Uso de aprendizado de máquina para prever performance estudantil

Gianluigi Dal Toso
Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brasil
gianluigi.toso@gmail.com

Raíssa Batista de Miranda Pimentel
Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brasil
raissabmpimentel@gmail.com

***Resumo*—Este trabalho tem o intuito de avaliar o desempenho de algoritmos de aprendizado de máquina para prever a performance estudantil a partir de dados escolares, socio-econômicos e comportamentais obtidos por uma pesquisa realizada com estudantes do ensino médio português. Inicialmente foram realizados uma análise exploratória da base de dados e o seu pré-processamento e, em seguida, foram avaliados, considerando várias combinações das notas passadas como atributos preditores, os desempenhos dos modelos *KNN*, redes neurais perceptron multicamadas (MLP), máquinas de vetores de suporte (SVM), árvores de decisão (DT) e *Random Forest* (RF) na previsão do desempenho final do estudante. Por fim, concluiu-se que as tarefas de aprendizado são bem sucedidas quando utiliza-se pelo menos uma nota de período para prever o desempenho final do estudante e que os algoritmos que apresentaram os melhores resultados foram o *Random Forest* e as redes neurais perceptron multicamadas, modelos de complexa interpretação devido à abordagem de "caixa preta" proposta por estes algoritmos.**

I. Introdução

O aprimoramento do sistema educacional afeta de diferentes maneiras o desenvolvimento de um país. A educação é ponto chave para o desenvolvimento científico e tecnológico da nação e também para atingir progresso econômico a longo prazo [1]–[3].

Um dos grandes desafios da educação é aumentar a taxa de alunos que se formam sem detrimento da qualidade de ensino [4]. Nesse sentido, é de suma importância identificar cedo alunos que por ventura venham ter dificuldades acadêmicas e intervir com alguma ferramenta de apoio (aulas de reforço, acompanhamento, monitoria, etc).

A dificuldade nesse cenário é saber previamente quais os fatores que impactam o desempenho dos estudantes. Seria possível prever a performance escolar de um aluno tendo em mãos dados sociais, econômicos e desempenho escolar prévio do indivíduo? Esta é a questão que será abordada neste artigo.

O sucesso desta análise pode colaborar muito para o mapeamento e compreensão dos fenômenos que influenciam o desempenho dos estudantes, além de fornecer métricas que devem auxiliar a administração do sistema educacional no aprimoramento das técnicas de intervenção e apoio ao ensino.

Diferentemente de outros estudos que utilizam somente fatores referentes à performance acadêmica prévia e abstenção nas aulas para a previsão da performance [6], esse estudo também analisa informações sobre a vida social e econômica do estudante, que podem influenciar na previsão de sua performance final.

A. Tarefas de aprendizados propostas

Diante do problema discutido anteriormente, iremos utilizar uma abordagem de aprendizagem supervisionada, pois ela é muito utilizada para resolução de problemas semelhantes e tem fornecido modelos preditivos eficientes com base em dados da performance passada de estudantes para estudos referentes ao ensino superior e ensino à distância ao redor do mundo [6]–[8].

As tarefas de aprendizados propostas terão base nos dados coletados de duas escolas públicas da região de Alentejo em Portugal durante o período escolar de 2005 a 2006. A coleta de dados foi feita por meio de um questionário preenchido pelos estudantes, cujos dados pessoais foram cruzados com seus dados de performance escolar, como notas e número de faltas, que foram fornecidos pelas escolas [5].

Os dados coletados contém 29 (vinte e nove) atributos respondidos no questionário, como sexo, idade, educação da mãe, tamanho da família, tempo de estudo, entre outros, que podem ter relação direta com o desempenho escolar. Além disso, há quatro atributos de desempenho escolar, sendo eles o número de faltas e três notas numéricas de 0 a 20: do primeiro período, do segundo período e a final, pois em Portugal os estudantes são avaliados em três períodos, sendo a nota final determinante para o estudante ser aprovado na classe (uma nota final acima ou igual a 10 indica aprovação) [5].

Por fim, os dados coletados compõem uma base de dados da classe de português, com 649 estudantes, que será utilizada como experiência das tarefas de aprendizados propostas. Para cada tarefa proposta será utilizada a indução para criar modelos de aprendizado preditivos com os dados coletados.

Ao total, serão propostas duas tarefas de aprendizado, diferenciadas pelo atributo-alvo que será previsto:

I. Classificação para prever se há aprovação ou não.
II. Regressão para prever a nota final.

Como as notas do primeiro e segundo período podem ser importantes para a previsão da nota final, uma vez que são indicativos da dedicação e aprendizado do estudante, serão utilizadas quatro combinações dos atributos de notas para a modelagem preditiva: considerando as notas do primeiro e segundo período, considerando cada uma individualmente (duas possibilidades) e desconsiderando todas as notas.

Para avaliar o desempenho dos modelos construídos, as bases de dados serão divididas em duas partes: uma para treinamento e outra para teste. A aplicação do modelo obtido para prever o atributo-alvo na base de teste é fundamental para avaliar o seu desempenho. Para a tarefa de classificação (I), a avaliação do desempenho será medida pela porcentagem de acertos de rotulação em relação às rotulações originais do atributo-alvo. Para a tarefa de regressão (II), a avaliação do desempenho será medida pelo erro quadrático médio dos valores previstos pelo modelo em relação aos valores originais do atributo-alvo.

B. Trabalhos relacionados

O aprendizado de máquina tem sido bastante utilizado em estudos de previsão de performance estudantil e evasão no ensino superior ao redor do mundo [6]–[9]. Outros estudos utilizam o aprendizado supervisionado para analisar dados de estudantes do ensino médio dos Estados Unidos com o intuito de prever alunos com tendência a reprovar de ano ([10], [11]), o que, além de gerar custos adicionais e sobrecarregar o sistema educacional, costuma gerar impactos negativos na motivação do estudante e comprometer seu futuro acadêmico [12]. Com exceção do estudo base [5] que coletou os dados que serão utilizados neste artigo, majoritariamente esses estudos não incluem dados externos ao ambiente escolar nas análises. No entanto, é possível que fatores familiares e sociais podem ter impacto significativo na previsão da performance estudantil.

II. Base de dados

Confirme descrito na seção de introdução, os dados utilizados nesse estudo foram extraídos da classe de português de duas escolas públicas da região de Alentejo em Portugal durante o período escolar de 2005 a 2006. Esses dados tinham o objetivo principal de serem usados em um trabalho acadêmico de Data Mining para prever a performance estudantil [5]. Posteriormente, eles foram disponibilizados para uso geral no repositório de aprendizado de máquina do UCI [13], onde foram obtidos para a realização desse estudo.

Os dados se basearam em duas fontes: relatórios do desempenho estudantil individual, fornecidos pelas escolas em papel, e um questionário de 37 (trinta e sete) questões respondidas por 788 estudantes. Das respostas obtidas, 111 respostas foram descartadas por falta de identificação dos alunos nos relatórios escolares. Além disso, algumas questões respondidas no questionário foram descartadas por falta de valor discriminativo relevante. Por fim, os dados coletados e filtrados foram separados em uma base de dados da classe de português, com 649 registros [5].

Foram extraídos 4 (quatro) atributos dos relatórios escolares: o número de faltas e três notas numéricas de 0 a 20 (do primeiro período, do segundo período e a final). Além disso, foram extraídos 29 (vinte e nove) atributos dos questionários, que estão relacionados com variáveis demográficas (trabalho e educação dos pais), sociais (consumo de álcool, frequência de saída com amigos) e escolares (tempo de estudo semanal, número de reprovações anteriores), que podem afetar a performance estudantil.

A. Descrição dos atributos

A tabela contida no Apêndice A apresenta a descrição de cada um dos 33 (trinta e três) atributos analisados nesse estudo, explicitando ainda para cada atributo seu tipo, escala e domínio.

III. Análise exploratória dos dados

Essa seção abordará a análise exploratória dos dados de todos os 33 atributos contidos na base de dados da classe de português, uma vez que diante das tarefas de aprendizado propostas todos os atributos poderão ser utilizados para gerar e validar os modelos de aprendizado de máquina.

A. Atributos qualitativos

Os atributos qualitativos considerados para a análise exploratória foram se há suporte educacional extra, suporte educacional familiar, aulas da disciplina extras pagas e atividades extra-curriculares. Além disso, também foram analisados o número de disciplinas com reprovação, o tempo livre após a escola, o tempo de estudo semanal e a condição de saúde atual. Esses dados sócio-econômicos são muito relevantes para o desempenho escolar do aluno, e por isso eles serão abordados mais detalhadamente. Os gráficos de frequência dos 8 atributos qualitativos mais relevantes para a base de dados estão contidos nas Figuras 1 e 2.

De acordo com esses gráficos, pode-se concluir que a maioria dos alunos tem poucas reprovações anteriores, tem suporte educacional familiar, não tem suporte educacional extra, tem tempo de estudo semanal de até 5 horas (classificação 1 e 2) e estão em boas condições de saúde. Outros pontos a serem destacados é que a classificação de ocorrência de atividades extra-curriculares está distribuída em quase 50% dos dados. Além disso, a classificação ordinal de tempo livre possui frequência semelhante à distribuição normal em torno do ponto central que é a classificação 3, de uma escala de 1 (“pouco”) a 5 (“muito”).

Além disso, outro ponto a ser destacado é que a maioria dos alunos da classe de português não paga aulas extras, o

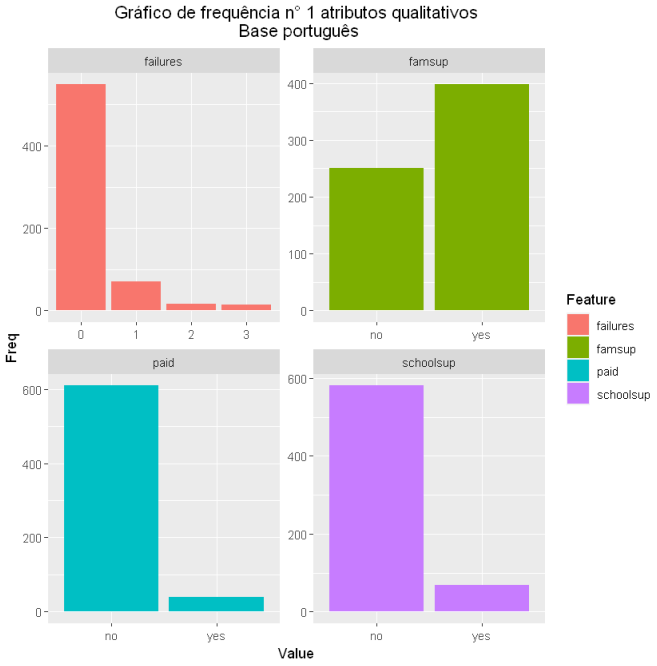

Figura 1. Gráfico de frequência de 4 atributos qualitativos da base de dados.

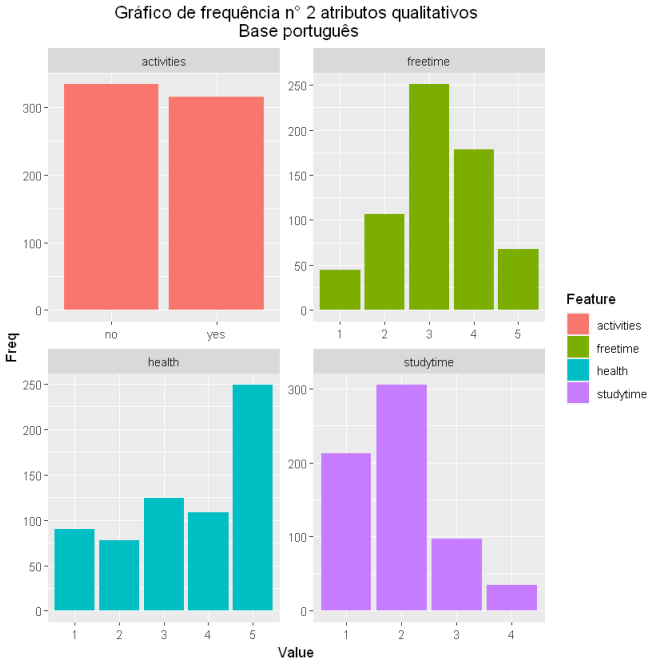

Figura 2. Gráfico de frequência de 4 atributos qualitativos da base de dados.

que reflete que muitos alunos não apresentam dificuldades nessa matéria a ponto de pagarem aulas extras dessa disciplina.

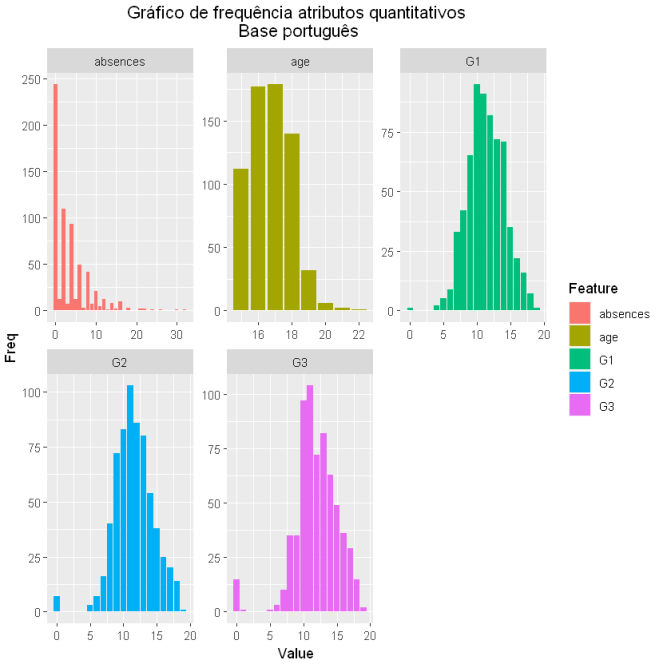

Figura 3. Gráfico de frequência dos atributos quantitativos da base de dados.

B. Atributos quantitativos

O gráfico da frequência dos atributos quantitativos para a base de dadosestá representado na Figura 3.

O gráfico mostra que a idade dos alunos está bem concentrada na faixa de 16 a 19 anos. Também nota-se que a maioria falta pouco e que, quanto maior o número de faltas, menor a sua frequência. Por fim, as notas dos alunos se distribuem de forma semelhante à uma distribuição normal centrada próximo da nota 10, que é a mínima para a aprovação. Contudo, há ainda poucos alunos que tiram a nota mínima (0), revelando que há ainda uma pequena porcentagem de alunos que apresentam muitas dificuldades na disciplina que estudam.

A matriz de correlação dos atributos quantitativos para a base de dados está representada na Figura 4.

Essa matriz mostra que há grande correlação entre os atributos **G1**, **G2** e **G3**, com módulo maior do que 0.8, enquanto as outras correlações são pouco relevantes, com módulo menor do que 0.2. Isso faz com que seja muito provável que os atributos **G1** e principalmente **G2**, por causa do seu alto valor de correlação, sejam bons previsores para o valor de **G3**. Essa relação já era esperada e justifica as tarefas de aprendizados escolhidas.

Outra correlação relevante, mas pequena, é entre o número de faltas (atributo **absences**) e a idade (atributo **age**). Ou seja, há uma pequena tendência de haver mais faltas com o aumento da idade, o que ocorre provavelmente porque pessoas mais velhas tenham outros compromissos, como trabalho, para conciliar com os estudos.

A Tabela I apresenta um compêndio de dados esta-

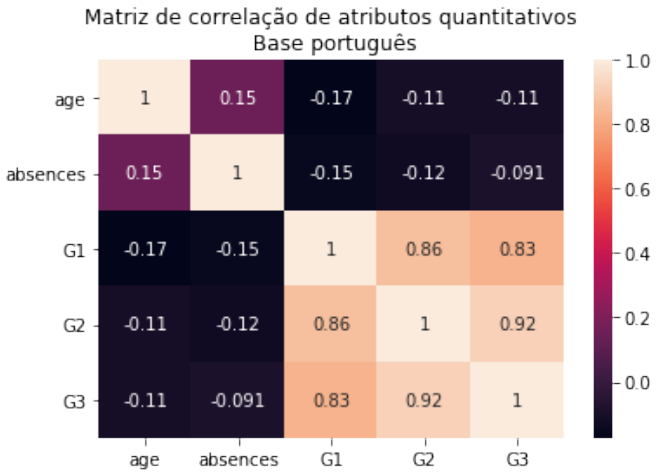

Figura 4. Matriz de correlação dos atributos quantitativos da base de dados.

tísticos referentes aos dados quantitativos da base de dados. Vale ressaltar que o segundo quartil é equivalente à mediana dos dados.

Tabela I
Estatísticas referentes aos atributos quantitativos da base de dados.

Feature	**absences**	**age**	**G1**	**G2**	**G3**
Valor Mínimo	0	15	3	0	0
Valor Máximo	75	22	19	19	20
Intervalo	75	7	16	19	20
Primeiro Quartil	0	16	8	9	8
Segundo Quartil	4	17	11	11	11
Terceiro Quartil	8	18	13	13	14
Desvio Padrão	8,00	1,28	3,32	3,76	4,58
Média	5,71	16,70	10,91	10,71	10,42
Variância	64,05	1,63	11,02	14,15	20,99
Obliquidade	3,644	0,463	0,239	-0,428	-0,727
Curtose	21,307	-0,031	-0,712	0,586	0,366

A Figura 5 apresenta o *boxplot* modificado para a base de dados. Um *boxplot* modificado é um diagrama de *Box* e *Whisker* com os limites inferior e superior da linha indo até os valores de mínimo e máximo apenas se esses valores não forem muito distantes do primeiro e terceiro quartis respectivamente, sendo essa distância no máximo o valor de 1.5 multiplicado pelo intervalo entre quartis. Nesse caso, os valores distantes são representados como pontos e são considerados *outliers*.

Percebe-se pelo diagrama de *boxplot* que, apesar da mediana para o número de faltas ser próximo de zero, ou seja a maioria dos alunos não costuma faltar muito, alguns poucos alunos que faltam tem um número muito grande de faltas. Esse padrão é percebido tanto pelo número de *outliers* nos gráficos de *boxplot* quando pelos valores de obliquidade notados na Tabela I.

É interessante também perceber o espalhamentos das notas. As notas são mais distribuídas próximas ao valor mediano e possuem alguns valores de *outliers*. Possivelmente a estrutura dos testes de português fazem com que

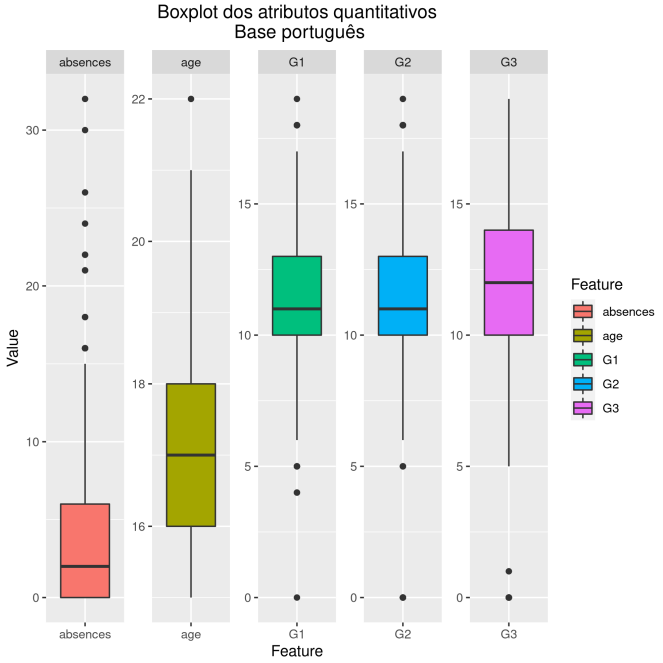

Figura 5. *Boxplot* modificado para os atributos quantitativos da base de dados.

seja mais fácil conseguir pontuar para não zerar a prova, no entanto seja difícil conquistar a nota mais alta possível. Isso é comum em testes avaliativos envolvendo redação por exemplo.

Analisando-se conjuntamente os gráficos de frequência para os atributos quantitativos (Figura 3) e os dados estatísticos para esses atributos (Tabela I) percebe-se que as notas comportam-se como distribuições normais, excetuando-se o fato de que existe um pico na nota zero, fato esse que foi mantido em mente ao avaliar os resultados obtidos pelas tarefas de aprendizado propostas.

IV. Pré-Processamento dos dados

Essa seção abordará o pré-processamento realizado em cima da bases de dados de modo a remover informações irrelevantes e torná-las mais apropriadas para a utilização de alguns algoritmos de aprendizado de máquina. Uma vez que o funcionamento e o desempenho de diversos algoritmos está diretamente relacionado aos dados de entrada, a etapa de pré-processamento é fundamental para uma análise adequada.

A base de dados a ser trabalhada já foi construída visando a sua utilização para estudos de aprendizado de máquina e portanto tem-se uma base de dados sem diversos problemas comuns decorrentes da agregação de dados como: inconsistências, redundâncias ou dados incompletos. Além disso, pela análise exploratória de dados, percebeu-se a qualidade dos dados em relação à distribuição e balanceamento dos mesmos.

Dessa forma, as atividades de pré-processamento realizadas visaram, em geral, a adequar a base de dados para a

sua utilização pelos algoritmos de aprendizado, removendo dados que não agregam informação e transformando atributos simbólicos em atributos numéricos.

A. Eliminação manual de atributos

A escola do estudante (atributo **school**) pode não ser uma informação relevante para as tarefas propostas, pois a escola em si não deve influenciar o desempenho do estudante, uma vez que seus atributos sociais, econômicos e escolares devem ser mais relevantes. Muitos desses atributos, por sua vez, estão bem distribuídos na base de dados, conforme discutido na seção da análise exploratória de dados. Assim, concluiu-se que esse atributo não contribui com informação relevante para a análise e por isso ele foi eliminado da base de dados.

B. Transformação de dados

A base de dados apresenta um grande número de atributos qualitativos e simbólicos. Alguns algoritmos de aprendizado exigem que os atributos sejam numéricos. Sendo assim, é preciso realizar uma transformação de atributos simbólicos para numéricos em cima desses dados. Será descrito inicialmente como foi feita a conversão de atributos simbólicos para atributos numéricos e posteriormente será descrito como se ajustou esses dados para que a escala numérica de um atributo não fique muito destoante das escalas dos demais atributos.

1) Atributos simbólicos para numéricos

O atributo **famsize** é binário e qualitativo ordinal. Neste caso, a conversão para numérico foi feita de tal forma que o valor com a menor hierarquia ("LE3") seja convertido para o valor 0, e o valor com a maior hierarquia (valor "GT3") seja convertido para o valor 1.

Os atributos qualitativos ordinais com mais de dois valores já estão em um formato numérico coerente. No entanto, os atributos qualitativos nominais são todos simbólicos e precisam ser transformados em atributos numéricos. Não existe hierarquia ou relação de ordem entre os valores desses atributos e é necessário que essa propriedade seja mantida após a conversão. Dessa forma, a transformação não pode ser realizada meramente atribuindo um valor inteiro para cada símbolo.

A transformação foi então realizada utilizando a codificação *One Hot Encoding*. Essa codificação prevê remover a coluna referente ao atributo simbólico e adicionar uma coluna para cada possível valor e então utilizar valores binários indicando a presença ou não daquela característica. Por exemplo o atributo **guardian** da lugar aos atributos **guardian_mother**, **guardian_father** e **guardian_other** e um exemplo em que o atributo **guardian** assumia o valor "father", assumirá, para os novos atributos, os valores 0, 1 e 0 respectivamente.

2) Normalização para ajuste de escala

Para que a escala dos novos atributos numéricos não fique discrepante, realizou-se uma normalização destes novos dados utilizando a normalização *Z-Score*. Sendo x o dado original, μ a média dos dados originais e σ o desvio padrão dos dados originais, o novo dado Z é calculado como sendo:

$$Z = \frac{x - \mu}{\sigma}$$

Ou seja, o valor normalizado por *Z-Score* é uma medida de quanto o dado se afasta da média da distribuição em termos de desvio padrão.

Por fim, executando todas as etapas citadas anteriormente, os dados pré processados se resumem a uma base com 649 linhas (observações) e 56 colunas (atributos).

V. Definição matemática das tarefas de aprendizado

Considerando dois conjuntos $X_{treinamento} = \{x_1, ..., x_l\}$ e $X_{teste} = \{x_{l+1}, ..., x_{l+u}\}$ de tal forma que $x_i \in \mathbb{R}^D$ para todo i. O valor de D pode variar de 53 a 55, pois ele depende da escolha de quais notas serão consideradas como atributos previsores para as tarefas. Por fim, o valor de l e u dependem da forma de particionamento dos dados em treinamento e teste.

Para a tarefa de classificação, cada ponto $x_i \in X_{treinamento} \cup X_{teste}$ é associado a um rótulo $y_i \in \{aprovado,\ não\ aprovado\}$. Assim, nessa tarefa o objetivo é utilizar os dados de $X_{treinamento}$ para treinar um modelo de aprendizado de máquina que associe rótulos corretos aos dados do conjunto X_{teste}. Para a tarefa de regressão, o raciocínio é semelhante, mas cada ponto $x_i \in X_{treinamento} \cup X_{teste}$ é associado com um número $y_i \in \mathbb{R}$. Portanto, nessa tarefa o objetivo é utilizar os dados de $X_{treinamento}$ para treinar um modelo de aprendizado de máquina que associe um número mais próximo possível dos números associados aos dados do conjunto X_{teste}.

VI. Avaliação e validação do aprendizado de máquina

A. Classificação para prever se há aprovação ou não

Para a validação do aprendizado da tarefa de classificação, foi utilizado o método de validação cruzada com 10 pastas estratificadas da base de dados. Para cada pasta utilizada para teste, foram utilizados os melhores hiperparâmetros encontrados para os modelos *KNN*, redes neurais perceptron multicamadas (MLP), máquinas de vetores de suporte (SVM), árvores de decisão (DT) e *Random Forest* (RF).

Os melhores hiperparâmetros foram escolhidos com base na validação cruzada dos modelos gerados por diferentes combinações de valores selecionados de forma arbitrária para os hiperparâmetros. Essa validação cruzada, por sua vez, utilizou 3 pastas estratificadas da base de dados sem a partição utilizada para teste. A tabela contida no Apêndice

B contempla as combinações de hiperparâmetros testadas. A medida de performance utilizada para todos os testes foi a porcentagem de acertos dos rótulos de aprovação.

Essa validação foi executada para 4 (quatro) combinações distintas de notas usadas como preditores: desconsiderando as notas do primeiro e segundo período ([ ]), contendo apenas a nota do primeiro período ([G1]), contendo apenas a nota do segundo período ([G2]) e contendo as notas do primeiro e segundo período ([G1, G2]). A porcentagem média de acertos e seu desvio padrão para cada combinação de atributos e cada modelo avaliado estão contidos na Tabela II.

Vale ressaltar que, em média, o uso das notas do primeiro e segundo período apresenta melhor resultado do que o uso da nota do segundo período, que apresenta melhor resultado do que o uso da nota do primeiro período e que, por fim, apresenta melhor resultado do que não considerar as notas. Isso pode ser justificado pelo fato de que a prova final é feita mais próxima do segundo período. Além disso, o uso de todas as notas do período pode oferecer uma visão mais ampla do desempenho passado do estudante, sendo um bom preditor para o seu desempenho futuro.

A partir dos resultados de desempenho obtidos para cada combinação de atributos, levantou-se a hipótese nula de que cada distribuição gerada por um modelo de aprendizado seja igual às demais distribuições geradas pelos outros modelos utilizados. Para confirmar ou refutar essa hipótese, foi realizado o teste de Friedman considerando uma significância de 0.05, ou seja, caso o *p-value* obtido para as distribuições de cada combinação de atributos seja menor do que 0.05, a hipótese nula é refutada.

Os valores de *p-values* obtidos para cada combinação de atributos foram:

I. $1.13 \cdot 10^{-1}$, desconsiderando as notas do primeiro e segundo período.
II. $1.09 \cdot 10^{-3}$, considerando a combinação de notas contendo apenas a nota do primeiro período.
III. $3.24 \cdot 10^{-5}$, considerando a combinação de notas contendo apenas a nota do segundo período.
IV. $1.90 \cdot 10^{-5}$, considerando a combinação de notas contendo as notas do primeiro e do segundo período.

Portanto, a hipótese nula é aceita apenas quando se desconsidera as notas do primeiro e segundo período, e se conclui que, para esse caso, todos os modelos apresentaram distribuições estatisticamente equivalentes. Além disso, nota-se que, para essa combinação de atributos, os resultados médios de desempenho estão muito próximos ou abaixo da porcentagem de aprovados contido na base de dados original, que é de 84.59%. Assim, a desconsideração das notas do primeiro e do segundo período fornece um modelo bem próximo do que o de classificar todos os alunos como aprovados. Por isso, desconsiderar as notas como atributos previsores não oferece um aprendizado satisfatório.

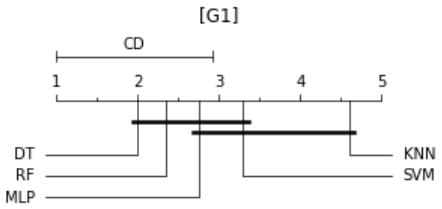

Figura 6. Diagrama da distância crítica de cada modelo para a combinação de notas contendo apenas a nota do primeiro período.

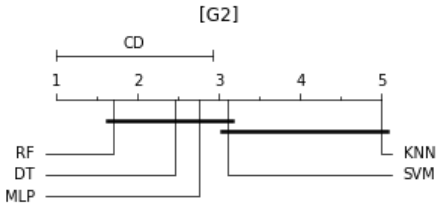

Figura 7. Diagrama da distância crítica de cada modelo para a combinação de notas contendo apenas a nota do segundo período.

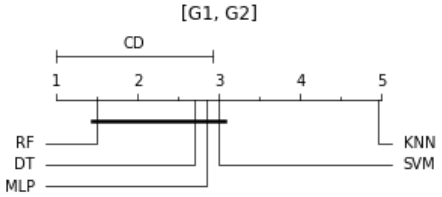

Figura 8. Diagrama da distância crítica de cada modelo para a combinação de notas contendo as notas do primeiro e segundo período.

Além disso, a hipótese nula é refutada para as outras combinações de atributos e se conclui que, nesses casos, as distribuições obtidas para cada modelo são distintas. Para analisar a independência estatística entre cada um dos modelos utilizados, executou-se o teste *post-hoc* de Nemenyi para as combinações de notas utilizadas, com exceção da que desconsidera as notas do primeiro e segundo período. Esse teste obteve os resultados representados nas Figuras 6, 7 e 8, que contêm os diagramas de distância crítica de cada modelo e de cada combinação de atributos.

A partir desses diagramas pode-se concluir que, apesar do método de *Random Forest* ter obtido os melhores resultados em média, não há diferença estatística significativa entre esse método, redes neurais perceptron multicamadas, árvores de decisão e máquina de vetores de suporte para as tarefas de classificação propostas. O que pode ser afirmado é que o método *KNN* é considerado estatisticamente distinto e de menor performance do que os quatro modelos citados anteriormente para a combinação contendo as notas do primeiro e segundo período.

Para a combinação contendo apenas a nota do primeiro período, não há diferença estatística entre os modelos

Tabela II
PORCENTAGEM MÉDIA DE ACERTOS E SEU DESVIO PADRÃO PARA CADA COMBINAÇÃO DE ATRIBUTOS E CADA MODELO AVALIADO NA TAREFA DE CLASSIFICAÇÃO.

Combinação	**MLP**	**SVM**	**KNN**	**DT**	**RF**
[ ]	83.98 ± 3.00	85.51 ± 1.82	82.90 ± 3.63	85.37 ± 3.40	84.75 ± 1.32
[G1]	88.75 ± 3.70	87.99 ± 4.14	84.13 ± 2.59	89.99 ± 4.17	89.53 ± 3.60
[G2]	90.60 ± 2.75	90.30 ± 2.71	83.83 ± 2.98	91.06 ± 3.30	93.07 ± 2.73
[G1, G2]	90.75 ± 2.28	90.60 ± 2.95	84.90 ± 2.35	91.22 ± 2.71	93.53 ± 2.38

KNN, redes neurais perceptron multicamadas e máquina de vetores de suporte. Isso pode ter ocorrido devido à menor performance apresentada pelo uso único das notas do primeiro período. Além disso, para a combinação contendo apenas a nota do segundo período, não há diferença estatística entre os modelos *KNN* e máquina de vetores de suporte, o que pode ter ocorrido devido à menor performance apresentada pelo uso único das notas do segundo período em comparação com o uso das duas notas em conjunto.

Portanto, qualquer um dos modelos de *Random Forest*, redes neurais perceptron multicamadas, árvores de decisão e máquina de vetores de suporte podem ser utilizados para resolver a tarefa de aprendizado proposta, uma vez que eles apresentaram um aprendizado bem sucedido, com porcentagem média de acertos significantemente maior do que a porcentagem de aprovados contido na base de dados original, que é de 84.59%. Além disso, considerando o desvio padrão a porcentagem de acertos também fica acima de 84.59% para esses modelos. Contudo, para o modelo *KNN* isso não ocorre. De fato, ele possui, para duas combinações de notas, desempenho inferior do que os três modelos citados anteriormente de acordo com o teste de Nemenyi.

B. Regressão para prever a nota final

Para a validação da tarefa de aprendizado de regressão para prever a nota final também será utilizado o método de validação cruzada utilizando-se 10 pastas. No entanto, como o atributo alvo é numérico, não serão utilizadas pastas estratificadas.

Ademais, de forma análoga ao realizado para a atividade de classificação, para cada modelo avaliado, dentre um conjunto de hiperparâmetros escolhidos arbitrariamente, utilizou-se validação cruzada com 3 partições da base de dados sem a partição de teste para selecionar os melhores hiperparâmetros de cada modelo. A tabela contida no Apêndice C contempla as combinações de hiperparâmetros testadas. Foram avaliados os modelos *KNN*, redes neurais perceptron multicamadas (MLP), máquinas de vetores de suporte (SVM), árvore de decisão (DT) e *Random Forest* (RF). Por se tratar de uma regressão, a medida de performance utilizada para esta análise foi o erro quadrático médio entre a nota final prevista pela regressão e a nota final real do estudante.

Conforme realizado para a tarefa de classificação, a validação foi executada para 4 (quatro) combinações distintas de notas usadas como preditores: desconsiderando as notas do primeiro e segundo período ([ ]), contendo apenas a nota do primeiro período ([G1]), contendo apenas a nota do segundo período ([G2]) e contendo as notas do primeiro e segundo período ([G1, G2]). A Tabela III contém os resultados médios do erro quadrático médio e o seu desvio padrão obtidos nas avaliações de cada modelo e para cada combinação de atributos.

Levando em conta os resultados ilustrados pela Tabela III, percebe-se que o erro quadrático médio para o caso em que não se leva nenhuma nota em consideração é muito grande, principalmente porque a nota é um atributo que pertence ao intervalo entre 0 e 20. Dessa forma, este caso não apresenta resultado satisfatório e análises feitas a partir dele devem ser menos relevantes para a avaliação e validação dos modelos.

Novamente percebe-se que utilizar somente a segunda nota traz resultados melhores do que utilizar somente a nota do primeiro período e que o melhor resultado é aquele obtido quando consideram-se as notas dos dois períodos.

Assim como na tarefa de classificação, levantou-se a hipótese nula de que cada distribuição gerada por um modelo de aprendizado seja igual às demais distribuições geradas pelos outros modelos utilizados. Para confirmar ou refutar essa hipótese, foi realizado o teste de Friedman considerando uma significância de 0.05.

Os valores de *p-values* obtidos para cada combinação de atributos foram:

I. $2.89 \cdot 10^{-6}$, desconsiderando as notas do primeiro e segundo período.
II. $7.45 \cdot 10^{-7}$, considerando a combinação de notas contendo apenas a nota do primeiro período.
III. $1.77 \cdot 10^{-7}$, considerando a combinação de notas contendo apenas a nota do segundo período.
IV. $3.01 \cdot 10^{-7}$, considerando a combinação de notas contendo as notas do primeiro e do segundo período.

Portanto, a hipótese nula é refutada para cada combinação de atributos, chegando à conclusão de que as distribuições obtidas para cada modelo são distintas. Para analisar a independência estatística entre cada um dos modelos utilizados, também se executou o teste *post-hoc* de Nemenyi. Esse teste obteve os resultados representados nas Figuras 9, 10, 11 e 12, que contêm os diagramas de distância crítica de cada modelo e de cada combinação de atributos.

Pelos diagramas obtidos por meio do teste de Nemenyi, conclui-se que, com exceção do caso em que não se leva

Tabela III
Resultados médios do erro quadrático médio e seu desvio padrão para cada combinação de atributos e cada modelo avaliado na tarefa de regressão.

Combinação	MLP	SVM	KNN	DT	RF
[ ]	13.18 ± 2.99	8.22 ± 3.44	15.15 ± 3.93	15.16 ± 3.99	7.96 ± 2.83
[G1]	4.93 ± 1.28	6.43 ± 3.07	12.59 ± 3.84	6.35 ± 3.68	3.27 ± 1.87
[G2]	3.33 ± 1.34	5.66 ± 2.65	10.92 ± 1.90	3.00 ± 1.90	1.73 ± 1.11
[G1, G2]	2.90 ± 1.22	4.50 ± 2.42	9.46 ± 1.78	3.32 ± 1.67	1.64 ± 1.03

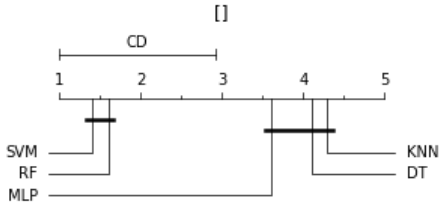

Figura 9. Diagrama da distância crítica de cada modelo para a tarefa de regressão sem levar em consideração as notas do primeiro e segundo período.

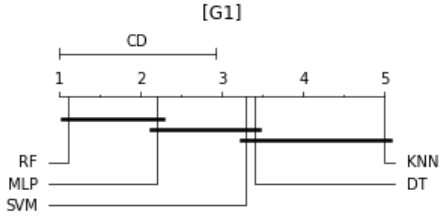

Figura 10. Diagrama da distância crítica de cada modelo para a tarefa de regressão contendo a combinação de notas contendo apenas a nota do primeiro período.

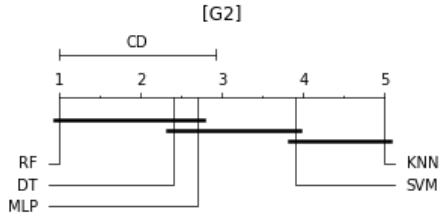

Figura 11. Diagrama da distância crítica de cada modelo para a tarefa de regressão contendo a combinação de notas contendo apenas a nota do segundo período.

nenhuma nota em consideração, os métodos de *Random Forest* e redes neurais perceptron multicamadas são estatisticamente idênticos e são, em geral, superiores aos demais métodos para a tarefa de regressão da nota final. As árvores de decisão apresentaram resultados estatisticamente independentes a esses dois modelos apenas para a combinação de atributos contendo a nota do primeiro período. Isso pode ter ocorrido devido à menor performance apresentada pelo uso único das notas do primeiro período, já que a nota do segundo período se mostrou mais relevante para a previsão da nota final.

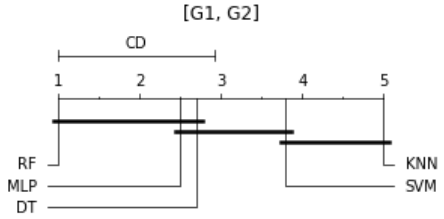

Figura 12. Diagrama da distância crítica de cada modelo para a tarefa de regressão contendo a combinação de notas contendo as notas do primeiro e segundo período.

Uma diferença ocorrida em relação aos resultados obtidos para a validação análoga durante a atividade de classificação é que o método de máquina de vetores de suporte para a atividade de regressão não foi considerada estatisticamente idêntica ao método *Random Forest* em nenhuma das combinações de atributos contendo pelo menos uma nota e foi, no entanto, consideradam nesses casos, estatisticamente idêntica ao método *KNN*, que obteve os piores resultados para todos os conjuntos de atributos.

Com relação à desconsideração das notas dos períodos, nota-se que os melhores modelos obtidos nesse caso foram os modelos de máquinas de vetores de suporte, que tem desempenho baixo nos outros casos discutidos, e *Random Forest*, estatisticamente iguais. Além disso, os modelos de redes neurais perceptron multicamadas, *KNN* e árvores de decisão, também estatisticamente iguais, possuíram desempenho bastante inferior aos dois modelos citados anteriormente. Essa grande diferença da comparação estatística dos modelos em relação às outras combinações de atributos pode ter direta relação com seu baixo desempenho, representado por altos valores de erros quadráticos médio.

A partir dos valores para os resultados médios do erro quadrático médio juntamente com os seus desvios padrão e também com os diagramas de distância crítica dos modelos, é possível concluir que tanto o modelo de *Random Forest* quanto o modelo de redes neurais perceptron multicamadas são as melhores escolhas para a realização da previsão da nota final por meio de regressão. Como as notas finais podem variar de 0 a 20, os erros quadráticos médios apresentados por esses métodos, considerando pelo menos uma nota como atributo preditor, são adequados. Portanto, a aprendizagem é satisfatória para essas combinações de atributos.

VII. Conclusões e trabalhos futuros

De acordo com as métricas de desempenho utilizadas, que são a porcentagem de acertos para a tarefa de classificação e o erro quadrático médio para a tarefa de regressão, pode-se concluir que as tarefas de classificação para previsão de aprovação dos estudantes e de regressão para previsão da nota final dos estudantes tiveram aprendizado bem sucedido para a combinação de atributos que possuía pelo menos uma nota (do primeiro ou segundo período) como atributo preditor. Os modelos que apresentaram os melhores resultados de desempenho foram os de *Random Forest* e redes neurais perceptron multicamadas, sendo seguidos pelas árvores de decisão.

Com relação aos algoritmos utilizados, sabe-se que as máquinas de vetores de suporte (SVM) são bastante sensíveis à escolha da função de *kernel*. Os resultados obtidos utilizaram as funções mais comuns, porém é possível utilizar uma função de *kernel* personalizada que obtivesse melhores resultados. No entanto, encontrar uma função de *kernel* e ajustar parâmetros para esta técnica não é uma tarefa simples. Com relação ao *KNN*, pode-se associar o seu baixo desempenho nas tarefas devido à maldição da dimensionalidade, visto que a base em questão possui um elevado número de atributos.

Já para os modelos que apresentaram a melhor performance, sendo eles *Random Forest* e redes neurais perceptron multicamadas, encontra-se o recorrente dilema na área de aprendizado de dados entre performance e facilidade de interpretação. Apesar de possuírem a melhor performance, devido à abordagem "caixa preta" destes algoritmos, a interpretação dos modelos elaborados é uma tarefa muito complexa e por consequência é difícil tanto descobrir quais atributos são mais relevantes para o problema e entender a associação entre eles, quanto identificar fatores que poderiam ser alterados para aumentar a performance dos algoritmos. Nesse sentido, o modelo de árvore de decisão (DT), mostra-se como uma boa alternativa entre performance e capacidade de interpretação caso seja desejável investigar mais a fundo o impacto de cada atributo no desempenho acadêmico final do estudante.

Para trabalhos futuros, este estudo pode ser estendido para incluir os resultados obtidos pela utilização de mais arquiteturas de redes neurais artificiais como, por exemplo, redes neurais convolucionais, redes neurais recorrentes ou redes neurais profundas.

Além das constatações discutidas anteriormente, notou-se que em todos os modelos utilizados o uso das notas do primeiro e segundo período apresenta melhor resultado do que o uso da nota do segundo período, que, por sua vez, apresenta melhor resultado do que o uso da nota do primeiro período. Isso pode ser justificado pelo fato de que a prova final é feita mais próxima do segundo período. Assim, os estudantes em geral repetem um desempenho semelhante ao obtido na nota do segundo período. Além disso, levar em consideração mais de uma nota oferece um maior histórico do desempenho escolar, o que contribui para produzir melhores previsões do desempenho futuro.

Contudo, quando não se levava em consideração nenhuma das notas anteriores, o aprendizado não foi satisfatório, obtendo-se, para a tarefa de classificação, modelos que apresentavam desempenho próximo à classificação de todos os alunos como aprovados e, para a tarefa de regressão, valores muito elevados de erro quadrático médio.

Por isso, pode-se concluir que o desempenho anterior dos alunos é um atributo fundamental para prever o seu desempenho futuro. Porém, a base de dados utilizada possui muitos atributos relacionados apenas à condição sócio-econômica dos estudantes, e os atributos mais relacionados com desempenho dos alunos são apenas as notas do primeiro e do segundo período, que se mostraram muito importantes para que o aprendizado fosse bem sucedido. Para que os desempenhos dos modelos serem ainda maiores, a base de dados poderia ter mais atributos relacionados ao desempenho passado dos estudantes, como a participação em sala de aula, o desempenho em atividades feitas em classe e em casa, entre outros possíveis atributos.

Outra abordagem possível a ser tomada a partir desse trabalho é aplicar as mesmas tarefas de aprendizado à base de dados da classe de matemática extraída para o trabalho acadêmico que abordou a mesma base de dados utilizada nesse trabalho [5]. Assim, pode-se avaliar se as conclusões acerca do aprendizado e dos atributos das bases de dados são as mesmas obtidas nesse estudo.

Referências

[1] L. Woessmann. "The economic case for education." em *Education Economics*, vol. 24.1, 2016, pp. 3-32.

[2] E. A. Hanushek. "Economic growth in developing countries: The role of human capital." em *Economics of Education*, vol. 37, 2013, pp. 204-212. [Online] Disponível em: https://doi.org/10.1016/j.econedurev.2013.04.005

[3] T. Zhu, P. Hua-Rong e Z. Yue-Jun. "The influence of higher education development on economic growth: evidence from central China." em *Higher Education Policy*, vol. 31.2, 2018, pp. 139-157.

[4] R. Schendel e M. Tristan. "Expanding higher education systems in low-and middle-income countries: the challenges of equity and quality." em *Higher education* vol. 72.4, 2016, pp. 407-411. [Online] Disponível em: https://link.springer.com/content/pdf/10.1007/s10734-016...

[5] P. Cortez e A. Silva. "Using Data Mining to Predict Secondary School Student Performance," em *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, Porto, Portugal, 2008, pp. 5-12. [Online] Disponível em: http://www3.dsi.uminho.pt/pcortez/student.pdf

[6] J. Xu, K. H. Moon e M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," em *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742-753, 2017.

[7] A. Ahadi, *et al.* "Exploring machine learning methods to automatically identify students in need of assistance," em *Proceedings of the eleventh annual International Conference on International Computing Education Research*, 2015.

[8] S. Kotsiantis, C. Pierrakeas e P. Pintelas. "Efficiency of machine learning techniques in predicting students' performance in distance learning systems," Educational Software Development Laboratory Department of Mathematics, University of Patras, Grécia, 2002.

[9] G. Kantorski, *et al.* "Predição da evasão em cursos de graduação em instituições públicas," em *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 27, no. 1, 2016. [Online] Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/6776

[10] H. Lakkaraju, *et al.* "A machine learning framework to identify students at risk of adverse academic outcomes," em *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015. [Online] Disponível em: https://dl.acm.org/doi/pdf/10.1145/2783258.2788620

[11] E. Aguiar, *et al.* "Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time," em *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 2015. [Online] Disponível em: https://dl.acm.org/doi/pdf/10.1145/2723576.2723619

[12] R. C. A. Duarte. "O efeito da reprovação na motivação dos alunos," Instituto Politécnico do Porto, Escola Superior de Educação, 2016. [Online] Disponível em: https://recipp.ipp.pt/bitstream/10400.22/8442/1/DM_Rute...

[13] UCI Machine Learning Repository. [Online] Disponível em: http://archive.ics.uci.edu/ml/index.php

[14] D. P. Kingma, J. L. Ba. "Adam : A method for stochastic optimization." em *ICLR 2015*, 2015. [Online] Disponível em: https://arxiv.org/pdf/1412.6980.pdf

Apêndice A
Descrição dos atributos com tipo e escala.

Atributo	Descrição	Tipo	Escala	Domínio
school	Escola do estudante	Qualitativo	Nominal	{"*GP*":Gabriel Pereira, "*MS*":Mousinho da Silveira}
sex	Sexo do estudante	Qualitativo	Nominal	{"*F*":Feminino, "*M*": Masculino}
age	Idade do estudante (anos)	Quantitativo Discreto	Racional	[15-22]
address	Tipo de endereço do estudante	Qualitativo	Nominal	{"*U*":Urbano , "*R*":Rural}
famsize	Tamanho da família	Qualitativo	Ordinal	{"*LE3*":Menor ou igual à 3, "*GT3*":Maior que 3}
Pstatus	Estatus de cohabitação dos pais	Qualitativo	Nominal	{"*T*":Juntos, "*A*":Separados}
Medu	Educação da mãe	Qualitativo	Ordinal	0:Nenhuma, 1:Ensino Fundamental (até a 4ª série), 2:Ensino Fundamental (entre 5ª e 9ª séries), 3:Ensino médio, 4:Ensino superior
Fedu	Educação do pai	Qualitativo	Ordinal	0:Nenhuma, 1:Ensino Fundamental (até a 4ª série), 2:Ensino Fundamental (entre 5ª e 9ª séries), 3:Ensino médio, 4:Ensino superior
Mjob	Emprego da mãe	Qualitativo	Nominal	"*teacher*":Professora, "*health*":Profissional da saúde, "*services*":Serviços civís (administrativo ou polícia), "*at_home*":Doméstica, "*other*":Outros empregos
Fjob	Emprego do pai	Qualitativo	Nominal	"*teacher*":Professor, "*health*":Profissional da saúde, "*services*":Serviços civís (administrativo ou polícia), "*at_home*":Doméstico, "*other*":Outros empregos
reason	Motivo de escolha da escola	Qualitativo	Nominal	"*home*": Proximidade da casa, "*reputation*": Reputação da escola, "*course*":Preferência por curso, "*other*":Outros motivos
guardian	Tutor do estudante	Qualitativo	Nominal	{"*mother*": Mãe, "*father*": Pai, "*other*":Outra pessoa}
traveltime	Tempo de deslocamento até a escola	Qualitativo	Ordinal	1:Menos de 15 minutos, 2:De 15 a 30 minutos, 3:De 30 minutos a 1 hora, 4:Mais de uma hora
studytime	Tempo de estudo semanal	Qualitativo	Ordinal	1:Menos de 2 horas, 2:De 2 a 5 horas, 3:De 5 a 10 horas, 4:Mais de 10 horas
failures	Número de disciplinas com reprovação	Qualitativo	Ordinal	1:Uma reprovação, 2:Duas reprovações, 3:Três reprovações, 4:Quatro ou mais reprovações
schoolsup	Suporte educacional extra	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
famsup	Suporte educacional familiar	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
paid	Aulas pagas extra da disciplina analisada	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
activities	Atividades extra-curríclares	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
nursery	Frequentou aulas de enfermagem?	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
internet	Possui acesso à internet em casa?	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
romantic	Em uma relação amorosa?	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
higher	Pretende cursar educação superior?	Qualitativo	Nominal	{"*yes*":Sim, "*no*":Não}
famrel	Qualidade da relação com a família	Qualitativo	Ordinal	[1:"*Muito ruim*"até 5:"*Excelente*"]
freetime	Tempo livre após a escola	Qualitativo	Ordinal	[1:"*Pouco*"até 5:"*Muito*"]
goout	Saídas com os amigos	Qualitativo	Ordinal	[1:"*Pouco*"até 5:"*Muito*"]
Dalc	Consumo diário de álcool	Qualitativo	Ordinal	[1:"*Pouco*"até 5:"*Muito*"]
Walc	Consumo semanal de álcool	Qualitativo	Ordinal	[1:"*Pouco*"até 5:"*Muito*"]
health	Condição de saúde atual	Qualitativo	Ordinal	[1:"*Muito ruim*"até 5:"*Muito boa*"]
absences	Número de faltas escolares (adimensional)	Quantitativo Discreto	Racional	[0-93]
G1	Nota do primeiro período (adimensional)	Quantitativo Discreto	Racional	[0-20]
G2	Nota do segundo período (adimensional)	Quantitativo Discreto	Racional	[0-20]
G3	Nota final (adimensional)	Quantitativo Discreto	Racional	[0-20]

Apêndice B

Hiperparâmetros utilizados em cada pasta para testar a eficiência dos algoritmos para a tarefa de classificação.

Algoritmo	**Hiperparâmetro**	**Valores**
MLP	Função de ativação	Identidade Logística Tangente Hiperbólica Linear retificada
	Otimização de pesos	L-BFGS Adam [14]
SVM	Regularização (C)	1 10
	Kernel	Linear RBF
KNN	Número de vizinhos	1 3 5
DT	Critério de divisão	Impureza de Gini Entropia
	Profundidade máxima	5 10 20
RF	Critério de divisão	Impureza de Gini Entropia
	Profundidade máxima	5 10 20
	Número de estimadores	30 50 100

Apêndice C

Hiperparâmetros utilizados em cada pasta para testar a eficiência dos algoritmos para a tarefa de regressão.

Algoritmo	**Hiperparâmetro**	**Valores**
MLP	Função de ativação	Identidade Tangente Hiperbólica Linear retificada
	Otimização de pesos	L-BFGS Adam [14]
SVM	*Kernel*	Linear RBF Polinomial Sigmoidal
KNN	Número de vizinhos	1 3 5
DT	Critério de divisão	Erro Quadrático Médio (EQM) EQM corrigido por Friedman Erro Absoluto Médio (EAM)
	Profundidade máxima	5 10 20
RF	Critério de divisão	Erro Quadrático Médio (EQM) Erro Absoluto Médio (EAM)
	Profundidade máxima	5 10 20
	Número de estimadores	30 50 100