TRABALHO FINAL - TRATAMENTO DE INCERTEZAS

A PREPRINT

Camila Gusmão **Marcos F. M. de Souza**
Instituto de Computação - IC
Universidade Federal Fluminense - UFF
Niterói, RJ
camilagusmao@id.uff.br
marcos.souza@iff.edu.br

February 12, 2022

ABSTRACT

This work aims to make Classification and Projection/Prediction experiments with the Student Perfomance dataset, available on UCI (University of California, Irvine) website. The experiments aim to find out the most relevant factors for good school performance in Portuguese and Math classes, and to find the best classifier based on the gathered parameters as well. Throughout the work the base concepts will be grounded, the used technics will be presented, just like the pre-processing and the performance evaluation metrics. Next, the classifiers execution proccess is presented. At last, each classifier metric is shown.

Este trabalho tem como objetivo realizar experimentos de Classificação e Projeção/Predição envolvendo a base de dados *Student Performance*, disponível no site da **UCI** (Universidade da Califórnia, Irvine). Os experimentos visam a obtenção de métricas de desempenho escolar de alunos em disciplinas de Português e Matemática. No decorrer do trabalho os conceitos base são fundamentados, sendo apresentadas as técnicas utilizadas, o pré-processamento adotado e as métricas de desempenho. Na sequência, é apresentado o processo de execução dos classificadores, as métricas dos classificadores escolhidos bem como as métricas de regressão linear para a projeção de notas dos alunos.

Keywords Classificação · Análise de Dados · Educação · Regressão Linear · Desempenho

1 Introdução

A base utilizada neste trabalho é a *Student Performance*, que está disponível no site da UCI[1] e contém informações sobre o aproveitamento dos alunos do ensino médio de duas escolas públicas de Portugal nas disciplinas de português e matemática. Os atributos incluem notas dos alunos, características demográficas, sociais e relacionadas à escola, sendo coletados por meio de relatórios escolares e questionários [Cortez et al. (2008)]. São fornecidos dois conjuntos de dados relativos ao desempenho em cada disciplina: O primeiro conjunto possui 649 registros com dados dos alunos da disciplina de língua portuguesa e o segundo conjunto de dados possui 395 registros contendo os dados dos alunos da disciplina de matemática. Estas informações foram coletadas entre 2005 e 2006 e os resultados deste estudo foram publicados em 2008 [Cortez et al. (2008)]. Vale destacar que ambos os conjuntos de dados possuem os mesmos atributos, sendo 33 atributos incluindo o atributo alvo, como pode ser visto na Figura 1.

[1] https://archive.ics.uci.edu/ml/datasets/Student+Performance

Atributo	Descrição	Valores Possíveis	Descrição dos Valores Possíveis
school	escola do aluno	[”GP”, ”MS”]	[”GP”: Gabriel Pereira, ”MS”: Mousinho da Silveira]
age	idade do aluno		
absences	número de ausências	Número discreto de 0 a 93	
sex	sexo do aluno	[”F”, ”M”]	[”F”: feminino, ”M”: masculino]
address	tipo de endereço residencial do aluno	[”U”, ”R”]	[”U”: meio urbano, ”R”: meio rural]
famsize	tamanho da família do aluno	[”LE3”, ”GT3”]	[”LE3- Menor ou igual a 3 ou ”GT3- Maior que 3]
Pstatus	situação de coabitação dos pais	[”T”, ”A”]	[”T”: mora junto dos pais, ”A”: mora separado dos pais]
failures	Número de reprovações	Número discreto n de 0 a 4 onde n varia de 1 a 3 caso n esteja entre 0 e 3 ou caso 4 caso seja superior a 3	
Medu	grau de instrução da mãe	Número discreto de 0 a 4	0 - Nenhuma, 1 - Educação Primária (4^o ano), $2 - 5^o$ a 9^o ano, 3 – Ensino Médio ou 4 – Ensino Superior)
Fedu	grau de instrução do pai	Número discreto de 0 a 4	0 - Nenhuma, 1 - Educação Primária (4^o ano), $2 - 5^o$ a 9^o ano, 3 – Ensino Médio ou 4 – Ensino Superior)
Mjob	tipo de ocupação da mãe	[Professor(a), Saúde , Serviços Sociais , Em Casa ou Outro]	Professor(a), Cuidado com Saúde, Serviços Sociais (e.g. Adminstrativo ou Polícia), Em Casa ou Outro
Fjob	tipo de ocupação do pai	[Professor(a), Saúde , Serviços Sociais , Em Casa ou Outro]	Professor(a), Cuidado com Saúde, Serviços Sociais (e.g. Adminstrativo ou Polícia), Em Casa ou Outro
reason	razão da escolha da escola	[”home”, ”reputation”, ”course”, ”other”]	[”home”: perto de casa, ”reputation”: reputação da escola, ”course”: preferência de curso, ”other”: outro motivo]
guardian	responsável pelo aluno	[”mãe”, ”pai”ou ”outro”]	[”mãe”, ”pai”ou ”outro”]
traveltime	Tempo de viagem da residência até a escola	[1, 2, 3, 4]	Tempo de Viagem entre Casa e Escola (numérico: 1 - $<$ 15 min., 2 - 15 até 30 min., 3 - 30 min. até 1 hora, ou 4 - $>$1 hora
studytime	Tempo de estudo semanal	[1,2,3,4]	Tempo de Estudo Semanal (numérico: 1 - $<$2 horas, 2 - 2 até 5 horas, 3 - 5 até 10 horas, ou 4 - $>$10 horas)
schoolsup	Suporte educacional extra	[ye,no]	[Suporte Educacional Extra (binário: sim ou não)]
famsup	Suporte educacional familar	[yes,no]	[Suporte Educacional Familiar (binário: sim ou não)]
paid	Aulas extras pagas	[yes,no]	[Aulas Extras Pagas dentro do assunto do curso (Matemática ou Português) (binário: sim ou não)]
activities	Atividades extra-curriculares	[yes,no]	[Atividades Extra Curriculares (binária: sim ou não)]
nursery		[yes,no]	[Frequentou a Creche (binário: sim ou não)]
higher	Quer ter educação superior		[Quer ter educação superior (binário: sim ou não)]
internet	Dispõe de acesso à internet em casa	[yes,no]	[Dispõe de acesso à internet em casa (binário: sim ou não)]
romantic	Está em um relacionamento romântico	[yes,no]	[Está em relacionamento romântico (binário: sim ou não)]
famrel	Qualidade do relacionamento familiar	[1,2,3,4,5]	[Qualidade do relacionamento familiar (numérico: de 1 - muito ruim até 5 - excelente)]
freetime	Tempo livre após a escola	[1,2,3,4,5]	[Tempo livre após a escola (numérico: de 1 - pequeno até 5 - grande)]
goout	Tempo com os amigos fora da escola	[1,2,3,4,5]	[Tempo com os amigos fora da escola (numérico: de 1 - muito baixo até 5 - muito alto)]
Dalc	Consumo de álcool em dias de semana	[1,2,3,4,5]	[Consumo de álcool em dias da semana (numérico: de 1 - muito baixo até 5 - muito alto)]
Walc	Consumo de álcool em fins de semana	[1,2,3,4,5]	[Consumo de álcool em finais de semana (numérico: de 1 - muito baixo até 5 - muito alto)]
health	Estado de saúde atual	[1,2,3,4,5]	[Estado de Saúde atual (numérico: de 1 - muito ruim até 5 - muito bom)]
absences	Nº de faltas na escola	[0-93]	[Nº de faltas na escola (numérico: de 0 até 93)]
G1	Nota no primeiro período	[0-20]	[Nota no primeiro período (numérico: de 0 até 20)]
G2	Nota no segundo período	[0-20]	[Nota no segundo período (numérico: de 0 até 20)]
G3	Nota final	[0-20]	[Nota final (numérico: de 0 até 20, saída desejada)]

Figure 1: **Atributos presentes no Dataset. Fonte: Autores**

1.1 Objetivos

Neste trabalho, o intuito é criar modelos preditivos capazes de prever se um aluno será ou não aprovado, levando em consideração todas as informações existentes sobre ele, como idade, grau de instrução dos pais, o local onde mora, dentre outras características.

De forma mais específica, tem-se como objetivo a construção de três modelos, nos quais os dois primeiros são classificadores; um para avaliar se um aluno será aprovado ou não e o outro para avaliar qual será o conceito final atingido por este aluno. Além destes classificadores, o terceiro modelo é um modelo de previsão da nota final do aluno, por meio de técnicas de regressão.

Estes três objetivos estão intrinsicamente ligados ao atributo G3 que contém a nota final de cada aluno, sendo um atributo quantitativo discreto, variando de 0 a 20, dado que o sistema educacional desta localidade adota este intervalo para avaliar os alunos. Para a construção dos dois classificadores, foram criados atributos auxiliares, onde um indica a aprovação, que é quando o aluno possui notal final igual ou superior a 10, e o outro indica em qual conceito aquela nota se encaixa [Cortez et al. (2008)]. Para a criação dos conceitos, as notas de 0 a 20 foram convertidas em 5 intervalos, aqui representando 5 classes de nota dos alunos, como pode-se observar na Tabela 1.

Intervalo da nota final	Conceito
$G3 \in [0, 10)$	F
$G3 \in [10, 12)$	D
$G3 \in [12, 14)$	C
$G3 \in [14, 16)$	B
$G3 \in [16, 20]$	A

Table 1: **Conceito do aluno de acordo com a nota final**

Efetuada esta conversão, temos a distribuição de notas nas disciplinas de Matemática e Português, presente nas Figuras 2 e 3.

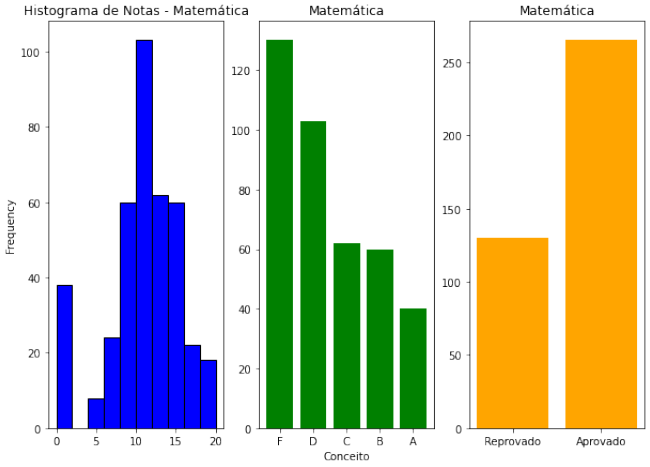

Figure 2: **Histograma de Atributos para a Base de Matemática. Fonte: Autores**

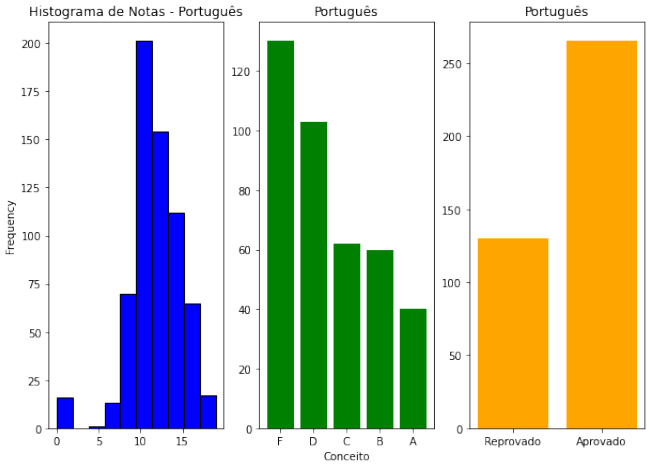

Figure 3: **Histograma de Atributos para a Base de Português. Fonte: Autores**

Na seção seguinte serão apresentados os modelos usados para efetuar as predições sobre as notas dos alunos deste estudo.

2 Referencial Teórico e Revisão da Literatura

Nesta seção são apresentados de forma breve os conceitos fundamentais sobre Aprendizado de Máquina (AM) bem como os algoritmos utilizados para a elaboração deste trabalho e uma revisão da literatura utilizada.

2.1 Revisão da Literatura

O artigo que deu origem à base Student Performance [Cortez et al. (2008)], utilizada neste trabalho, já foi citado por mais de 500 trabalhos[2], mostrando que este trabalho foi relevante para a comunidade científica, pois desencadeou novas pesquisas no âmbito educacional utilizando o aprendizado de máquina como ferramenta. Embora ele seja de 2008, temos diversos trabalhos recentes que o referenciam. Alguns trabalhos recentes no âmbito educacional que podemos destacar são:

[Kaura et al. 2015] que apresenta uma análise utilizando ferramentas como o *Weka* e métodos como Multilayer Perception, Naive Bayes e SMO para construir 5 classificadores para identificar estudantes com lentidão no aprendizado em uma escola de ensino médio.

Outro artigo relevante é [Injadat et al. 2020] que utiliza técnicas gráficas, estatísticas e quantitativas, além de abordagens utilizando índice Gini e *p-value* para estimar parâmetros e escolher o melhor algoritmo de aprendizado para analisar bases de dados de desempenho escolar.

2.2 Aprendizado de Máquina

O Aprendizado de Máquina (AM) ou *Machine Learning* em inglês, é um subcampo da Inteligência Artificial (IA) responsável por construir modelos que, submetidos a um grande volume de dados, são capazes de aprender, tomar decisões e identificar padrões com o mínimo de interferência humana. A construção destes modelos requer a implementação de algoritmos que processam um alto volume de dados de modo que tal quantidade não seja um embargo para se obter respostas rápidas e precisas [SAP 2020]. A seguir são apresentados os principais conceitos de AM e os tipos de aprendizado.

2.2.1 Principais conceitos

Para que um modelo de aprendizado de máquina seja construído, é necessário antes de mais nada **coletar** e **preparar** os dados que serão utilizados pelo modelo, através de coleta e pré-processamento, procedimentos que serão apresentados mais adiante. Os **dados** são representados como vetores, onde cada dimensão desse vetor é chamada de **atributo**. Um conjunto de valores para esses atributos é uma **instância**.

Os dados selecionados podem ser divididos em dois conjuntos. O primeiro é o **conjunto de treinamento**, sendo utilizado na construção do modelo, para que os dados sejam usados pelo algoritmo na fase de aprendizagem. Nesta fase, chamada de fase de treinamento, o modelo é investigado, a partir das parametrizações que cada modelo dispõe, e, em geral, o modelo que tiver maior métrica é escolhido para ser usado na previsão de classificações futuras. Para medir o desempenho do modelo construído, é utilizado o segundo conjunto, que pode ser chamado de **conjunto de teste ou validação**, desconhecido até então pelo modelo. Para a escolha do modelo, deve-se observar o problema a ser resolvido bem como o tipo de dado existente, de modo a escolher um modelo que melhor se adeque a estas condições. Um ponto a destacar é que o sucesso do aprendizado do modelo requer atenção na escolha dos parâmetros usados no treinamento, ou seja, muitas vezes ótimos resultados nessa fase podem estar relacionados a ***overfitting***, isto é, o modelo se especializou nos dados usados no treinamento, e com isso não é capaz de generalizar adequadamente ao avaliar novos dados. Por outro lado, não ter resultados satisfatórios nas predições de treino pode indicar o ***underfitting***, que mostra que o modelo não está conseguindo aprender e, portanto, não é capaz de fazer inferências sobre os dados.

Um fator importante na escolha de um modelo é a sua capacidade de generalização, para garantir que classificações futuras, sobre dados desconhecidos, sejam o mais acertadas possível. A partir dos ajustes feitos nos modelos durante a fase de treinamento, um modelo pode se adaptar excessivamente àquele conjunto, gerando uma falsa sensação de êxito. Para analisar o comportamento do modelo de forma mais robusta, pode ser aplicada a técnica de **validação cruzada**. Na validação cruzada (ou *K-fold Cross Validation*), a base de dados é dividida em k conjuntos de dados. A cada iteração, um modelo é desenvolvido utilizando o conjunto de dados total - k, e o conjunto k é usado como validação [Cesar et al. 2017]. Ao final do processo é selecionado o modelo que possui o melhor desempenho médio (média das k execuções).

[2]Resultado obtido após busca por [Cortez et al. (2008)] no *Google Scholar*

2.3 Algoritmos de classificação utilizados

Nesta seção serão apresentados os três algoritmos utilizados para geração dos modelos classificadores deste trabalho.

2.3.1 Árvores de Decisão

As Árvores de Decisão (AD) configuram um método de aprendizado supervisionado caracterizado pelo uso da estrutura de árvore para representar possíveis decisões que podem ser feitas sobre um conjunto de dados (Ver Figura 3). Em outras palavras, As AD classificam instâncias através de uma divisão baseada nos valores dos atributos [Kotsiantis, 2007]. Nas árvores existem os nós internos, que possuem filhos e os nós presentes no último nível da árvores, chamados de **folhas**. Os nós internos representam os **atributos** de uma base de dados, as arestas (ou ramos) são os **predicados**, que são os valores (ou um intervalo de valores) que um determinado atributo pode assumir. As folhas das árvores são os valores das **classes** [Han et al. 2013].

Este modelo de aprendizado é **orientado a regras**, de modo que um nó interno com um predicado determina uma condição sobre o conjunto de dados. As instâncias são classificadas a partir da raiz da árvore, que é o primeiro nó – e divididas pelos valores dos seus atributos até chegar à folha [Kotsiantis, 2007]. Cada caminho da raíz até a folha representa uma regra, definida como a conjunção das condições percorridas, implicando no valor da classe encontrada na folha em questão. A árvore deve ser definida de forma que, para um mesmo registro, haja um e apenas um caminho da raiz até a folha. Dentre os algoritmos de AD, os mais conhecidos são ID3, CART e o C4.5, no qual cada um utiliza critérios distintos para separar as instâncias recebidas entre os atributos como o ganho de informação através da entropia e do método Gini [Han et al. 2013].

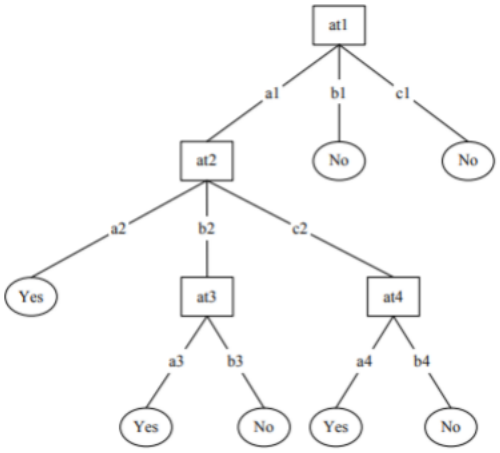

Figure 4: **Estrutura de uma Árvore de Decisão. Fonte: [Kotsiantis 2007]**

2.3.2 SVM

O algoritmo de Máquinas de Vetores de Suporte (SVM), ou *Support Vector Machines* em inglês, é um algoritmo de aprendizado supervisionado que foi desenvolvido por [CORTES & VAPNIK] em 1995 e, essencialmente, consiste em criar classificadores lineares capazes de separar conjuntos de dados através de um hiperplano [Vapnik et al. 1995]. Embora o SVM tenha sido criado originalmente para classificações binárias, ele também pode ser usado em conjuntos de dados com múltiplas classes.

O objetivo é encontrar um hiperplano que gere a melhor separação possível entre as classes observadas no conjunto de treinamento, onde cada registro será representado por um ponto no espaço n-dimensional. Quando se tem um espaço com duas dimensões, a função do hiperplano é a equação da reta. No entanto, infinitas retas podem ser traçadas de forma que os dois conjuntos fiquem separados. Desta forma, o SVM considera os dados e parâmetros para definir o hiperplano separador, por meio de **vetores de suporte** e a **margem** ótima associada ao hiperplano [Goldschmidt et al. 2015].

Para um exemplo bidimensional, como pode-se ter inúmeras retas para separar o conjunto de dados, e para cada uma delas são traçadas duas retas paralelas, na qual cada uma se move a partir da reta candidata (separadora) até encontrar o primeiro ponto (registro da base de dados) de interseção. A distância entre as retas paralelas indica a margem entre

elas e os pontos interceptados pelas retas paralelas são os vetores de suporte do classificador, como pode ser visto no exemplo em 2 dimensões na Figura 5. O objetivo é encontrar um classificador com a maior margem possível, através da otimização de uma função Lagrangeana escrita em função dos parâmetros do hiperplano separador.

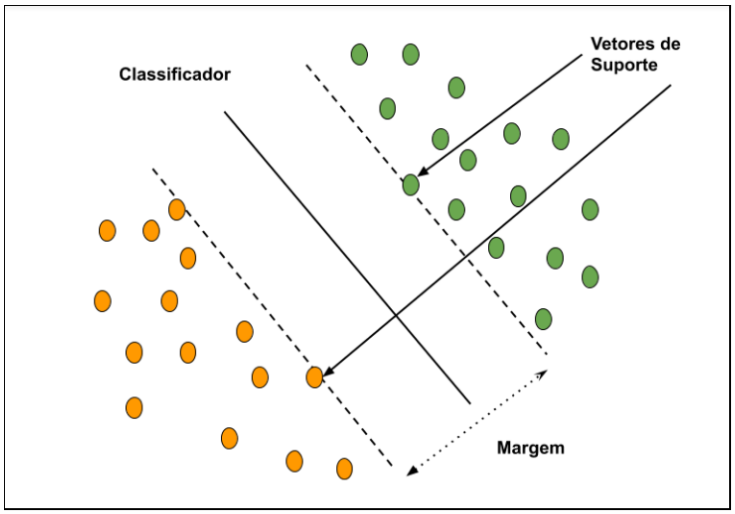

Figure 5: **Divisão dos dados pelo hiperplano que define o classificador, onde os vetores de suporte são os 3 pontos que interceptam as retas paralelas tracejadas. Fonte: Autores**

Além disso, pode-se considerar a inclusão de variáveis de folga que permitem o relaxamento das restrições, restringindo a existência de dados de treinamento entre as margens que separam as classes [CARVALHO et al. 2007]. Esse modelo, denominado classificador SVM com margens suaves, permite que alguns dados possam violar a restrição de separação entre as classes, conforme mostra a Figura 6. As variáveis de folga indicam a distância de um padrão em relação à margem correspondente a sua classe.

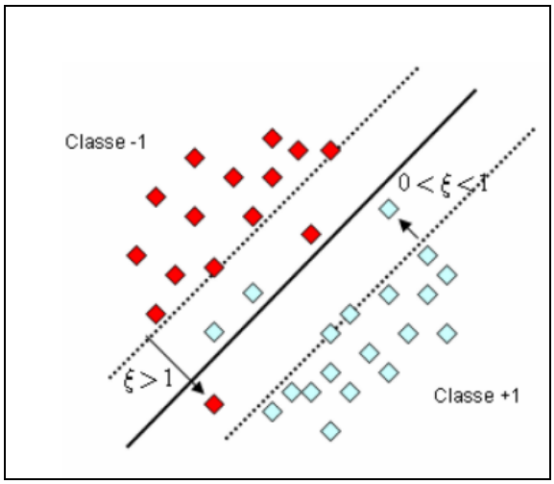

Figure 6: **Inclusão de variáveis de folga, $\xi_i \geq 0$ visando a construção de modelos com margens suaves. Fonte: Autores.**

Na Figura 6 pode-se observar que um dos elementos vermelhos está além da margem separadora, mas também há casos em que há elementos entre o hiperplano e a margem. Este procedimento possibilita aceitar padrões que se situam fora da região de sua classe, impedindo que esses desviem o hiperplano separador. A soma das variáveis de folga ξ_i representa um limite no número de erros de treinamento. Dessa forma, surge o parâmetro de custo "C", que pondera a importância, no processo de minimização dada a esses elementos que violam a classificação, possibilitando que o modelo do SVM seja menos sensível à presença de pontos "mal comportados" no conjunto de treinamento. Valores do parâmetro C com valores altos, penalizará mais uma classificação incorreta, reduzindo a quantidade de erros (minimizando mais a

margem), e com valores baixos maximiza a margem, de modo que o hiperplano é menos sensível a erros do conjunto de aprendizado. No entanto, isso poderá causar *overfitting* ao modelo, reduzindo sua capacidade de generalização.

Embora o SVM tenha sido criado originalmente para conjuntos de dados linearmente separáveis, ele também pode ser utilizado no treinamento de dados não linearmente separáveis. Conjuntos não linearmente separáveis são conjuntos em que não existe um hiperplano capaz de separar completamente os registros de classes distintas. Para conjuntos não linearmente separáveis, ou seja, com uma alta quantidade de registros ruidosos, são utilizadas as **funções núcleo**, conhecidas também como *kernel functions* [Smola et al. 2002]. As funções núcleo mais utilizadas são apresentadas na Tabela 2.

Função Núcleo	Fórmula
Linear	$x_i \cdot x_i$
Polinomial	$(\delta(x_i \cdot x_i) + \kappa)^d$
Radial Basis Function (RBF)	$exp\left(-\sigma\|\|x_i - x_j\|\|^2\right)$
Sigmoidal	$tanh(\delta(x_i \cdot x_i) + \kappa)$

Table 2: **Funções de núcleo mais utilizadas no SVM. Fonte: Adaptado de [Smola and Scholkopf 2002]**.

2.3.3 MLP

As Redes Neurais Artificiais (RNAs) são inspiradas nos neurônios biológicos e consistem em modelos matemáticos cuja estrutura costuma ser representada como um grafo, no qual os nós representam os neurônios e as arestas as conexões entre eles [Haykin, 2001] . Um neurônio pode possuir várias entradas, e para cada entrada X de um neurônio, um peso sináptico w é aplicado, sendo que os valores de todas as entradas são somados pela função soma Σ, e então uma função de ativação φ é aplicada ao resultado, o que gera uma saída, como pode ser visto na Figura 7. Essa saída pode ter como destino a entrada de um próximo neurônio quanto pode ser o valor final da rede neural [Giacomel, 2016]

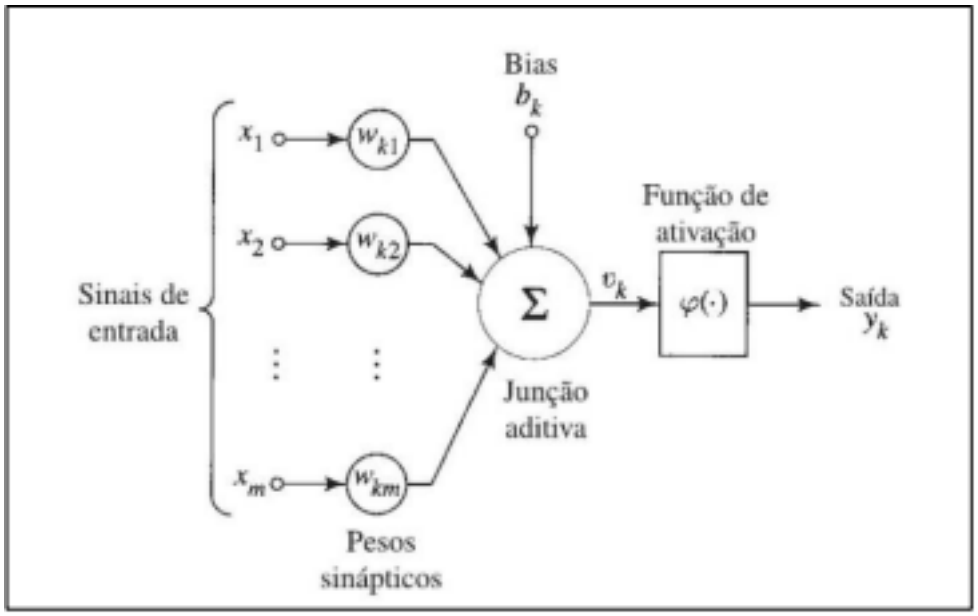

Figure 7: **Modelo de neurônio. Fonte: Adaptado de [Haykin, 2001]**

O modelo Multilayer Perceptron (MLP) é um modelo de aprendizagem supervisionada por correção de erro, possui arquitetura com mais de uma camada (também chamada de multicamadas), é acíclica (quando a saída de um neurônio não pode servir de entrada para algum neurônio anterior) e conectada (cada entrada é processada por todos os neurônios). A regra de propagação é dada pelo produto interno das entradas ponderadas pelos pesos com adição do termo *bias* e, a saída da camada anterior é a entrada da camada atual [Zavadzki, 2020].

O treinamento de um modelo MLP possui as seguintes etapas: *feedforward, backpropagation* e ajuste dos pesos. Na primeira etapa ocorrem os treinamentos de padrões com as entradas, na segunda etapa ocorre a retropropagação por

correção dos erros e por fim, na terceira etapa os pesos são ajustados [Fausett, 1994]. Estas etapas podem ser vistas na Figura 8.

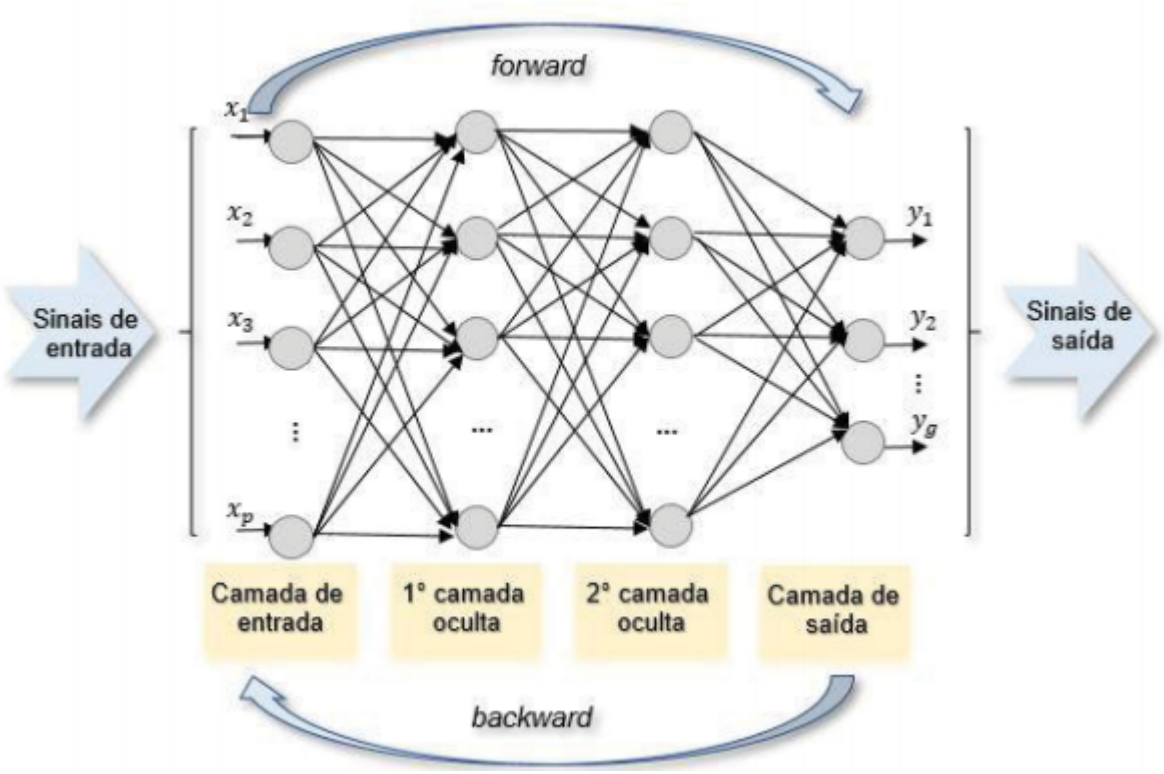

Figure 8: **Representação da rede neural MultiLayer Perceptron. Fonte: [Zavadzki, 2020]**

A seguir serão apresentados os passos de cada etapa de execução do algoritmo MLP, onde:

- *j* refere-se aos neurônios da camada de entrada,
- *k* refere-se aos neurônios da camada oculta,
- *h* refere-se aos neurônios da camada de saída,
- *t* representa o número da iteração, iniciando em 1.

Propagação Forward

1. Os pesos são iniciados com valores randômicos $|w| \leq 1$ e $|b| \leq 1$
2. São calculados os valores da função aditiva, dada pelas equações (1) e (2), para cada neurônio k da camada oculta e na sequência o valor de transferência resultante da equação (3); Nesta etapa também é definida a função de ativação que será usada pelo modelo.

$$u_k^{(t)} = \Sigma_{j=1}^{p} x_j^{(t)} w_{jk}^{(t)} \tag{1}$$

$$v_k^{(t)} = u_k^{(t)} + b_k^{(t)} \tag{2}$$

$$a_k^{(t)} = \varphi\left(v_k^{(t)}\right) \tag{3}$$

3. Para cada neurônio h da camada de saída faz-se conforme as equações (4), (5) e (6), nos casos em que há mais de um neurônio nesta camada. A função de ativação para esta etapa pode ser a mesma do passo anterior ou uma função diferente

$$u_h^{(t)} = \sum_{j=1}^{n} a_k^{(t)} w_{kh}^{(t)} \tag{4}$$

$$v_h^{(t)} = u_h^{(t)} + b_h^{(t)} \tag{5}$$

$$a_h^{(t)} = \varphi\left(v_h^{(t)}\right) \tag{6}$$

Propagação Backward

4. O erro de previsão é então calculado e um algoritmo de aprendizagem é utilizado para o ajuste dos pesos a fim de minimizar o erro
5. Os pesos são atualizados e é possível seguir para a próxima iteração **t + 1** com os novos pesos. Isso ocorre até que o critério de parada seja atendido.

2.4 Algoritmo de Regressão Linear

De acordo com [Montgomery et al. 2012], a Regressão Linear trata-se de uma técnica estatística (provavelmente, a técnica estatística mais utilizada) para investigação e modelagem da relação entre duas variáveis. Essa técnica é extremamente aplicada em muitas áreas, dentre elas, na engenharia, física e química, entre outras.

Existem diversos exemplos que podem ser utilizados para expressar a utilização da Regressão Linear. O exemplo utilizado por [Montgomery et al. 2012] aborda um problema de um engenheiro que é empregado de uma empresa de refrigerantes. O engenheiro suspeita que o tempo gasto pelo entregador está relacionado com o volume entregue. Dessa forma, o engenheiro visita aleatoriamente 25 pontos de venda e os analisa.

A Figura 9 abaixo é chamado de Diagrama de Dispersão, nela é possível verificar claramente que existe uma relação entre o tempo de entrega e o volume entregue. Isso fica claro pelo seguinte fato: os pontos do gráfico estão espalhados e, quase sempre, se distribuem em torno de uma reta (não exatamente no traçado da reta, mas bem próximo).

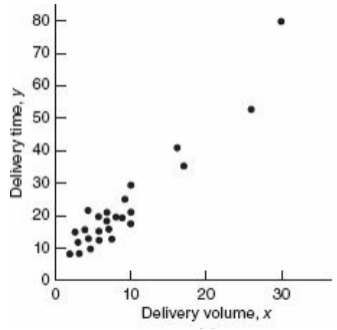

Figure 9: **Diagrama de Dispersão relacionado ao problema anterior. Fonte:[Montgomery et al. 2012]**

Na Figura 10 abaixo é possível perceber que os pontos, de fato, se encontram próximos à uma reta plotada.

Como trata-se de uma reta, podemos dizer essa reta pode ser dada através da seguinte fórmula matemática:

$$y = \beta_0 + \beta_1 x \tag{7}$$

A Equação 7 é uma similar à uma função afim, onde β_0 se comporta como o **coeficiente linear** (onde corta o eixo y) e β_1 se comporta como o **coeficiente angular** que é responsável pela inclinação da reta.

Em alguns casos (com o intuito de encontrar o melhor resultado), leva-se em consideração o erro e pode ser visto na Equação 8:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{8}$$

Esse valor de ϵ pode apresentar qualquer valor, de forma a otimizar (se aproximar da realidade) o valor da equação em relação aos dados apresentados.

Figure 10: **Diagrama de Dispersão relacionado ao problema anterior com a Reta que relaciona o tempo e volume de entrega. Fonte:[Montgomery et al. 2012]**

3 Metodologia da Análise Experimental

Conforme já foi mencionado anteriormente, neste trabalho são utilizadas duas bases de dados contendo informações sobre os alunos e seu desempenho em duas disciplinas: matemática e língua portuguesa. Como as bases não apresentam valores faltantes e possuem os mesmos atributos, todos os passos da metodologia foram adotados em ambas as bases.

3.1 Pré-processamento

Inicialmente cada base foi submetida ao pré-processamento, para que os dados pudessem ser tratados antes de serem submetidos à etapa de mineração propriamente dita. As técnicas de pré-processamento aplicadas na base podem ser consultadas a seguir:

1. Normalização dos atributos numéricos para o intervalo [0.0, 1.0]
2. Discretização dos atributos categóricos binários (que só assumem dois valores possíveis) para os valores 0 e 1
3. Discretização dos atributos categóricos para valores numéricos discretos correspondentes
4. Transformação do atributo alvo para o atributo *finalGrade* com base no campo *G3* conforme tabela 1
5. Seleção de variáveis através do método de seleção de variáveis supervisionado Relief-F executado na ferramenta Weka versão 3.8 com número de vizinhos igual a 10, removendo os 5 atributos menos relevantes para a tarefa de classificação em cada base de dados

Após o pré-processamento, cada base foi dividida por meio de *holdout* estratificado, em duas partições: 70% dos registros compuseram a base de treinamento e 30% compuseram a base de testes. Ao efetuar esta divisão da base de modo estratificado, a proporção das classes é mantida tanto no treino quanto no teste, garantindo assim um treinamento com um cenário mais próximo do real, melhorando a qualidade do treino do modelo de classificação. A seguir será apresentada a metodologia do processo de mineração propriamente dito.

3.1.1 Validação Cruzada

Com os dados pré-processados, foi aplicada a validação cruzada sobre a base de treinamento com 10 partições em conjunto com uma escolha de hiperparâmetros para cada modelo de classificação. A busca pela melhor parametrização dos modelos foi feita por meio do método do sklearn GridSearchCV, que retornava qual a melhor configuração encontrada para determinado algoritmo de classificação. Na Figura 11 encontram-se os parâmetros fornecidos para o Grid Search em cada modelo.

A partir da obtenção do melhor modelo de cada classificador, a base era treinada e o seu desempenho avaliado sobre a base de teste, cujos dados não foram apresentados ao modelo anteriormente. Para cada base de dados foram calculadas as seguintes métricas por classificador selecionado:

- *Accuracy*
- *Precision (Micro e Macro)*
- *Recall (Micro e Macro)*
- *F1 (Micro e Macro)*

Classificador	Parâmetros
Árvores de decisão	{ ”criterion”: [”gini”, ”entropy”], ”min_samples_split”: [5, 7, 9, 10, 20], ”max_depth”: [6, 7, 8], ”min_samples_leaf”: [4, 5, 6, 8, 10], ”max_leaf_nodes”: [5, 10, 20], ”max_features”: [8, 9, 10, 11, 12, 13, 14] }
SVM	[{’C’:COSTS_LIST, ’kernel’: [’linear’]}, {’C’:COSTS_LIST, ’gamma’: GAMMA_LIST, ’kernel’: [’rbf’]}, {’C’:COSTS_LIST, ’gamma’: GAMMA_LIST, ’coef0’: COEFFICIENTS, ’degree’: [1,2,3], ’kernel’: [’poly’]}, {’C’:COSTS_LIST, ’gamma’: GAMMA_LIST, ’coef0’: COEFFICIENTS, ’kernel’: [’sigmoid’]}]
MLP	{’solver’: [’lbfgs’], ’max_iter’: [35000, 40000, 45000, 50000, 55000], ’alpha’: 10.0 ** -np.arange(1, 10), ’hidden_layer_sizes’:np.arange(10, 15)}

Figure 11: **Parâmetros Grid Search por Classificador. Fonte: Autores**

4 Resultados Obtidos

Serão apresentados os resultados obtidos na fase de teste para os modelos de classificação e para o modelo de regressão.

4.1 Classificadores

Aqui apresentaremos o resultado dos classificadores criados para obter as métricas de desempenho do aluno.

4.1.1 Seleção de Variáveis no Weka

Na ferramenta Weka, utilizamos o método Relief-F para o ranqueamento de atributos, dado que são muitas colunas. Com base no resultado, foram removidos os 5 atributos de menor ranque para cada base, como podem ser vistos nas Figuras 12 e 13.

Ranque	Atributo	Ranque	Atributo
0.124064	G2	-0.001067	absences
0.081045	G1	-0.002861	romantic
0.031982	failures	-0.004114	sex
0.01224	goout	-0.004145	health
0.0121	higher	-0.004346	schoolsup
0.011009	studytime	-0.005654	Mjob
0.007367	nursery	-0.006759	guardian
0.00556	paid	-0.006887	Pstatus
0.005228	famrel	-0.008531	Dalc
0.004389	traveltime	-0.010621	activities
0.00364	school	-0.011623	Walc
0.001473	freetime	-0.013097	Fedu
0.001048	Medu	-0.019034	famsize
0.000792	age	-0.020012	famsup
0.000702	internet	-0.024621	Fjob
-0.000851	address	-0.039525	reason

Figure 12: **Ranque de Atributos para a base de Matemática. Fonte: Autores**

Ranque	Atributo	Ranque	Atributo
0.131428	G2	0.0115405	famsize
0.0973252	G1	0.0109854	romantic
0.0462798	school	0.0106637	health
0.0343094	higher	0.0098875	Mjob
0.0314783	sex	0.0095335	nursery
0.0303417	address	0.0094289	Dalc
0.023638	Walc	0.007796	freetime
0.0221811	Medu	0.0076197	reason
0.0217427	failures	0.0065532	absences
0.0193272	internet	0.0053384	age
0.0191442	famsup	0.0033358	Pstatus
0.0189613	Fedu	0.0029468	traveltime
0.0172904	studytime	0.0019692	Fjob
0.0161883	activities	0.0006745	famrel
0.0159086	schoolsup	0.0000514	guardian
0.0134514	goout	-0.0019729	paid

Figure 13: **Ranque de Atributos para a base de Português. Fonte: Autores**

4.1.2 Resultados de Teste

Como dito anteriormente, foram criados dois modelos de classificação, um para predizer se um aluno seria aprovado ou não e outro para indicar qual seria o conceito deste aluno. Para o modelo de aprovação do aluno, foram gerados os resultados de teste presentes nas Tabelas 3 e 4 e para o modelo de classificação por conceito, foram gerados os resultados das Tabelas 5 e 6. Nos testes de cada classificador foi selecionado sempre o melhor modelo com base na validação cruzada.

Métrica	Decision Tree	SVM	MLP
accuracy	0,938	0,923	0,923
precisionMicro	0,938	0,923	0,923
precisionMacro	0,868	0,873	0,853
recallMicro	0,938	0,923	0,923
recallMacro	0,889	0,791	0,821
f1Micro	0,938	0,923	0,923
f1Macro	0,878	0,824	0,836

Table 3: **Métricas de classificação dos alunos em Português quanto à aprovação ou não**

Métrica	Decision Tree	SVM	MLP
accuracy	0,890	0,882	0,873
precisionMicro	0,890	0,882	0,873
precisionMacro	0,869	0,864	0,866
recallMicro	0,890	0,882	0,873
recallMacro	0,891	0,864	0,837
f1Micro	0,890	0,882	0,873
f1Macro	0,878	0,864	0,849

Table 4: **Métricas de classificação dos alunos em Matemática quanto à aprovação ou não**

A partir dos resultados de teste exibidos nas tabelas acima, verificamos que o modelo de classificação em relação à aprovação do aluno nas disciplinas teve os melhores resultados, com destaque para a Árvore de Decisão, que alcançou uma acurácia de 93,8% para a base de Português e 89,0% para a base de Matemática, o que nos mostra que se trata de um modelo de classificação altamente confiável para obtenção da métrica de desempenho aprovação do aluno. Outras métricas foram calculadas para uma análise além da acurácia, que em bases não balanceadas pode gerar viés de interpretação, mas tanto os números de recall quanto de precisão também são muito satisfatórios.

Métrica	Decision Tree	SVM	MLP
accuracy	0,829	0,857	0,770
precisionMicro	0,829	0,857	0,770
precisionMacro	0,499	0,720	0,564
recallMicro	0,829	0,857	0,770
recallMacro	0,534	0,594	0,614
f1Micro	0,829	0,857	0,770
f1Macro	0,510	0,627	0,584

Table 5: **Métricas de classificação dos alunos em Português por conceito**

Métrica	Decision Tree	SVM	MLP
accuracy	0,720	0,758	0,742
precisionMicro	0,720	0,758	0,742
precisionMacro	0,711	0,747	0,734
recallMicro	0,720	0,758	0,742
recallMacro	0,610	0,741	0,747
f1Micro	0,720	0,758	0,742
f1Macro	0,615	0,739	0,736

Table 6: **Métricas de classificação dos alunos em Matemática por conceito**

Já para os modelos de classificação dos alunos por conceito os valores de acurácia foram um pouco mais baixos, o que é compreensível dado que ao invés de duas classes (foi aprovado ou não) existem cinco, uma para cada conceito. Com uma quantidade maior de classes o SVM foi o modelo que obteve os melhores resultados, com 85,7% de acurácia para a base de Português e 75,8% para a base de Matemática. Outro aspecto a ser observado é que em ambos os modelos de classificação os resultados foram melhores com a base de Português do que com a base de Matemática, o que pode nos indicar que provavelmente a base de Matemática necessite de um pré-processamento um pouco mais aprofundado. Outro ponto que pode justificar esta diferença esta no próprio tamanho das duas bases, visto que a base de Português possui 649 alunos e a base de Matemática possui 395 alunos, o que indica que mais dados auxiliam na aprendizagem dos modelos. Mesmo com estas ressalvas os valores encontrados para classificação por conceito ainda são satisfatórios.

4.2 Previsão e Projeção dos Dados

Nesta seção, abordaremos a previsão e projeção dos dados obtidos. Em sua maioria, os dados foram analisados no software Weka (já abordado anteriormente). Inicialmente, utilizamos o Modelo de Regressão Linear disponível no Weka para determinar a nossa equação e os resultados podem ser vistos nas figuras 14 e 15. Vale ressaltar que para a tarefa de predição não foi realizada seleção de variáveis.

Na Figura 14 é possível perceber que a Equação para a Nota Final (G3) leva em consideração alguns atributos (tais como idade, ausências, G1, G2 entre outras). Podemos perceber também que, como resultado da validação cruzada, o Coeficiente de Correlação encontrado foi de 0.902 para a nossa base que apresenta 395 instâncias. Resultado similar encontramos na Figura 15, que apresenta um Coeficiente de Correlação de 0.9148.

Um dos objetivos iniciais era tentar efetuar uma previsão de aprovação dos alunos dados alguns atributos. Diante disso, apresentamos a Figura 16, que nos fornece uma importante análise no que tange a previsão de aprovação ou reprovação do aluno para a Base de Matemática. Na figura, é possível perceber que existe uma clara correlação entre a Nota da 1^a avaliação (G1) com o resultado final esperado (G3-Predict). Da mesma forma, podemos dizer que fica clara a correlação entre a Nota da 2^a avaliação (G2) com o resultado final esperado (G3-Predict). Se formos interpretar de acordo com o Coeficiente de Correlação de Pearson (que abordaremos mais profundamente logo em seguida), os valores dos ρ_1 (quando relacionamos G1 x G3) e ρ_2 (quando relacionamos G2 x G3) podem ser dados por: $\rho_1 = 0.8$ e $\rho_2 = 0.90$, ou seja, apresentam correlação forte e correlação muito forte, respectivamente.

Apresentamos a Figura 17, que nos fornece uma importante análise no que tange a previsão de aprovação ou reprovação do aluno para a Base de Português. Na figura, é possível perceber que existe uma clara correlação entre a Nota da 1^a avaliação (G1) com o resultado final esperado (G3-Predict). Da mesma forma, podemos dizer que fica clara a correlação entre a Nota da 2^a avaliação (G2) com o resultado final esperado (G3-Predict). Se formos interpretar de acordo com o Coeficiente de Correlação de Pearson (que abordaremos mais profundamente logo em seguida), os valores dos ρ_1 (quando relacionamos G1 x G3) e ρ_2 (quando relacionamos G2 x G3) podem ser dados por: $\rho_1 = 0.83$ e $\rho_2 = 0.92$, ou seja, apresentam correlação forte e correlação muito forte, respectivamente.

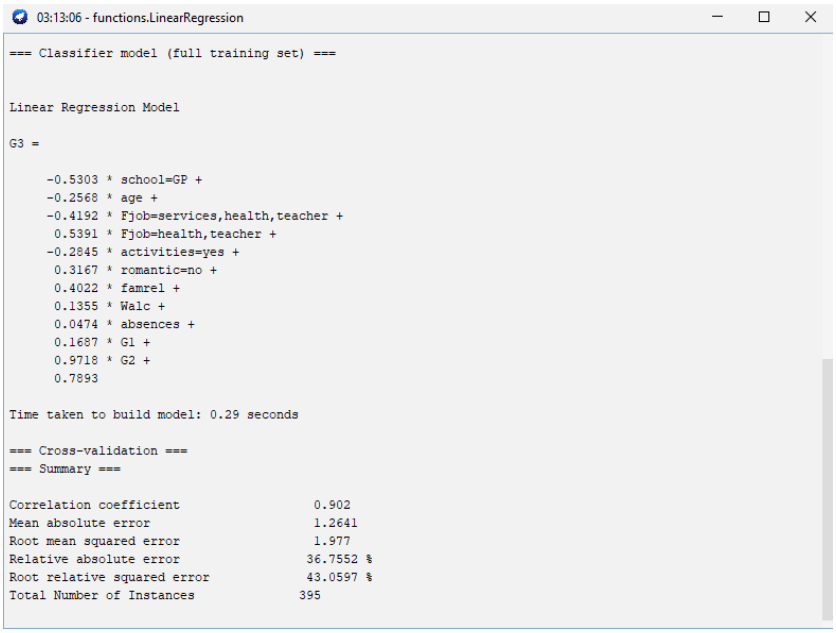

Figure 14: **Resultado da Regressão Linear para a Base de Matemática. Fonte: Autores**

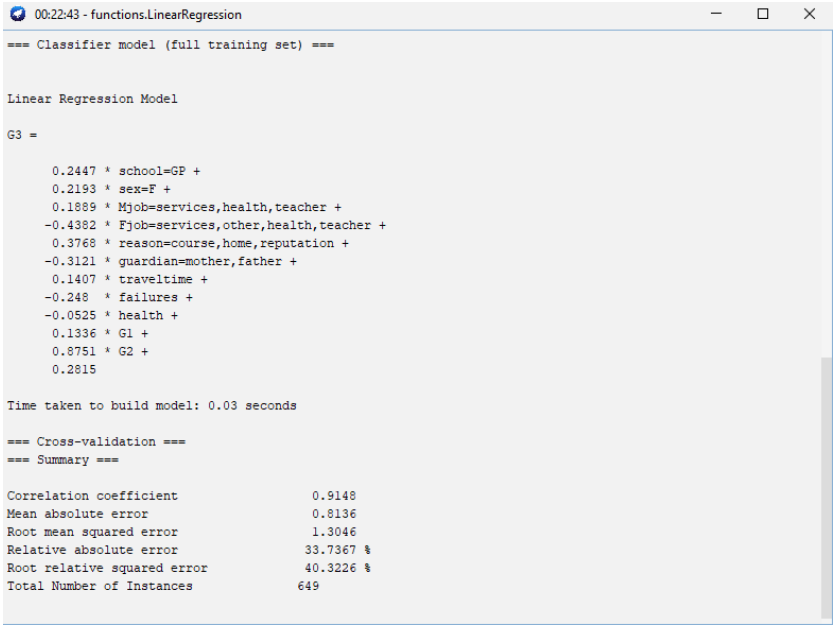

Figure 15: **Resultado da Regressão Linear para a Base de Português. Fonte: Autores**

4.3 Outras Análises

Para ampliar a nossa investigação utilizamos os atributos **Tempo de Viagem até a escola**, **Tempo Livre**, **Número de Faltas** e **Tempo de Estudo**. Para facilitar a visualização e o entendimento dos resultados obtidos em cada base, as Figuras estarão dispostas da seguinte maneira: À esquerda utilizou-se a base de Matemática e à direita a base de Português.

A nossa análise estará relacionada com o Coeficiente de Correlação de Pearson ρ (que varia entre 0 e 1 para uma correlação positiva e entre -1 e 0 para uma correlação negativa), no qual podemos interpretá-lo em 5 classes que podem ser dadas abaixo:

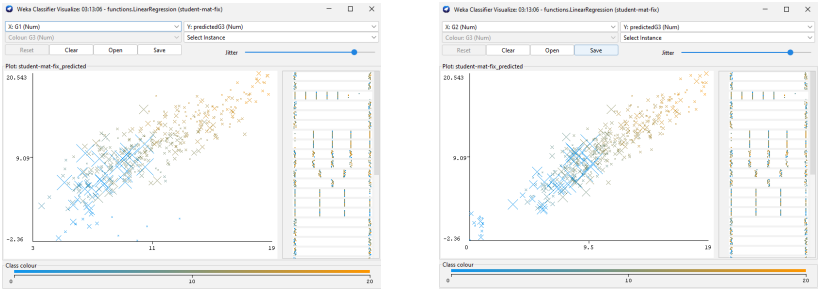

Figure 16: **Gráfico de Correlação entre G1 x G3 esperado e G2 x G3 esperado, respectivamente, para a base de Matemática. Fonte: Autores**

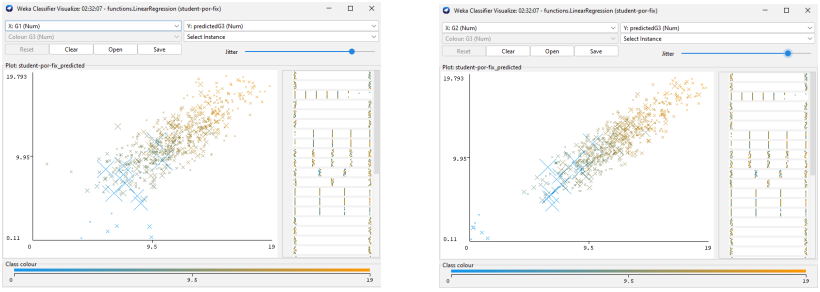

Figure 17: **Gráfico de Correlação entre G1 x G3 esperado e G2 x G3 esperado, respectivamente, para a base de Português. Fonte: Autores**

- $\rho > 0.9$ indica uma correlação muito forte;
- $0.7 < \rho < 0.9$ indica uma correlação forte;
- $0.5 < \rho < 0.7$ indica uma correlação moderada;
- $0.3 < \rho < 0.5$ indica uma correlação fraca;
- $0 < \rho < 0.3$ indica uma correlação desprezível;

Dos atributos analisados, foram separados 4 que, em nossa percepção deveriam ter correlação e (adiantando) apresentam correlação desprezível. Analisando a Figura 18 ($\rho = 0.1$) podemos perceber alguns pontos:

1. Apesar de pensarmos que seria **bem evidente** que, pessoas com mais horas de estudo apresentariam um rendimento melhor, não fica tão claro a partir dos gráficos (tanto para a Base de Matemática quanto para a Base de Português);
2. É possível sim perceber uma pequena tendência (apesar do baixo número de pessoas com uma maior carga horária de estudos) de relação. Isso se deve pelo fato de, apesar do número pequeno de pessoas, quase na totalidade dos casos as pessoas foram aprovadas;
3. Essa tendência também pode ser percebida também para os alunos que tiveram 3 horas semanais de estudo;
4. O que podemos perceber através do Gráfico é que, as pessoas com mais tempo de estudo, na grande maioria, foram aprovados (apenas alguns poucos pontos estão abaixo da "média" que são 10 acertos);

Analisando a Figura 19 ($\rho = 0.03$) podemos perceber alguns pontos:

1. Assim como no item analisado anteriormente, apesar de pensarmos que seria **bem evidente** que, pessoas com maior número de faltas apresentariam um rendimento menor, não fica tão claro a partir dos gráficos (tanto para a Base de Matemática quanto para a Base de Português);
2. O que podemos perceber através do Gráfico é que, as pessoas com mais tempo de estudo, na grande maioria, foram aprovados (apenas alguns poucos pontos estão abaixo da "média" que são 10 acertos);

Analisando a Figura 20 ($\rho = 0.12$) podemos perceber alguns pontos:

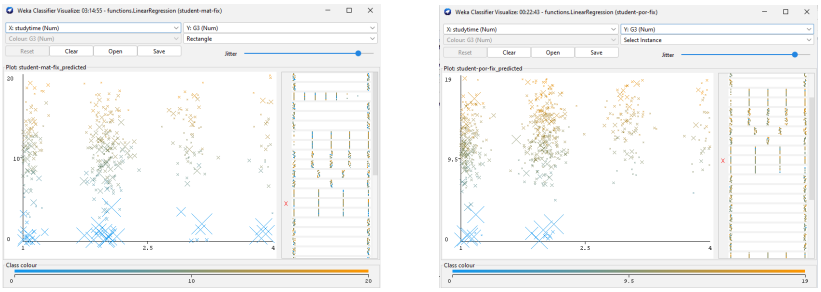

Figure 18: **Gráfico relacionando Tempo de Estudo com Nota Final. Fonte: Autores**

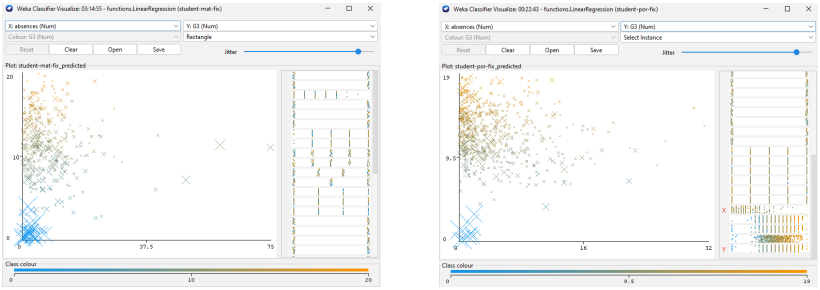

Figure 19: **Gráfico relacionando Número de Faltas com Nota Final. Fonte: Autores**

1. Nesses gráficos, é possível verificar que pessoas os alunos em sua grande maioria (tanto para a Base de Matemática quanto para a Base de Português) residem relativamente próximo da escola, gastando até 2 horas por dia no trajeto de ida e volta para a escola;
2. Em geral esperamos que exista uma tendência/relação entre alunos com maior tempo gasto no trajeto para a escola apresentem um rendimento menor. Analisando o gráfico, não fica tão evidente que isso acontece. Percebemos que os alunos que tem ficam menos tempo no trajeto apresentam notas maiores (de uma forma geral) mas não que essa tendência seja completamente correlacionada;
3. Chegamos à essa conclusão analisando os alunos que "perdem" 4 horas no trajeto e apresentam nota maior do que a Média (10 acertos);
4. Uma possibilidade aventada por nós é a de que, pelos alunos gastarem mais tempo no trajeto, eles podem utilizar esse período para estudar;

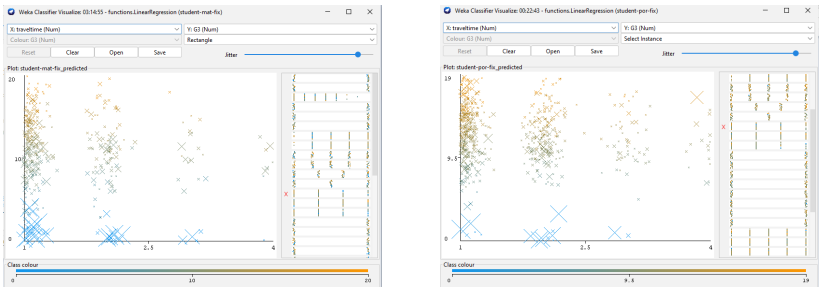

Figure 20: **Gráfico relacionando Tempo Gasto em Viagem com Nota Final. Fonte: Autores**

Analisando a Figura 21 ($\rho = 0.01$) podemos perceber alguns pontos:

1. Nesses gráficos, é possível verificar que pessoas os alunos em sua grande maioria (tanto para a Base de Matemática quanto para a Base de Português) dispõe de tempo livre (acima de 3 horas semanais);
2. Em geral esperamos que exista uma tendência/relação entre alunos com maior tempo livre, poderão se dedicar à alguma outra atividade que possa permitir uma melhoria no seu aprendizado (em geral), mas não existe uma tendência que podemos afirmar isso.

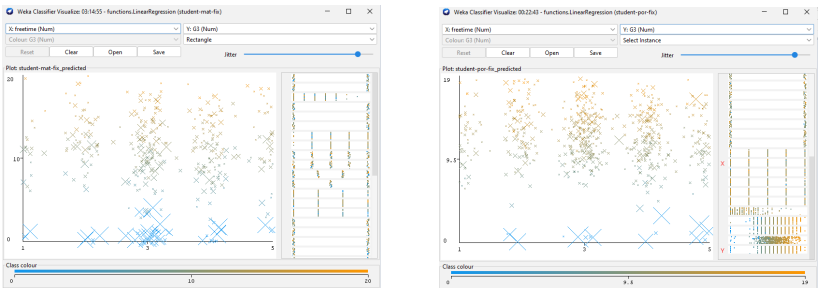

Figure 21: **Gráfico relacionando Tempo Livre com Nota Final. Fonte: Autores**

De uma forma geral, os atributos que apresentam maior correlação com a Nota Final/Aprovação são: **Educação dos Pais** com $\rho_{FormacaoMae} = 0.24$ e $\rho_{FormacaoPai} = 0.21$ e **Expectativa de ingressar no Ensino Superior** com $\rho_{EnsinoSuperior} = 0.33$.

5 Conclusão e Trabalhos Futuros

Pode-se concluir que os modelos construídos podem auxiliar na obtenção das métricas de desempenho dos alunos nas disciplinas de Português e Matemática, seja indicando a aprovação ou reprovação, o conceito obtido ou estipulando qual será a nota final. Com base nos resultados obtidos, os modelos tiveram boa capacidade de generalização e apresentaram resultados satisfatórios, tanto na classificação quanto na predição dos dados.

Para estudos futuros, uma possível extensão deste trabalho seria o emprego de tarefas descritivas, como clusterização e regras de associação, para que fosse possível se aprofundar nas relações que existem entre estes atributos e quais regras ou grupos poderíamos encontrar sobre este aspecto. Essa análise mais aprofundada pode ser de grande ajuda no estudo das questões envolvidas na evasão escolar. Como se trata de uma base relativamente pequena e com dados de outro sistema de ensino, adicionar dados locais também seria interessante. Por fim, para fins de melhoria o pré-processamento poderia ser aprimorado com outras técnicas avaliando-se quais seriam mais indicadas para esta base.

6 Referências

- CESAR, M. V. G. (2017). Classificac¸ao de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.
- Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.
- de Miranda, S. P. (2021). Predicting drug sensitivity of cancer cells based on genomic data. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.
- Fausett, L. (1994). Fundamentals of Neural Network. Prentice Hall.
- Giacomel, F. S. (2016). Um metodo algoritmo para operações na bolsa de valores baseado em ensembles de redes neurais para modelar e prever os movimentos dos mercados de ações. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre. ~
- Goldschmidt, R., Passos, E., and Bezerra, E. (2015). Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações. Elsevier.
- Han, J., Kamber, M., and Kaufmann, M. (2013). Data Mining: Concepts and Techniques. 3rd edition.
- Haykin, S. (2001). Redes Neurais: princ´ıpios e praticas ´ . Bookman, 2nd edition.
- Injadat, M., Moubayeda, A., Nassif, A. B., and Shamia, A. (2020). Systematic ensemble model selection approach for educational data mining.
- Kaura, P., Singhb, M., and Singh, S. J. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques.
- Lorena, A. C. and de Carvalho, A. C. P. L. F. (2007). Uma introdução as support vector machines.
- SAP. O que e machine learning ou aprendizagem de máquina? ´

- Smola, A. J. and Scholkopf, B. (2002). ¨ Learning with Kernels. The MIT Press.
- Zavadzki, S. T. (2020). Diferentes abordagens para o aprendizado da rede neural artificial multilayer perceptron. Master's thesis, Universidade Federal do Parana, Curitiba.
- Montgomery, D. C; PECK, E. A.; VINING, G. G.; Introduction to Linear Regression Analysis - Fifth Edition. Wiley.