeto permitiu a colocação em prática de quase todos os conceitos e conteúdos lecionados ao longo do semestre na Unidade Curricular de *Machine Learning*, desde o processo de Análise dos dados do *Dataset*, a alteração dos dados presentes nos *Datasets* através de técnicas de *Feature* Engineering, a aplicação de algoritmos de aprendizagem supervisionada e aprendizagem não supervisionada com o *Tunning* de Hiperparâmetros realizado. Contudo existiram conceitos que não foram utilizados, dado que o presente trabalho não aporda *Time Series*, Sistemas de Recomendação ou *Text Mining*. Faltou ainda colocar os algoritmos realizados em *Containers* através do *Docker*.

Este projeto foi bastante enriquecedor, no sentido de proporcionar a obtenção de conhecimentos nas diversas etapas de um Projeto de *Data Mining*. Contudo, ainda existem diversas melhorias que poderiam ter sido implementadas. Primeiramente e possivelmente a mais fulcral, uma melhoria na análise dos Dados, bem como uma melhor limpeza dos dados, ou mesmo a utilização de *Data Augmentation*, de forma a se obter um *Dataset* superior. Por outro lado, poderiam se ter utilziado mais algoritmos de *Machine Learning*, tanto de aprendizagem supervisionada como de aprendizagem não-supervisionada ou mesmo a utilização de Redes Neuronais.

Outra implementação interessante seria a de utilização de Regras de Associação, de forma a perceber como os atributos se relacionam, isto é, se é possível a obtenção de sequencias de estabeleciemento de atributos.

Em suma, o seguinte trabalho foi bastante enriquecedor, dado que nos permitiu perceber acerca das diversas etapas de um Projeto de *Data Mining*, além de perceber os diversos problemas que alguém do Ramo da Ciência de Dados ou de *Machine Learning* pode enfrentar nos seus diversos projetos.

Referências

[1] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.