PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Instituto de Ciências Exatas e de Informática

Aprendizado de Máquina - Trabalho Sistemas Inteligentes*

Machine Learning

Resumo

O trabalho foi desenvolvido com a finalidade de aprofundar os tópicos vistos na matéria de Sistemas Inteligentes, da PUC Minas. Ao utilizar técnicas de inteligência artificial, por meio do software Orange Mining, foi possível colocar em prática tudo visto anteriormente. Dessa forma, utilizamos essas tecnologias ao nosso favor para aprofundar questões da área da educação.

Palavras-chave: Inteligência artificial. Educação.

*Artigo apresentado como pré-requisito para conclusão da matéria Sistemas Inteligentes do curso de Ciência da Computação.

1 IDENTIFICAÇÃO DE UMA BASE DE DADOS DE INTERESSE

Em primeiro lugar, é notório que a educação é essencial para formação do cidadão e transformação da sociedade. Ela é responsável é a responsável pela multiplicação do conhecimento e pelo desenvolvimento de habilidades úteis para a atuação do indivíduo em sua comunidade. Dessa forma, levando em consideração a importância desse tema, foi escolhida um banco de dados relacionada a esse tema.

A base de dados escolhida aborda o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos de dados incluem notas dos alunos, características demográficas, sociais e relacionadas à escola e foram coletados por meio de relatórios e questionários escolares. São fornecidos dois conjuntos de dados relativos ao desempenho em duas disciplinas distintas: Matemática (mat) e Língua Portuguesa (por), entretanto, visando uma possível simplificação do estudo, apenas a primeira foi considerada para o treinamento, sendo que a segunda foi usada novo para testes dos modelos de classificação em uma base de dados nova, tudo isso com auxilio do software Orange Mining.

2 PREPARAÇÃO/PRÉ-PROCESSAMENTO DOS DADOS

É importante lembrar que existe uma fase intermediaria entre a identificação dos dados e a classificação do mesmo, denominada de preparação ou pré-processamento dos dados. Essa etapa é crucial para obter resultados de análise de melhor qualidade e veracidade. Para uma melhor analise nesse caso, foram realizadas etapas para o processamento dos dados, por meio do Orange Mining. Antes de tudo, foi realizado a limpeza dos dados, removendo registros com atributos nulos. Além disso, visando o aumento da eficiência e diminuição dos custos foi feito a redução em cinquenta por cento dos registros de toda base de dados, sendo que o restante foi utilizado apenas para testes de qualidade e confiança dos modelos criados. Por fim, foi efetuada uma seleção das colunas e do objetivo que seria utilizado para classificação, assim foi escolhido as seguintes colunas, por serem consideradas as mais relacionadas com o alvo: Studytime, Famrel, Romantic, Fedu, Medu, Traveltime, Freetime e Absences para o objeto de classificar se o aluno foi ou não aprovado, através da coluna “Aprovado”.

2.1 Dicionário de Dados

Dicionário dos Dados

Columns	Description	Tipo de Dado	Intervalo
school	escola do estudante	binário	'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira
sex	sexo do aluno	binário	'F' - feminino ou 'M' - masculino
age	idade do aluno	numérico	de 15 a 22
address	tipo de endereço residencial do aluno	binário	'U' - urbano ou 'R' - rural
famsize	tamanho família	binário	'LE3' - menor ou igual a 3 ou 'GT3' - maior que 3
Pstatus	estado de coabitação dos pais	binário	'T' - vivendo juntos ou 'A' - separados
Medu	educação da mãe	numérico	0 - nenhum a 4 ensino superior
Fedu	educação do pai	numérico	0 - nenhum a 4 ensino superior
Mjob	trabalho de mãe	nominal	'professor', 'cuidados de saúde', 'serviços' civis
Fjob	Trabalho do pai	nominal	'professor', 'cuidados de saúde', 'serviços' civis
reason	razão para escolher esta escola	nominal	próximo a 'casa', escola 'reputação', preferência 'curso' ou 'outro'
guardian	guardião do aluno	nominal	'mãe', 'pai' ou 'outro'
traveltime	tempo de viagem de casa para escola	numérico	1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - > 1 hora
studytime	tempo de estudo semanal	numérico	1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 - >10 horas
failures	número de reprovações em aulas anteriores	numérico	n se 1<=n<3, senão 4
schoolsup	apoio educacional extra	binário	sim ou não
famsup	apoio educacional familiar	binário	sim ou não
paid	aulas extras pagas dentro da disciplina do curso	binário	sim ou não
activities	atividades extracurriculares	binário	sim ou não
nursery	frequentou a creche	binário	sim ou não
higher	quer fazer o ensino superior	binário	sim ou não
internet	acesso à internet em casa	binário	sim ou não
romantic	com um relacionamento amoroso	binário	sim ou não
famrel	qualidade das relações familiares	numérico	de 1 - muito ruim a 5 - excelente
freetime	tempo livre depois da escola	numérico	de 1 - muito baixo a 5 - muito alto
goout	sair com amigos	numérico	de 1 - muito baixo a 5 - muito alto
Dalc	consumo de álcool no dia de trabalho	numérico	de 1 - muito baixo a 5 - muito alto
Walc	consumo de álcool no fim de semana	numérico	de 1 - muito baixo a 5 - muito alto
health	estado de saúde atual	numérico	de 1 - muito ruim a 5 - muito bom
absences	número de faltas escolares	numérico	de 0 a 93
G1	nota do 1 período	numérico	de 0 a 20
G2	nota do 2 período	numérico	de 0 a 20
G3	nota final	numérico	de 0 a 20
Aprovado	aprovado ou não	Binário	falso ou verdadeiro

2.2 Amostra de Dados

aluno	1	2	3	4	5
school	GP	GP	GP	GP	GP
sex	F	F	F	F	F
age	18	17	15	15	16
address	U	U	U	U	U
famsize	GT3	GT3	LE3	GT3	GT3
Pstatus	A	T	T	T	T
Medu	4	1	1	4	3
Fedu	4	1	1	2	3
Mjob	at_home	at_home	at_home	health	other
Fjob	teacher	other	other	services	other
reason	course	course	other	home	home
guardian	mother	father	mother	mother	father
traveltime	2	1	1	1	1
studytime	2	2	2	3	2
failures	0	0	3	0	0
schoolsup	yes	no	yes	no	no
famsup	no	yes	no	yes	yes
paid	no	no	yes	yes	yes
activities	no	no	no	yes	no
nursery	yes	no	yes	yes	yes
higher	yes	yes	yes	yes	yes
internet	no	yes	yes	yes	no
romantic	no	no	no	yes	no
famrel	4	5	4	3	4
freetime	3	3	3	2	3
goout	4	3	2	2	2
Dalc	1	1	2	1	1
Walc	1	1	3	1	2
health	3	3	3	5	5
absences	6	4	10	2	4
G1	5	5	7	15	6
G2	6	5	8	14	10
G3	6	6	10	15	10
Aprovado	FALSO	FALSO	FALSO	VERDADEIRO	FALSO

3 TAREFA DE APRENDIZADO DE MÁQUINA: CLASSIFICAÇÃO

Como dito anteriormente, para tal tarefa, foi utilizado o Orange, que é um kit de ferramentas de visualização de dados de código aberto, aprendizado de máquina e mineração de dados. Dessa forma, após conclusão foi disponibilizado o fluxograma da ferramenta.

3.1 Árvore de Decisão

A Árvore de Decisão é uma técnica de aprendizado supervisionado que pode ser usada tanto para problemas de classificação quanto de regressão, mas é preferida principalmente para resolver problemas de classificação. É um classificador estruturado em árvore, onde os nós internos representam as características de um conjunto de dados, os ramos representam as regras de decisão e cada nó folha representa o resultado.

3.1.1 Definição

A Árvore de Decisão é uma técnica de aprendizado supervisionado que pode ser usada tanto para problemas de classificação quanto de regressão, mas é preferida principalmente para resolver problemas de classificação. É um classificador estruturado em árvore, onde os nós internos representam as características de um conjunto de dados, os ramos representam as regras de decisão e cada nó folha representa o resultado.

3.1.2 Modelo Criado

O modelo criado para esse trabalho levou em consideração cinquenta por cento da base de dados para treinamento, ou seja, 199 registros e o restante para validação e testes. Além disso, é notório revelar que todas as opões de parâmetros foram deixadas do mesmo jeito que a plataforma recomenda, ou seja, houve uma indução para criação de uma árvore binária, tive um mínimo de duas instâncias nas folhas, conjuntos menores que cinco não foram divididos, e por fim a profundidade máxima da árvore foi definida para cem.

3.1.3 Precisão do Modelo

É significante analisar e compreender as métricas do modela, para realiza-se uma avaliação do mesmo. No caso, dessa arvore de decisão nos dados treinados previamente foi obtido uma acurácia de 87,9%, uma precisão de 91,2% e uma revocação de 77,5%. Entretanto, nos

dados considerados como novos, ou seja , de uma disciplina diferente, em uma tentativa de prever os aprovados, esse mesmo modelo teve uma acurácia de 58,3 %, uma precisão de 61,2% e uma revocação de 58,3%.

3.2 Rede Neural

3.2.1 Definição

Uma rede neural é uma série de algoritmos que se esforça para reconhecer relacionamentos subjacentes em um conjunto de dados por meio de um processo que imita a maneira como o cérebro humano opera. Nesse sentido, as redes neurais referem-se a sistemas de neurônios, sejam de natureza orgânica ou artificial.

3.2.2 Modelo Criado

O modelo criado para esse trabalho levou em consideração 50% da base de dados para treinamento, ou seja, 199 registros e o restante para validação e testes. Além disso, é notório revelar que todas as opões de parâmetros foram alteradas, ou seja, houve a utilização de 1500 neurônios nas camadas da rede, foi utilizada a ativação denominada de ReLu e o solucionador Adam, a regularização foi definida como zero, e finalmente, o número de passos máximos do modelo foi ajustada para 2000.

3.2.3 Precisão do Modelo

Neste caso, foi obtida métricas que indicam uma maior eficiência do que o outro modelo. Em um conjunto de dados que foi utilizado para o treinamento, foi alcançado uma acurácia de 99,5%, uma precisão de 91,2% e uma revocação de 100%. Porém, é notório que existe uma grande diferença quando analisado com dados novos, assim conquistado uma acurácia de 58%, uma precisão de 60,2% e uma revocação de 58%.

4 ANÁLISE CRÍTICA E CONCLUSÃO

Em primeiro lugar, é importante ressaltar a diferença das métricas analisadas entre os modelos citados anteriormente, quando analise os dados da disciplina de matemática que foram usados no treinamento. Em relação a acurácia da classificação, precisão e revocação possui uma distinção de 11,6%, 7,6% e 22,5%, respectivamente, sendo que a rede neural teve melhores

resultados em todas métricas. Todavia, quando essa análise é feita em uma base de dados nova, no caso da disciplina de português, essa diferença modifica-se a 0,03%, 1,1%,0,03%, respectivamente, sendo que a arvore de decisão teve melhores resultados em todas métrica, diferentemente do outro cenário. Portanto, nota-se que o modelo de rede neural foi muito mais eficiente quando analisado com uma base de dados previamente treinada, e um pouco menos eficiente com dados novos, porém devido a grande diferença no primeiro cenário, conclui-se que este foi o melhor modelo de classificação considerando o contexto e o desempenho.

REFERÊNCIAS

CHAUHAN, Aman. **Student performance**. 2022. Disponível em: <https://www.kaggle.com/datasets/whenamancodes/student-performance>.

(CHAUHAN, 2022)