

Missão: Formar e aperfeiçoar cidadãos e prestar serviços atendendo às necessidades tecnológicas da sociedade com agilidade, dinâmica e qualidade.

MINERAÇÃO DE DADOS

Projeto - Riscos Sobre Investimentos



“Sound strategy starts with having the right goal.” – Michael Porter



INTEGRANTES:

Juliana Almeida Morroni RA:200372

Arthur Briganti Gini RA:213253

Cleyvesson G. da Silva RA:195743

Sumário:

Introdução	4
1.1 O que é mineração de dados?	4
1.2 Base de dados	5
1.3 Preparação ou Pré-processamento de dados	5
1.3.1 Limpeza de dados	5
1.3.2 Integração dos dados	5
1.3.3 Redução dos dados	5
1.3.4 Transformação dos dados	6
2. Tarefas de mineração de dados	6
2.1 Tarefas	6
3. Métodos (ou Técnicas)	12
3.1 Classificações	12
3.1.1 Árvores de Decisão (Decision Trees)	12
3.1.2 Redes Neurais (Neural Networks)	13
3.2 Predições Numéricas	13
3.2.1 Regressão Linear	13
3.2.2 Regressão Não-Linear	14
3.3 Agrupamento	14
3.3.1 Métodos de Particionamento (Partitioning Methods)	15
3.3.1.1 k-Means	15
3.3.1.2 K - Medoids	15
3.3.2 Métodos Hierárquicos (Hierarchical Methods)	16
3.3.2.1 Aglomerativos	16
3.3.2.2 Divisivos	17
3.4 Associações	17
3.4.1 Mineração de Itens Frequentes (Frequent Itemset Mining)	17
3.4.1.1 Apriori	17
4. Descrição do problema	17
4.1 Mercado financeiro	18
4.2 Base de dados	18
4.2.1: Atributos	20
4.2.2: Conformidade da base?	21
4.3 Escolha da tarefa para o problema	21
5. Metodologia	21
5.1 Pré-processamento	22

.....

5.1.1: Dados faltantes	23
5.1.2: Estratégia de limpeza	23
5.2 Algoritmos utilizados	24
5.2.1: Regressão Linear	25
5.2.2 Multilayer Perceptron	26
5.3 Avaliação da previsão	27
6. Apresentação e Discussão dos resultados	28
6.1 Execução sem pré-processamento	28
6.1.1 Regressão Linear	29
6.1.1.1 Modelo Preditivo para Fechamento	30
6.1.1.2 Modelo Preditivo para Máxima	31
6.1.1.3 Modelo Preditivo para Mínima	32
6.1.1.4 Medidas de precisão da previsão	33
6.1.2 Previsão dos modelos	35
6.1.2.1 Regressão Linear	35
6.1.2.2 Multilayer Perceptron	35
6.2 Execução com pré-processamento	37
6.2.1 Regressão Linear	37
6.2.1.1 Modelo Preditivo para Fechamento	38
6.2.1.2 Modelo Preditivo para Máxima	39
6.2.1.3 Modelo Preditivo para Mínima	40
6.2.1.4 Medidas de precisão da previsão	41
6.2.1 Previsão dos modelos	43
6.2.1.1 Regressão Linear	43
6.2.1.2. Multilayer Perceptron	44
7. Conclusão	45
8. Referências Bibliográficas	45

1. Introdução

Em 1965, Gordon Moore, um dos fundadores da Intel, publicou um artigo no qual observou que a quantidade de componentes em um circuito integrado (CI) estava dobrando aproximadamente a cada ano desde sua invenção. E essa taxa permaneceria por pelo menos mais dez anos. Moore atualizou sua estimativa para períodos de dois anos, em vez de um ano. Essa elevada taxa de crescimento na quantidade de componentes do CI está diretamente relacionada à velocidade de processamento e capacidade de memória dos computadores e também tem servido de meta para a indústria de hardware computacional.

Paradoxalmente, esses avanços da tecnologia, têm produzido um problema de superabundância de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimento destes. Nesse contexto, é necessária a aplicação de técnicas e ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento para uma tomada de decisão estratégica nos negócios e até no dia a dia de cada um de nós.

1.1 O que é mineração de dados?

Data Mining ou Mineração de Dados ¹ é o processo de explorar grandes quantidades de dados à procura de padrões consistentes. Como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Data mining é formada por um conjunto de ferramentas e técnicas que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística. Estes são capazes de explorar um conjunto de dados, extraíndo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. O conhecimento em Data Mining pode ser apresentado por essas ferramentas de diversas formas: agrupamentos, hipóteses, regras, árvores de decisão, grafos, ou dendrogramas.

A novidade da era do computador é o volume enorme de dados que não pode mais ser examinado à procura de padrões em um prazo de tempo razoável. A solução é instrumentalizar o próprio computador para detectar relações que sejam novas e úteis. Data Mining (DM) surge para essa finalidade e pode ser aplicada tanto para a pesquisa científica como para impulsionar a lucratividade da empresa madura, inovadora e competitiva. Por fim, o conhecimento retirado dos dados é algo que permite uma tomada de decisão para agregação de valor.

A mineração de dados ⁵, no entanto, é parte de um processo mais amplo, chamado como Descoberta de Conhecimento em Base de Dados (KDD), que se refere a todo o processo de extração de conhecimentos a partir de dados, incluindo também, a seleção e integração das bases de dados, a limpeza da base, a seleção e transformação dos dados, a mineração e a avaliação de dados.

.....

1.2 Base de dados

Coleção organizada de dados, valores quantitativos ou qualitativos referentes a um conjunto de itens. O nível mais básico de abstração onde informações e conhecimentos podem ser extraídos.

1.3 Preparação ou Pré-processamento de dados

Etapa anterior à mineração, que visa preparar os dados para uma análise mais eficiente. Nessa etapa, inclui a limpeza da base (remoção de ruídos e dados inconsistentes), a integração (combinação de dados obtidos a partir de múltiplas fontes), a seleção ou redução (escolha dos dados relevantes à análise) e a transformação (transformação ou consolidação dos dados em formatos apropriados para a mineração).

1.3.1 Limpeza de dados

Frequentemente, os dados são encontrados com diversas inconsistências: registros incompletos, valores errados e dados inconsistentes. A etapa de limpeza dos dados visa eliminar estes problemas de modo que eles não influenciam no resultado dos algoritmos usados na tarefa de mineração. As técnicas usadas nesta etapa vão desde a remoção do registro com problemas, passando pela atribuição de valores padrões para atributos incompletos, até a aplicação de técnicas de agrupamento para auxiliar na descoberta dos melhores valores.

1.3.2 Integração dos dados

É comum obter-se os dados a serem minerados de diversas fontes: banco de dados, arquivos textos, planilhas, data warehouses, vídeos, imagens, entre outras. Surge então, a necessidade da integração destes dados de forma a termos um repositório único e consistente. Para isto, é necessária uma análise aprofundada dos dados observando redundâncias, dependências entre as variáveis e valores conflitantes (categorias diferentes para os mesmos valores, chaves divergentes, regras diferentes para os mesmos dados, entre outros).

1.3.3 Redução dos dados

O volume de dados usado na mineração costuma ser alto. Em alguns casos, este volume é tão grande que torna o processo de análise dos dados e da própria mineração impraticável. Nestes casos, as técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, porém, sem perder a representatividade dos dados originais. Isto permite que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado. As estratégias adotadas

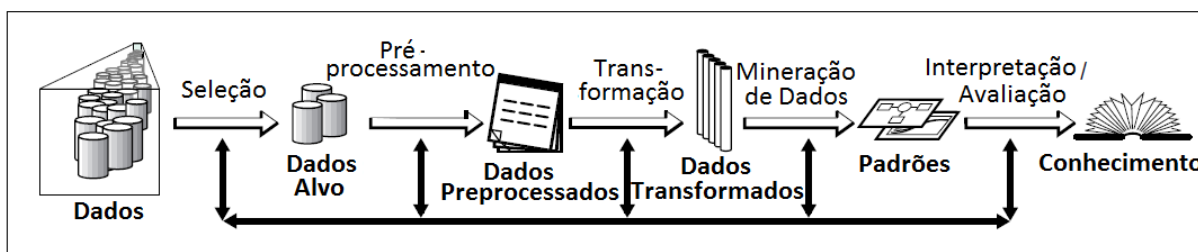
.....

nesta etapa são: criação de estruturas otimizadas para os dados (cubos de dados), seleção de um subconjunto dos atributos, redução da dimensionalidade e discretização.

1.3.4 Transformação dos dados

A etapa de transformação dos dados merece destaque. Alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Nestes casos, é necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes).

2. Tarefas de mineração de dados



2.1 Tarefas

As tarefas da mineração de dados são os tipos de descoberta que se pretende realizar em uma base de dados, ou seja, são informações que se deseja extrair. Para determinar qual tarefa a ser resolvida, deve-se ter um conhecimento do domínio da aplicação e saber o tipo de informação que se quer obter.

A Mineração de Dados ³ é comumente classificada pela sua capacidade em realizar determinadas tarefas. As tarefas mais comuns são:

Análise descritiva de dados



Os algoritmos de aprendizagem de máquina são ferramentas poderosas para a descoberta de conhecimentos em base de dados. A análise descritiva é uma etapa inicial que não requer elevado nível de sofisticação. É uma tarefa utilizada para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido. Especificamente, essa análise permite investigar a distribuição de frequência, as medidas de centro e variação, e as medidas de posição relativa e associação dos dados.

Predição: Classificação e Estimação



É uma tarefa preditiva da mineração de dados que associa objetos a determinadas classes, ou seja, ela pode prever automaticamente a classe de um novo dado. Sob essa perspectiva, a classificação e estimação constituem os dois principais problemas de predição, sendo que a

classificação é usada para prever valores discretos, ao passo que a estimação é usada para prever valores contínuos.

Uma das tarefas mais comuns, a *Classificação*: Visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa: Perfil Técnico, Perfil Negocial e Perfil Gerencial. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa. A tarefa de classificação pode ser usada por exemplo para:

- Determinar quando uma transação de cartão de crédito pode ser uma fraude;
- Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
- Diagnosticar onde uma determinada doença pode estar presente;
- Identificar quando uma pessoa pode ser uma ameaça para a segurança.

Estimação ou Regressão: A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor. A tarefa de estimação pode ser usada por exemplo para:

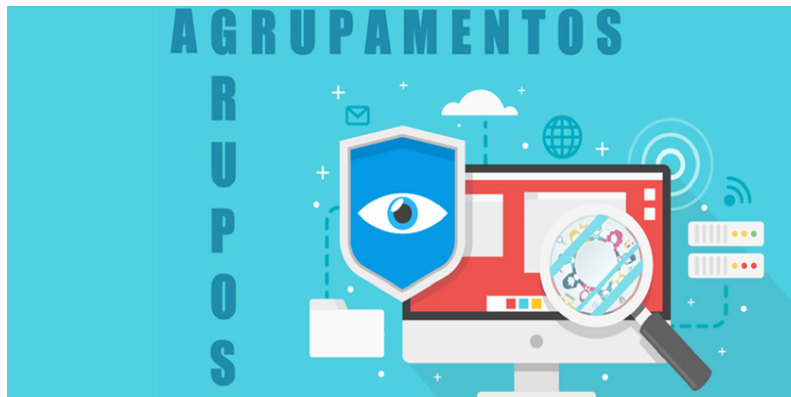
- Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
- Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

Predição: A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Exemplos:

- Prever o valor de uma ação três meses adiante;
- Prever o percentual que será aumentado de tráfego na rede se a velocidade aumentar;
- Prever o vencedor do campeonato baseando-se na comparação das estatísticas dos times. Alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

.....

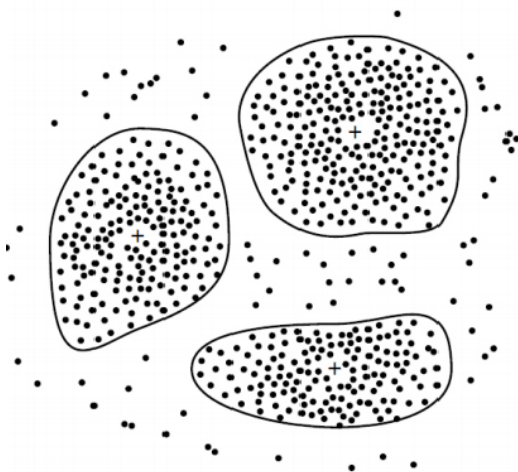
Análise de grupos: Agrupamento



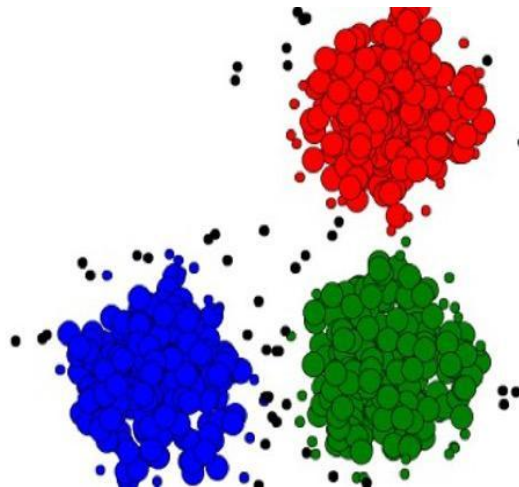
Agrupamento (Clustering) é o nome dado ao processo de separar um conjunto de objetos em grupos de objetos similares. Diferente da tarefa de classificação, o agrupamento considera dados de entrada não rotulados (a classe do objeto não é conhecida), esse processo é denominado treinamento não supervisionado ou aprendizagem não supervisionada. No entanto, esse processo é utilizado para identificar tais grupos, assim, cada grupo formado pode ser visto como uma nova classe de objetos.

A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa não tem a pretensão de classificar, estimar ou prever o valor de uma variável, ela apenas identifica os grupos de dados similares. Exemplos:

- Segmentação de mercado para um nicho de produtos;
- Para auditoria, separando comportamentos suspeitos;
- Reduzir para um conjunto de atributos similares registros com centenas de atributos.



Registros agrupados em três *clusters*



As aplicações das tarefas de agrupamento são as mais variadas possíveis: pesquisa de mercado, reconhecimento de padrões, processamento de imagens, análise de dados, segmentação de mercado, taxonomia de plantas e animais, pesquisas geográficas, classificação de documentos da Web, detecção de comportamentos atípicos (fraudes), entre outras. Geralmente a tarefa de agrupamento é combinada com outras tarefas, além de serem usadas na fase de preparação dos dados.

Associação (Association)

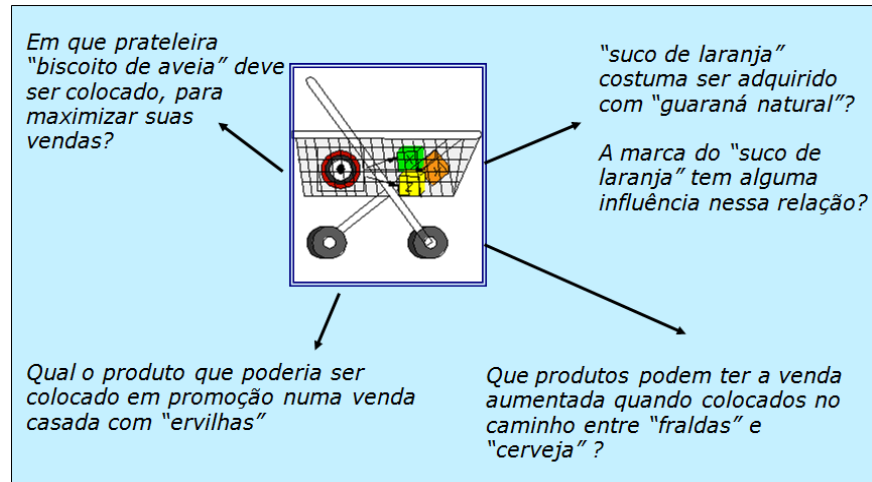
A \Rightarrow B

A análise por associação, também conhecida por mineração de regras de associação, é a tarefa que corresponde à descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados. Costuma ser usado em ações de marketing e para o estudo de bases de dados transacionais. Há dois aspectos centrais na mineração de regras de associação: a proposição ou construção eficiente das regras de associação e a quantificação da significância das regras propostas. Ou seja, um bom algoritmo de mineração de regras de associação precisa ser capaz de propor associações entre itens que sejam esteticamente relevantes para o universo representado pela base de dados.

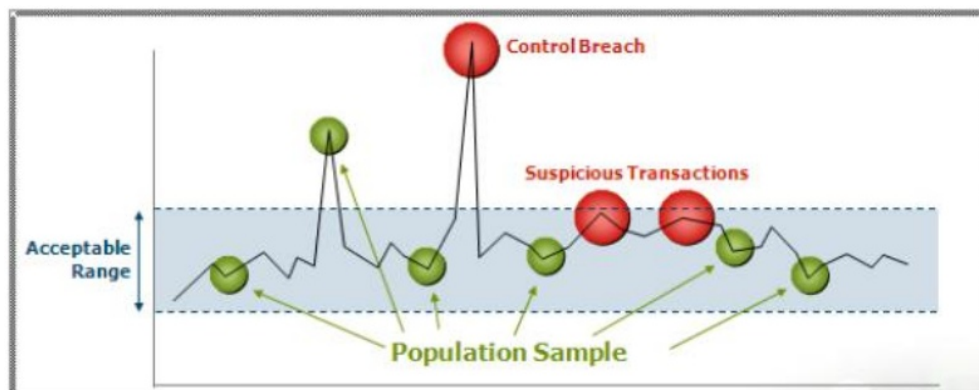
A tarefa de associação tem como objetivo principal encontrar padrões do tipo $X \rightarrow Y$, ou seja, o quanto X implica em Y . onde X e Y são conjuntos distintos. Por exemplo, um cliente que compra o item A , frequentemente compra também o item B . Através dessa tarefa pode-se estimar que um conjunto de item " X " possui uma tendência a se repetir frequentemente em conjunto com o item " Y ".

É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras" (Market Basket), onde identificamos quais produtos são levados juntos pelos consumidores. Alguns exemplos:

- Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais;
- Identificar os usuários de planos que respondem bem a oferta de novos serviços.



Detecção de anomalias



Uma base de dados ⁵ pode conter objetos que não possuem as características comuns dos dados, esses dados são conhecidos como anomalias ou valores discrepantes (outliers). A maioria das ferramentas de mineração acabam descartando as anomalias, entretanto, em algumas aplicações, como na detecção de fraudes, os eventos raros podem ser mais informativos do que aqueles que ocorrem regularmente e até mesmo em banco de dados referentes à saúde pública.

Uma característica marcante das anomalias é que elas compõem uma classe que ocorre com frequência bem inferior. Isso faz com que os algoritmos de classificação sejam impactados, forçando o uso de algoritmos e medidas de desempenho específicos para tratar tais problemas.

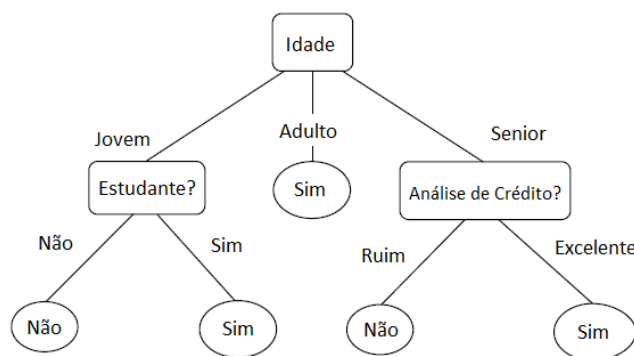
3. Métodos (ou Técnicas)

Durante o processo de mineração, diversas técnicas devem ser testadas e combinadas a fim de que comparações possam ser feitas e então a melhor técnica (ou combinação de técnicas) seja utilizada. Em seguida será demonstrado alguns tipo de técnicas mais recorrentes:

3.1 Classificações

As técnicas de classificação podem ser supervisionadas e não-supervisionadas. São usadas para prever valores de variáveis do tipo categóricas.

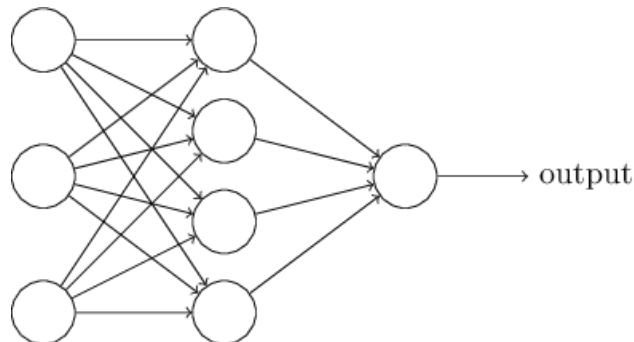
3.1.1 Árvores de Decisão (Decision Trees)



O método de classificação por Árvore de Decisão, funciona como um fluxograma em forma de árvore, onde cada nó indica um teste feito sobre um valor (por exemplo, idade > 20). As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Após a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore, começando no nó raiz até chegar a uma folha. Pela estrutura que formam, as árvores de decisões podem ser convertidas em Regras de Classificação.

O sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade. Apesar de ser uma técnica extremamente poderosa, é necessário uma análise detalhada dos dados que serão usados para garantir bons resultados.

3.1.2 Redes Neurais (Neural Networks)

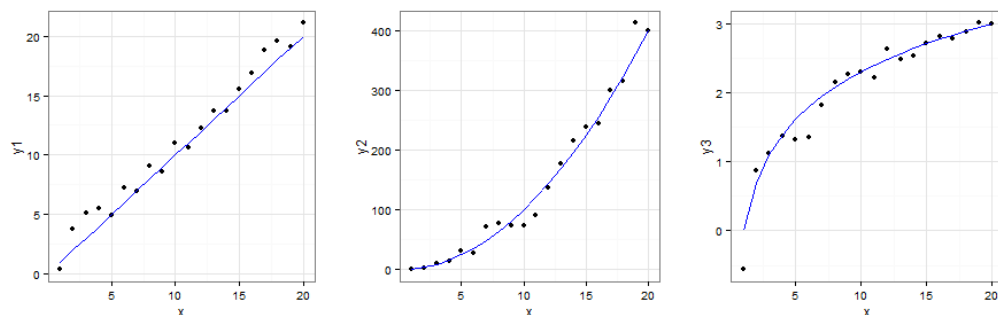


É uma técnica que tem origem na psicologia e na neurobiologia. Consiste basicamente em simular o comportamento dos neurônios. De maneira geral, uma rede neural pode ser vista como um conjunto de unidades de entrada e saída conectados por camadas intermediárias e cada ligação possui um peso associado. Durante o processo de aprendizado, a rede ajusta estes pesos para conseguir classificar corretamente um objeto. É uma técnica que necessita de um longo período de treinamento, ajustes finos dos parâmetros e é de difícil interpretação, não sendo possível identificar de forma clara a relação entre a entrada e a saída. Em contrapartida, as redes neurais conseguem trabalhar de forma que não sofram com valores errados e também podem identificar padrões para os quais nunca foram treinados.

3.2 Predições Numéricas

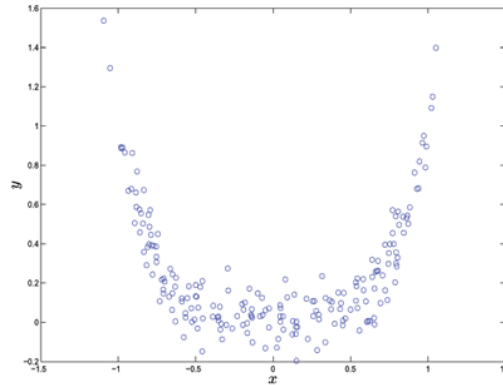
Os métodos de predição visam descobrir um possível valor futuro de uma variável. As predições numéricas visam prever valores para variáveis contínuas.

3.2.1 Regressão Linear



As regressões são chamadas de lineares quando a relação entre as variáveis preditoras e a resposta segue um comportamento linear. Neste caso, é possível criar um modelo no qual o valor de Y é uma função linear de X. Como: $y = b + wx$. Pode-se utilizar o mesmo princípio para modelos com mais de uma variável preditora.

3.2.2 Regressão Não-Linear



Nos modelos de regressão não-linear, a relação entre as variáveis preditoras e a resposta não segue um comportamento linear. Por exemplo, a relação entre as variáveis pode ser modelada como uma função polinomial. Ainda, para estes casos (Regressão Polinomial), é possível realizar uma conversão para uma regressão linear.

3.3 Agrupamento

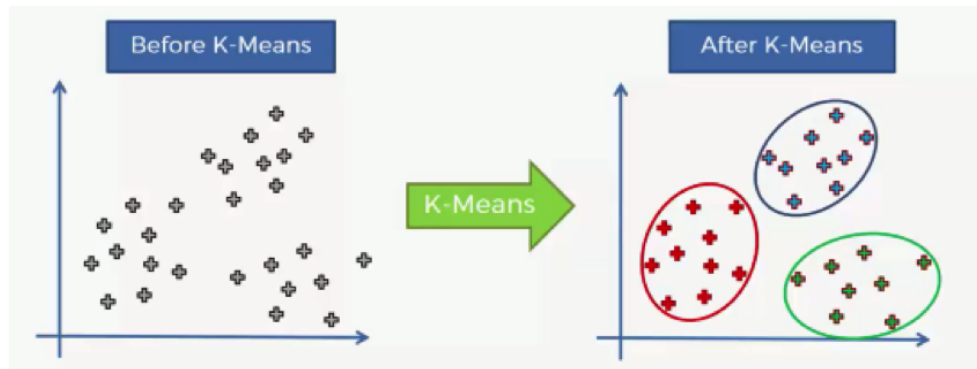
As técnicas de agrupamento são consideradas como não supervisionadas. Dado um conjunto de registros, são gerados agrupamentos (ou cluster), contendo os registros mais semelhantes. Em geral, as medidas de similaridade usadas são as medidas de distâncias tradicionais (Euclidiana, Manhattan, etc).

Por trabalhar com o conceito de distância (similaridade) entre os registros, geralmente é necessário realizar a transformação dos diferentes tipos de dados (ordinais, categóricos, binários, intervalos) para uma escala comum, exemplo [0.0, 1.0]. Podemos classificar os algoritmos de agrupamento nas seguintes categorias:

3.3.1 Métodos de Particionamento (Partitioning Methods)

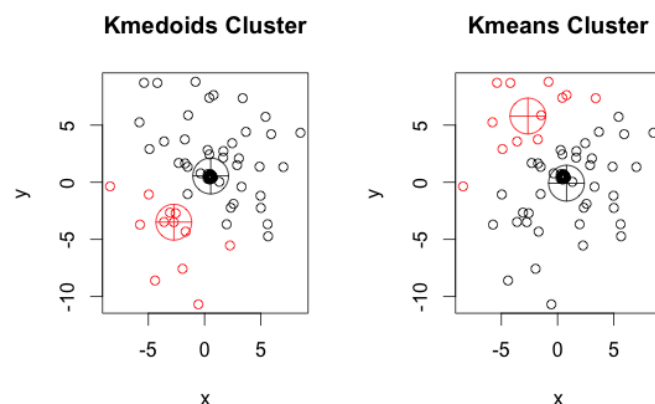
Dado um conjunto D de dados com n registros e k o número de agrupamentos desejados, os algoritmos de particionamento organizam os objetos em k agrupamentos, tal que $k \leq n$. Os algoritmos mais comuns de agrupamento são: k-Means e k-Medoids.

3.3.1.1 k-Means



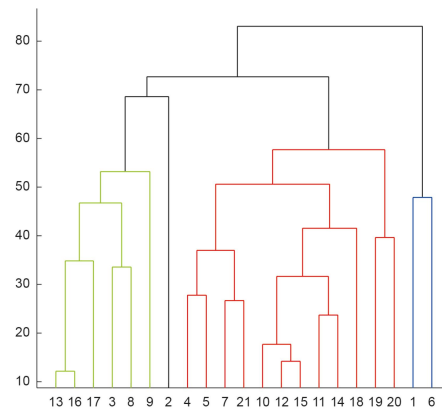
Esse algoritmo usa o conceito da *centróide*. Dado um conjunto de dados, o algoritmo seleciona de forma aleatória k registros, cada um representando um agrupamento. Para cada registro restante, é calculada a similaridade entre o registro analisado e o centro de cada agrupamento. O objeto é inserido no agrupamento com a menor distância, ou seja, maior similaridade. O centro do cluster é recalculado a cada novo elemento inserido. Diferentes variações surgiram: implementando otimizações para escolha do valor do k, novas medidas de dissimilaridade e estratégias para o cálculo do centro do agrupamento. Uma variação bem conhecida do k-Means é o k-Modes. Nesse caso, ao invés de calcular o centro do agrupamento através da média de distância dos registros, ele usa a moda.

3.3.1.2 K - Medoids



É uma variação do k-Means. Neste algoritmo, ao invés de calcular o centro do agrupamento e usá-lo como referência, trabalha-se com o conceito do objeto mais central do agrupamento. As variações mais conhecidas são os algoritmos PAM (Partitioning Around Medoids) e CLARA (Clustering LARge Applications).

3.3.2 Métodos Hierárquicos (Hierarchical Methods)



A idéia básica dos métodos hierárquicos é criar o agrupamento por meio da aglomeração ou da divisão dos elementos do conjunto. A forma gerada por estes métodos é um dendrograma (gráfico em formato de árvore). Dois tipos básicos de métodos hierárquicos podem ser encontrados: Aglomerativos e Divisivos.

3.3.2.1 Aglomerativos

Adotam uma estratégia bottom-up onde, inicialmente, cada objeto é considerado um agrupamento. A similaridade é calculada entre um agrupamento específico e os outros agrupamentos. Os agrupamentos mais similares vão se unindo e formando novos agrupamentos. O processo continua, até que exista apenas um agrupamento principal. Os algoritmos AGNES (AGglomerative NESTing) e CURE (Clustering Using Representatives) utilizam esta estratégia.

3.3.2.2 Divisivos

Adotam uma estratégia top-down, onde inicialmente todos os objetos estão no mesmo agrupamento. Os agrupamentos vão sofrendo divisões, até que cada objeto representa um agrupamento. O algoritmo DIANA (Dlvisive ANALysis) utiliza esta estratégia.

3.4 Associações

É uma das técnicas mais conhecidas de mineração de dados, devido ao problema da Análise da Cesta de Compras. Consiste em identificar o relacionamento dos itens mais frequentes em um determinado conjunto de dados, e permite obter resultados do tipo: SE compra leite e pão TAMBÉM compra manteiga. Esta construção recebe o nome de Regra de Associação (Association Rules).

3.4.1 Mineração de Itens Frequentes (Frequent Itemset Mining)

Essa técnica pode ser visualizada em duas etapas: primeiro, um conjunto de itens frequentes (Frequent Itemset) é criado, respeitando um valor mínimo de frequência para os itens. Depois, as regras de associação são geradas pela mineração desse conjunto. Para garantir resultados válidos, os conceitos de suporte e confiança são utilizados em cada regra produzida. A medida de suporte indica o percentual de registros (dentro de todo o conjunto de dados) que se encaixam nessa regra. Já a confiança mede o percentual de registros que atendem especificamente a regra, por exemplo, o percentual de quem compra leite e pão e também compra manteiga.

Para uma regra ser considerada forte, ela deve atender a um certo grau mínimo de suporte e confiança.

3.4.1.1 Apriori

Um dos mais tradicionais algoritmos de mineração utilizando ⁶ a estratégia de itens frequentes é o Apriori. Diversas variações deste algoritmo, envolvendo o uso de técnicas de hash, redução de transações, particionamento e segmentação podem ser encontrados

Apesar de cada método possuir suas peculiaridades e apresentar melhor resultado com um certo tipo de dado, não existe uma classificação única para a escolha e aplicação destes métodos.

4. Descrição do problema

O objetivo do projeto é a mitigação do risco de investidores a partir do uso de técnicas de mineração de dados, como a técnica de predição. Obtendo uma ferramenta de análise de ativos financeiros.

A partir disso, o projeto inclui estimar valores, e comparar a porcentagem de erros gerados nos dois métodos abordados (Regressão Linear e Multilayer Perceptron) que serão explicados neste relatório, também foram feitas as análises com e sem Pré-processamento para assim, após as comparações, chegarmos a um resultado final mais confiável.

4.1 Mercado financeiro

A base foi retirada das ações da petrobras (petr4 h1) bovespa, pois pela nossa pesquisa, avaliamos que essa é uma das ações mais movimentadas (negociadas) e mais influentes no mercado brasileiro. Muitas pessoas acabam se interessando nelas, pois abordam diversas áreas. Assim, influenciando cada vez mais pessoas em sua compra.

É uma grande Estatal de interesse de todos, e seu viés está relacionado com vários aspectos da nossa sociedade atual.

.....

Mais negociadas

PETR4.SA	-0,56%	R\$ 26,68
VALE3.SA	+1,9%	R\$ 49,46
VVAR3.SA	-2,74%	R\$ 4,61
KROT3.SA	+5,1%	R\$ 10,10
ITSA4.SA	-0,35%	R\$ 11,45

Exemplo das ações mais negociadas

4.2 Base de dados

A base de dados possui atributos como, abertura, fechamento, máxima, mínima, volume, data e hora referentes a movimentações e o volume de ações a cada hora do dia. Porém foram selecionados atributos específicos para nossa análise, que serão abordados logo em seguida.

INDEX	DATA	HORA	ABERTURA	MÁXIMA	MÍNIMA	FECHAMENTO	VOLUME	TENDÊNCIA	
			R\$5.287.816.000	R\$5.315.156.000	R\$5.257.776.000	5287148000		SOMA DOS VALORES	
1	2018.01.02	10:00	1.562.000	1.571.000	1.562.000	1.570.000	3144	Dados	Valores
2	2018.01.02	11:00	1.570.000	1.578.000	1.567.000	1.573.000	3690	Index	4
3	2018.01.02	12:00	1.574.000	1.581.000	1.571.000	1.579.000	5114	Data	2018.01.02
4	2018.01.02	13:00	1.578.000	1.592.000	1.577.000	1.591.000	6957	Hora	13:00:00
5	2018.01.02	14:00	1.591.000	1.596.000	1.589.000	1.592.000	6242	Abertura	R\$ 1.578.000
6	2018.01.02	15:00	1.591.000	1.594.000	1.588.000	1.591.000	5192	High	R\$ 1.592.000
7	2018.01.02	16:00	1.591.000	1.595.000	1.589.000	1.592.000	5728	Low	R\$ 1.577.000
8	2018.01.02	17:00	1.592.000	1.597.000	1.591.000	1.597.000	4591	Close	R\$ 1.591.000
9	2018.01.02	18:00	1.597.000	1.597.000	1.597.000	1.597.000	0	Volume	6957
10	2018.01.03	10:00	1.591.000	1.596.000	1.579.000	1.591.000	4780		
11	2018.01.03	11:00	1.591.000	1.604.000	1.587.000	1.600.000	4096		
12	2018.01.03	12:00	1.601.000	1.604.000	1.592.000	1.598.000	4583		
13	2018.01.03	13:00	1.597.000	1.605.000	1.593.000	1.603.000	4446		
14	2018.01.03	14:00	1.602.000	1.605.000	1.598.000	1.604.000	5628		
15	2018.01.03	15:00	1.603.000	1.609.000	1.601.000	1.605.000	6046		
16	2018.01.03	16:00	1.606.000	1.613.000	1.602.000	1.611.000	5159		
17	2018.01.03	17:00	1.612.000	1.613.000	1.605.000	1.611.000	6693		
18	2018.01.03	18:00	1.611.000	1.611.000	1.611.000	1.611.000	0		
19	2018.01.04	10:00	1.619.000	1.636.000	1.619.000	1.633.000	4327		





Tabela gerada no excel, a partir da base de dados retirada da petr4 h1

A partir disso, utilizamos tabelas do excel para que nos auxiliasse na visualização da massa de dados. Assim, fora criado um index para selecionarmos objetos específicos, para que fosse possível realizar as comparações necessárias.

Dados ▾	Valores ▾
Index	55
Data	2018.01.10
Hora	14:00:00
Abertura	R\$ 1.634.000
High	R\$ 1.638.000
Low	R\$ 1.628.000
Close	R\$ 1.628.000
Volume	4143

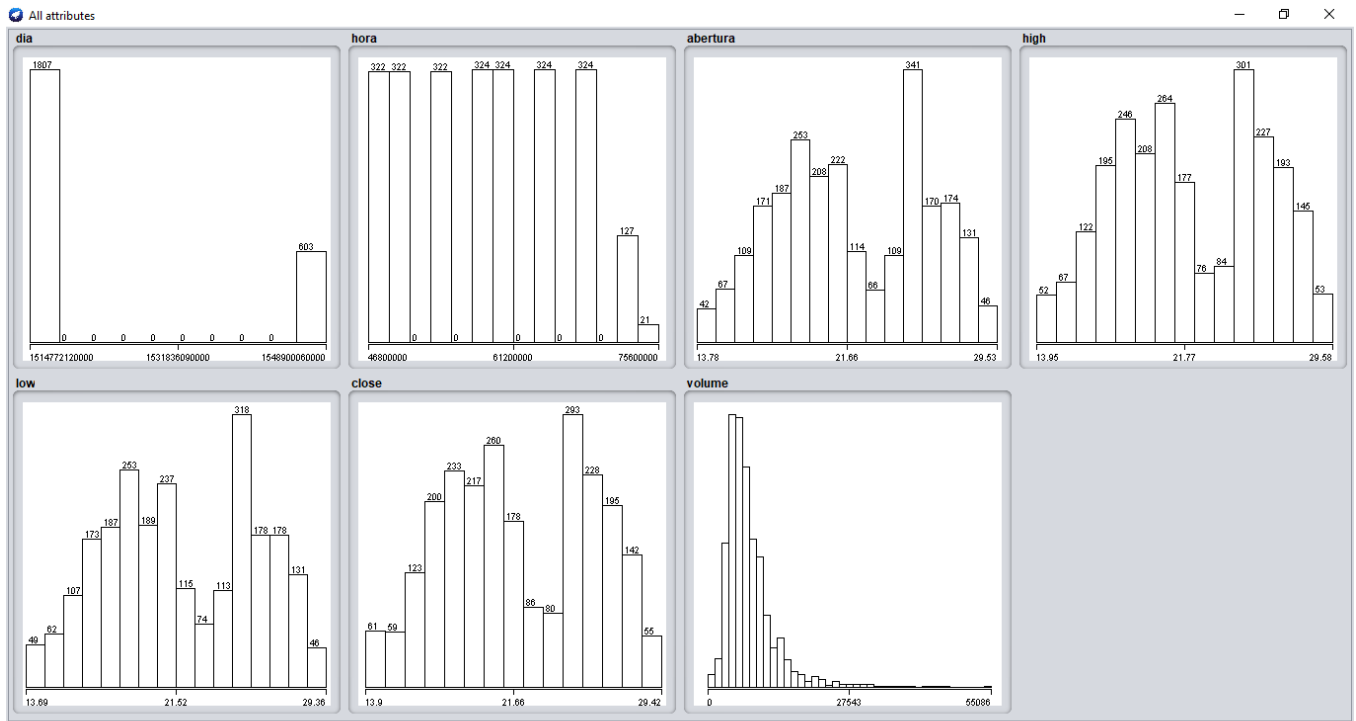
Tabela selecionadora de atributos aleatórios

Neste mesmo âmbito, na tabela do excel utilizada, utilizamos minigráficos para demonstração das movimentações e sua linha de progressão. Assim, deixando mais evidente suas tendências.

ABERTURA ▾	MÁXIMA ▾	MÍNIMA ▾	FECHAMENTO ▾	VOLUME ▾	TENDÊNCIA
					
R\$5.287.816.000	R\$5.315.156.000	R\$5.257.776.000	5287148000		SOMA DOS VALORES

Minigráficos gerados para identificação de tendências

4.2.1: Atributos



Gráficos referentes à movimentação dos valores dos atributos

Os atributos da base são:

- **Abertura**: início das atividades no mercado de ações, valores iniciais;
- **Máxima**: o valor máximo atingido pelas transações durante o dia;
- **Mínima**: o menor valor atingido pelas ações;
- **Dia**: a data que ocorreram as transações;
- **Hora**: o horário em que as transações ocorreram, em nosso caso, a cada uma hora;
- **Volume**: a quantidade de transações que ocorreram durante o dia;
- **Fechamento**: o encerramento das atividades do mercado, o valor remanescente das últimas transações;

Nos gráficos acima, podemos observar a dispersão dos dados para cada atributo. Fica evidente um padrão em Abertura, Máxima, Mínima e Fechamento. Já os atributos Dia e Hora foram desconsiderados para a geração dos modelos, pois apresentavam um comportamento anômalo pelo fato de possuírem métricas diferentes que destoava do padrão dos demais utilizados em nosso processo.

4.2.2: Conformidade da base?

Nossa base de dados, inicialmente se encontra em um bom estado para as análises que pretendemos efetuar, porém possui alguns ruídos (existência de 0 no atributo “volume”), sendo pouco significativa para nosso resultado. Pois, sua quantidade pode ser considerada mínima.

4.3 Escolha da tarefa para o problema

Foi determinada a utilização da tarefa de Predição, pois estimamos um valor numérico. Assim foi selecionada a técnica de Predições numéricas, especificamente utilizando a Regressão Linear, e complementando foi utilizado o método de Redes Neurais (Multilayer Perceptron).

Assim, as duas técnicas utilizadas foram de grande importância para comparação de valores futuros, com a garantia de melhores resultados possíveis.

5. Metodologia

Para cumprir os objetivos do projeto utilizamos o programa Open Source Weka na seguinte versão:



5.1 Pré-processamento

Em nosso Pré-processamento utilizamos o método de exclusão, pois quando o atributo “volume” possuir valor menor que 1, essa tupla será excluída. Esse valor ocorre devido ao horário de encerramento das atividades do mercado, às 18H de cada dia da semana.

A partir disso, descobrimos por meio das tabelas que selecionam atributos aleatórios, que o conteúdo dos atributos desta mesma tupla, sempre obtêm valores fixos e iguais. Assim, a decisão tomada de excluir tais tuplas convém ao processo, a fim de diminuir erros futuros.

DATA	HORA	ABERTURA	MÁXIMA	MÍNIMA	FECHAMENTO	VOLUME
		R\$5.287.816.000	R\$5.315.156.000	R\$5.257.776.000	5287148000	
2018.01.02	10:00	1.562.000	1.571.000	1.562.000	1.570.000	3144
2018.01.02	11:00	1.570.000	1.578.000	1.567.000	1.573.000	3690
2018.01.02	12:00	1.574.000	1.581.000	1.571.000	1.579.000	5114
2018.01.02	13:00	1.578.000	1.592.000	1.577.000	1.591.000	6957
2018.01.02	14:00	1.591.000	1.596.000	1.589.000	1.592.000	6242
2018.01.02	15:00	1.591.000	1.594.000	1.588.000	1.591.000	5192
2018.01.02	16:00	1.591.000	1.595.000	1.589.000	1.592.000	5728
2018.01.02	17:00	1.592.000	1.597.000	1.591.000	1.597.000	4591
2018.01.02	18:00	1.597.000	1.597.000	1.597.000	1.597.000	0
2018.01.03	10:00	1.591.000	1.596.000	1.579.000	1.591.000	4780
2018.01.03	11:00	1.591.000	1.604.000	1.587.000	1.600.000	4096
2018.01.03	12:00	1.601.000	1.604.000	1.592.000	1.598.000	4583
2018.01.03	13:00	1.597.000	1.605.000	1.593.000	1.603.000	4446
2018.01.03	14:00	1.602.000	1.605.000	1.598.000	1.604.000	5628
2018.01.03	15:00	1.603.000	1.609.000	1.601.000	1.605.000	6046
2018.01.03	16:00	1.606.000	1.613.000	1.602.000	1.611.000	5159
2018.01.03	17:00	1.612.000	1.613.000	1.605.000	1.611.000	6693
2018.01.03	18:00	1.611.000	1.611.000	1.611.000	1.611.000	0
2018.01.04	10:00	1.619.000	1.636.000	1.619.000	1.633.000	4327

Exemplo de ruídos e seus respectivos valores da tupla

5.1.1: Dados faltantes

Os dados faltantes da base de dados são os valores contidos no atributo "volume" visto que os valores contidos eram 0, eles não impactaram ou tiveram relevância significativa nos resultados durante a utilização dos algoritmos de processamento.

5.1.2: Estratégia de limpeza

O método de limpeza selecionado foi com base no processo de exclusão visto que a base já apresentava uma boa organização dos dados, não foi necessário aplicar outros métodos de limpeza apenas 21 tuplas foram ignoradas. Totalizando uma porcentagem de não conformidade das linhas da base de 0,0099%.

Current relation	
Relation: airline_passengers	Attributes: 7
Instances: 2410	Sum of weights: 2410

Sem Pré-processamento

Current relation

Relation: airline_passengers-weka.filters.unsupervised...
Instances: 2389

Attributes: 7
Sum of weights: 2389

Com Pré-processamento

weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.RemoveWithValues

About

Filters instances according to the value of an attribute.

More

Capabilities

attributeIndex

7

debug

False

doNotCheckCapabilities

False

dontFilterAfterFirstBatch

False

invertSelection

False

matchMissingValues

False

modifyHeader

False

nominalIndices

first-last

splitPoint

1.0

Open...

Save...

OK

Cancel

A configuração do algoritmo decorre dos valores padrões advindos do programa, foi alterado apenas o *attributeIndex* que localiza o atributo a ser pesquisado para a regra de exclusão da base e o *splitPoint* valor arbitrário que separa as linhas que serão excluídas e as que serão mantidas na base de dados.

5.2 Algoritmos utilizados

Optamos pela escolha de dois algoritmos:

O Multilayer Perceptron e a Regressão linear simples.

Porque?

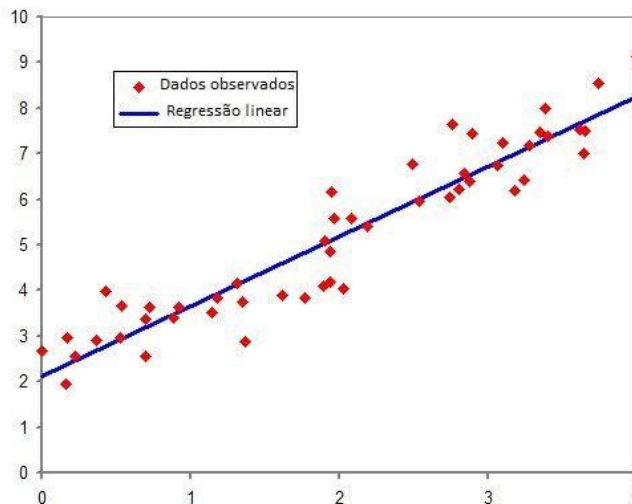
O algoritmo de Regressão linear, tem como objetivo fornecer uma previsão de dados com base em uma série histórica que deve seguir um modelo linear, ou seja, deve se 'encaixar' melhor por uma reta que representa os dados.

Geralmente, os problemas que a Regressão Linear auxilia estão relacionados à tarefas de predição de quantidades ou valores expressivos para determinados propósitos, este algoritmo possui uma eficiência melhor de acordo com a tarefa do nosso trabalho que é prever o comportamento de mercado com base na análise de ativos financeiros e como também, é recomendado para casos que envolvem problemas linearmente separáveis, ou seja, casos em que é possível predeterminar uma separação de grupos apenas tracejando uma reta no plano cartesiano.

Já o Multilayer Perceptron (MLP), por sua vez, é uma rede neural, que utiliza múltiplas camadas de neurônios interligadas uns aos outros por sinapses, tais sinapses possuem pesos, que podem ser aleatórios ou pré definidos. E esses valores influenciam na hora da tomada de decisões ao fim do processamento do algoritmo, a vantagem do uso do MLP é que graças a sua atribuição de pesos para as camadas neurais, os resultados gerados pelo uso desse algoritmo tendem a ser mais próximos de um resultado real esperado.

5.2.1: Regressão Linear

A Regressão Linear é um algoritmo de correlação que verifica a existência de um relacionamento entre duas variáveis distintas, a regressão linear utiliza pontos de dados para traçar uma linha de ajuste ideal para modelar essa relação entre variáveis.



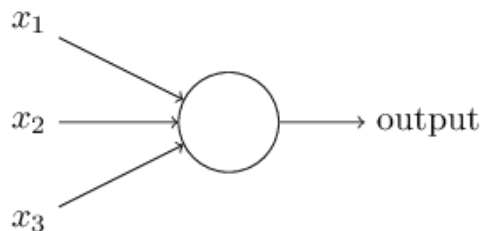
A equação que representa o cálculo de regressão se dá pela seguinte função:

$$y = a * x + b$$

Onde **Y** é uma variável dependente de **X** a variável independente o objetivo dessa equação é encontrar valores para **a** e **b** dado o conjunto de dados

5.2.2 Multilayer Perceptron

Um Perceptron ² é um modelo matemático que recebe várias entradas, x_1, x_2, \dots, x_n que produz uma única saída binária:



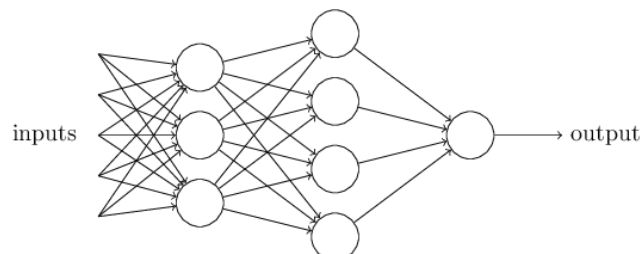
No figura é demonstrado que o modelo possui 3 entradas (x_1, x_2, x_3). A partir daí, nós iremos utilizar uma regra simples para calcular a saída., introduzindo assim, pesos (w_1, w_2, w_3) que serão números reais expressando a importância das respectivas entradas para a saída.

A saída do neurônio, 0 ou 1, é determinada pela soma ponderada, $\sum w_j x_j$, menor ou maior do que algum valor limiar (threshold). Assim como os pesos, o threshold é um número real que é um parâmetro do neurônio. Para colocá-lo em termos algébricos mais precisos:

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{threshold} \end{cases}$$

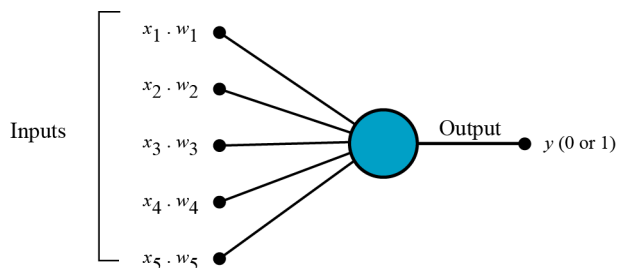
Esse é o modelo matemático básico. Uma maneira de pensar sobre o Perceptron é que é um dispositivo que toma decisões ao comprovar evidências.

O Perceptron não é um modelo completo de tomada de decisão humana. Mas o que o exemplo ilustra é como um Perceptron pode pesar diferentes tipos de evidências para tomar decisões. E deve parecer plausível que uma rede complexa de Perceptrons possa tomar decisões bastante sofisticadas.²



Na rede acima, os Perceptrons parecem ter múltiplos resultados. Na verdade, eles ainda são de saída única. As setas de saída múltiplas são uma maneira útil de indicar que a saída de um Perceptron está sendo usada como entrada para vários outros Perceptrons. Assim, devemos deixar claro que Redes de apenas uma camada só representam funções linearmente separáveis e a utilização de Redes de múltiplas camadas solucionam essa restrição.

Um Perceptron segue o modelo “feed-forward”, o que significa que as entradas são enviadas para o neurônio, processadas e resultam em uma saída. No diagrama abaixo, isso significa que a rede (um neurônio) lê da esquerda para a direita.



O processo de treinamento de um modelo Perceptron consiste em fazer com que o modelo aprenda os valores ideais de pesos e bias. Com o modelo treinado, podemos apresentar novos dados de entrada e o modelo será capaz de prever a saída.

O Perceptron é um classificador linear, ou seja, os problemas solucionados por ele devem ser linearmente separáveis. Além disso, é usado na aprendizagem supervisionada e pode ser usado para classificar os dados de entrada fornecidos. O gráfico a seguir mostra um conjunto de pontos bidimensional que pode ser separado linearmente, sendo possível identificar isso passando uma linha reta entre os dois grupos de cores diferentes:

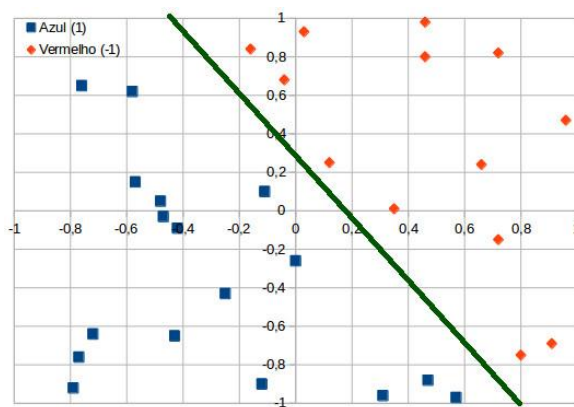


Gráfico com problema linearmente separável em duas dimensões

5.3 Avaliação da previsão

O processo de avaliação da precisão dos algoritmos se baseia nas descrições dos tipos de erros e dos coeficientes abaixo:

.....

- **Coeficiente de correlação:** Esse indicador mede o grau de correlação entre as variáveis de escala métrica. O resultado muito próximo a 1 indica que os valores das variáveis do modelo estão fortemente e positivamente relacionados. Este coeficiente é essencial para o propósito do trabalho, pois este valor indica a eficiência da tarefa preditiva.
- **Erro absoluto médio:** É a medida da diferença média entre duas variáveis contínuas. O resultado próximo a 0 demonstra que o erro do modelo é extremamente baixo e que suporta o resultado do coeficiente de correlação.
- **Erro quadrático médio:** É uma das medidas usadas para descobrir a diferença entre os valores preditos por um modelo, e os valores observados. O resultado próximo a 0 demonstra que o modelo possui uma acurácia alta e está muito próximo a linha ideal.
- **Erro relativo:** É um demonstrativo do quão impreciso o modelo se torna em relação ao valor real buscado na análise, trata-se da diferença entre um valor real e um valor aproximado pode ser afetado por arredondamentos do software.
- **Erro quadrático relativo:** É usado para expressar a acurácia dos resultados numéricos, com sua utilidade ajudará na comparação dos estimadores com o parâmetro ao quadrado sua vantagem de que apresenta valores do erro nas mesmas dimensões da variável analisada.
- **Total de instâncias testadas:** 482

6. Apresentação e Discussão dos resultados

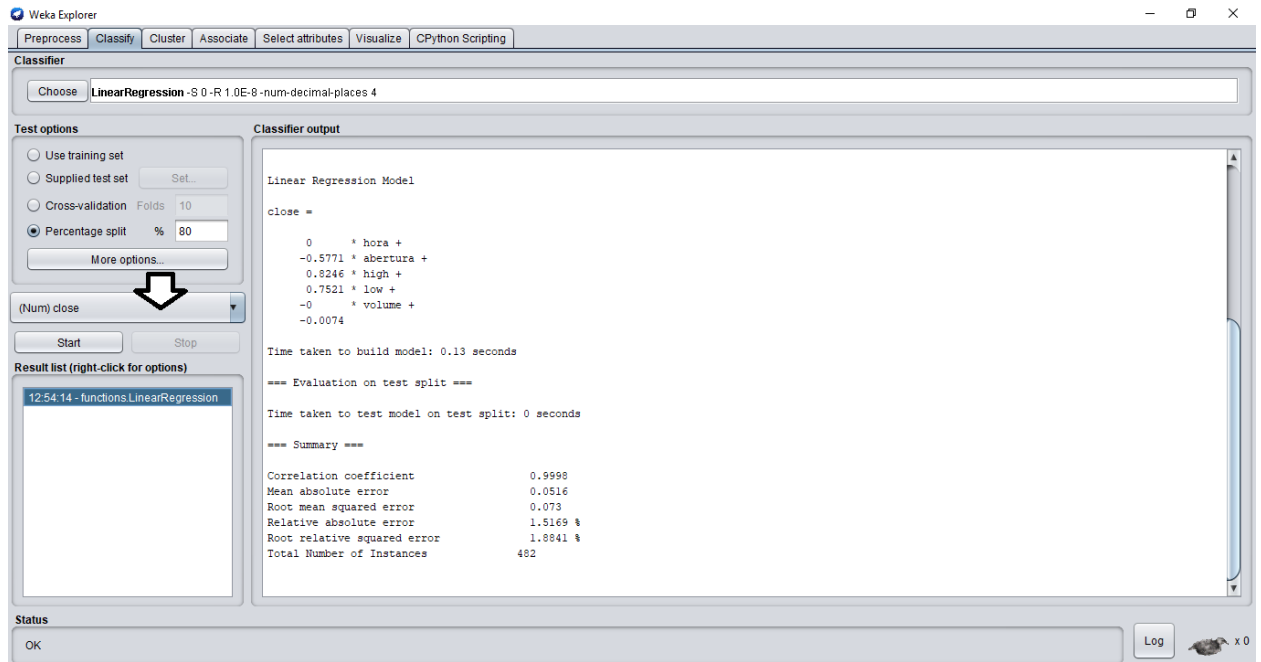
Destacaremos a partir desse tópico os modelos resultantes do processo de mineração de dados, assim como as previsões estatísticas e seus resultados relacionados.

6.1 Execução sem pré-processamento

A partir da base inicialmente obtida e sem realizar processos de pré-processamento dos dados, separamos para ser utilizado 80% da base no treinamento dos modelos e 20% para validação dos mesmos.

Na imagem abaixo está a tela dos resultados gerados a partir do programa WEKA para o primeiro tópico apresentado sobre os modelos preditivos obtidos, após o carregamento da base fomos até a aba de classificação para selecionar o algoritmo a ser aplicado na base, a seta indica o atributo objetivo da predição, essa é a única configuração que é alterada para os demais modelos gerados nesse tópico sem a realização de limpeza da base.

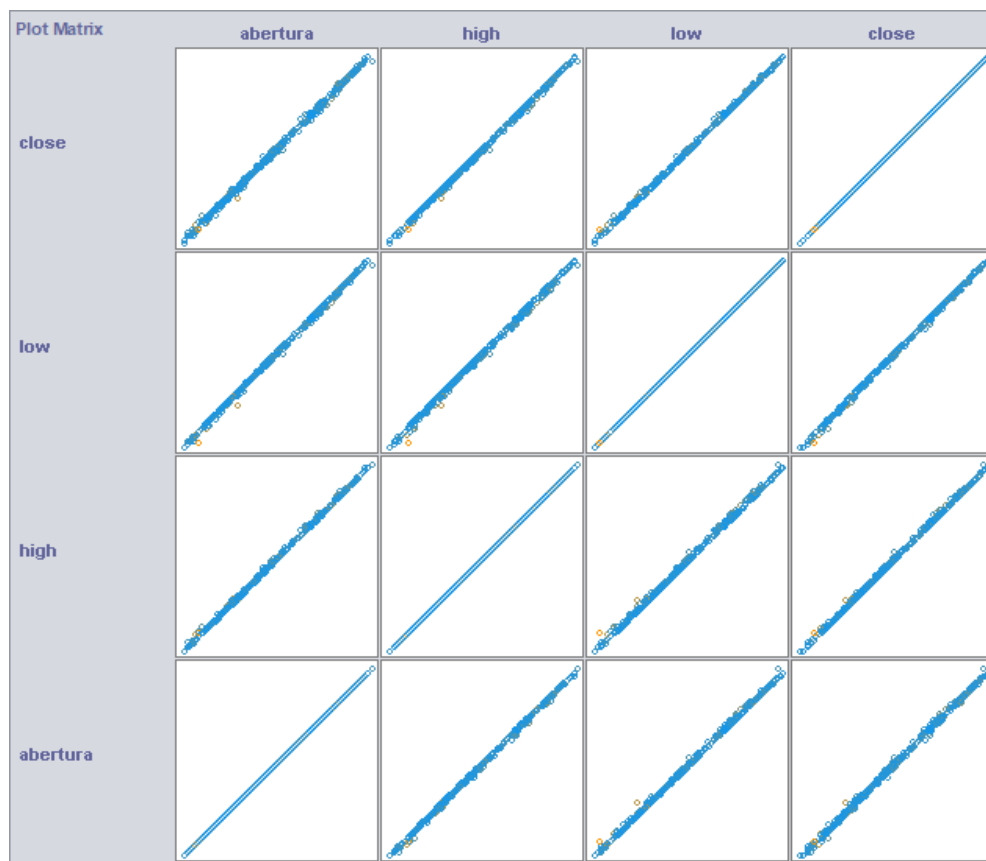
.....



6.1.1 Regressão Linear

Nessa sessão discutiremos a partir do método de regressão linear os modelos e previsões realizadas para os atributos de Fechamento, Máxima e Mínima respectivamente. Após a explicitação do modelo iremos analisar as medidas de precisão das previsões a fim de comparar sua eficácia em representar a realidade.

O gráfico a seguir não possui pré processamento e mostra como a correlação dos atributos é alta e como aos valores respeitam a linha ideal da regressão (diagonal dos gráficos) isso é perceptível pela distância dos pontos à linha principal.

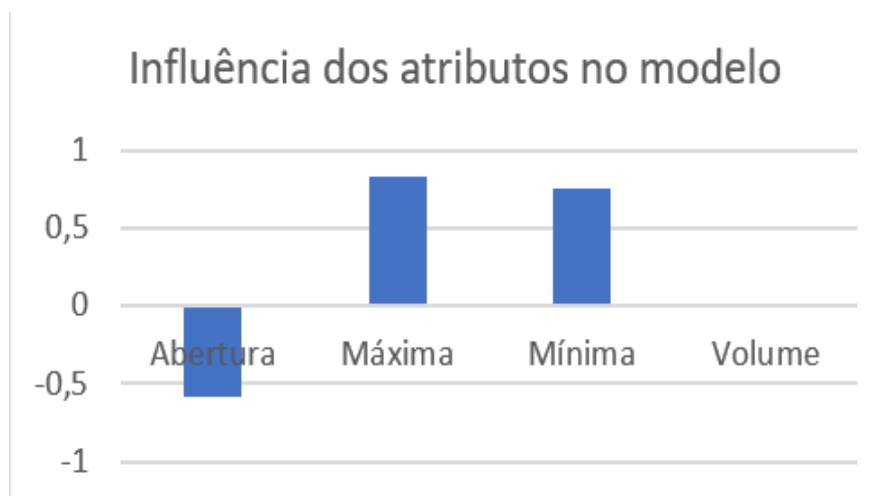


6.1.1.1 Modelo Preditivo para Fechamento

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Fechamento
Dia	Não relevante
Hora	Não relevante
Abertura	-0,5771
Máxima	0,8246
Mínima	0,7521
Volume	-0,0074

Os atributos dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos serem referências de tempo, enquanto os demais são valores reais precificados pelo mercado, caso fosse realizada a conversão dos valores de tempo para números inteiros iria distorcer o modelo tendo em vista que são métricas diferentes de medição.



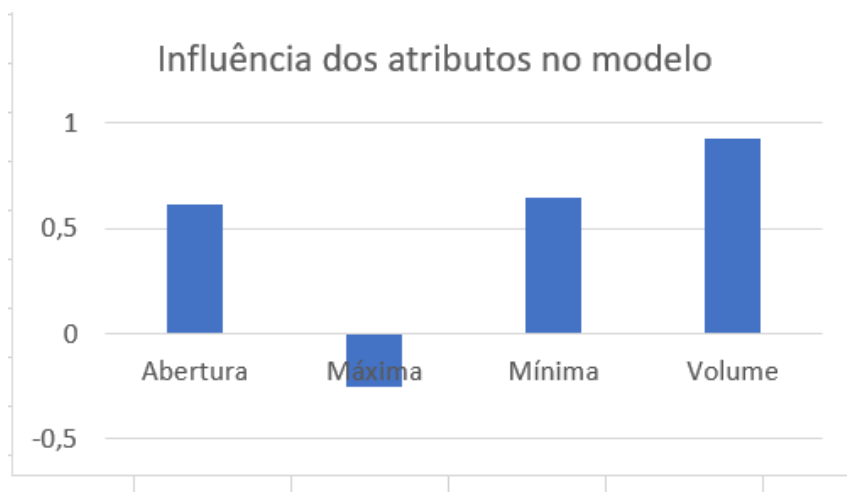
Podemos observar que a influência do volume nesse modelo preditivo é praticamente nulo, isso se dá ao fato de que no fechamento do mercado a quantidade de transações está naturalmente mais baixa, devido ao gerenciamento dos investidores da exposição ao risco, pelo receio de não completar a transação e ela ter que ser feita no dia seguinte, na qual o valor alvo do ativo não é certo.

6.1.1.2 Modelo Preditivo para Máxima

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Máxima
Dia	Não relevante
Hora	Não relevante
Abertura	0,6089
Mínima	-0,2517
Fechamento	0,6445
Volume	0,9254

Vale reiterar que os atributos dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos, se fosse realizada a conversão deles para números inteiros como o restante dos atributos iria distorcer o modelo.



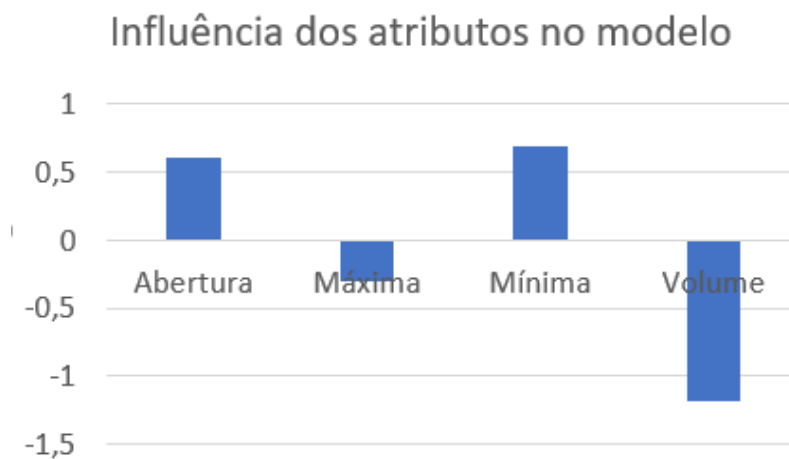
Vale ressaltar que para o modelo preditivo da máxima o volume possui uma influência positiva maior do que para os outros modelos, isso é um reflexo sobre como o mercado funciona, é um fator essencial para valores máximos o aumento do volume, se o valor se consolidar ou se houver uma correção do preço o volume de transações tende a estar mais alto devido a atenção e emoção dos investidores, tendo em vista que o risco que a volatilidade traz como influência aos ativos.

6.1.1.3 Modelo Preditivo para Mínima

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Mínima
Dia	Não relevante
Hora	Não relevante
Abertura	0,6024
Máxima	-0,2981
Fechamento	0,6935
Volume	-1,1773

Como dito acima o dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos, se fosse realizada a conversão deles para números inteiros como o restante dos atributos iria distorcer o modelo.



Vale reiterar que para o modelo preditivo da mínima o volume possui uma influência negativa maior do que para os outros modelos, isso é um reflexo sobre como o mercado funciona, sendo um fator importante a diminuição do volume quando o mercado está em queda, estabilizando naturalmente o preço do mercado e diminuindo a volatilidade.

6.1.1.4 Medidas de precisão da previsão

Neste tópico abordaremos as medidas de avaliação dos modelos obtidos para os atributos de Fechamento, Máxima e Mínima, os valores não variam muito de um modelo para outro devido a acurácia alta obtida em todos, demonstrando que os modelos apesar de variar não são significativamente diferentes, isso é explicitado também pela correlação praticamente perfeita entre os atributos da amostra de dados. Ao final das tabelas iremos fazer comentários gerais sobre o que significa cada métrica de previsão e o que o valor que ela retorna corresponde com os objetivos do trabalho

Validação do modelo para Fechamento

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0516
Erro quadrático médio	0,073
Erro relativo	1,53%
Erro quadrático relativo	1,88%
Total de instancias testadas	482

Validação do modelo para Máxima

Indicadores	Valores
Coeficiente de correlação:	0,9999
Erro absoluto médio	0,046
Erro quadrático médio	0,0669
Erro relativo	1,35%
Erro quadrático relativo	1,72%
Total de instancias testadas	482

Validação do modelo para Mínima

Indicadores	Valores
Coeficiente de correlação:	0,9999
Erro absoluto médio	0,0472
Erro quadrático médio	0,0648
Erro relativo	1,39%
Erro quadrático relativo	1,68%
Total de instancias testadas	482

Coeficiente de correlação: Esse indicador mede o grau de correlação entre as variáveis de escala métrica. O resultado muito próximo a 1 indica que os valores das variáveis do modelo são fortemente e positivamente relacionados.

Erro absoluto médio: É a medida da diferença média entre duas variáveis contínuas. O resultado próximo a 0 demonstra que o erro do modelo é extremamente baixo.

Erro quadrático médio: É uma das medidas usadas para medir a diferença entre os valores preditos por um modelo e os valores observados. O resultado próximo a 0 demonstra que o modelo possui uma acurácia alta e está muito próximo a linha ideal.

Erro relativo: É um demonstrativo do quão impreciso o modelo se torna em relação ao valor real buscado na análise, trata-se da diferença entre um valor real e um valor aproximado pode ser afetado por arredondamentos do software

Erro quadrático relativo: É usado para expressar a acurácia dos resultados numéricos, com sua utilidade ajudará na comparação dos estimadores com o parâmetro ao quadrado sua vantagem de que apresenta valores do erro nas mesmas dimensões da variável analisada

Total de instâncias testadas: 482

.....

6.1.2 Previsão dos modelos

Nessa sessão abordaremos os modelos nos quais não foram aplicados pré-processamento aplicados a dados aleatórios da base, apresentando a diferença entre o valor real e o predito, assim como seus erros percentuais.

6.1.2 Regressão Linear

A seguir apresentamos uma tabela que compila os modelos gerados e aplicá eles a um exemplo aleatório da base de dados utilizada.

Atributos	Valores	Classe predita	Diferenças	Erro
Index	55			
Data	2018.01.10			
Hora	00/01/1900			
Abertura	R\$1.634.000			
Máxima	R\$1.638.000	R\$1.634.422	R\$3.578	0,22%
Mínima	R\$1.628.000	R\$1.625.051	R\$2.949	0,18%
Fechamento	R\$1.628.000	R\$1.632.132	-R\$4.132	-0,25%
Volume	4143			

A classe predita é baseada na multiplicação do coeficiente obtido e o respectivo valor do atributo.

Fórmula:

Previsão = (Peso 1 * Valor 1)+(Peso 2 *Valor 2)+(Peso 3 *Valor 3)+Peso 4

Exemplo:

Previsão da Máxima = (Peso Mínima * Mínima da base) + (Peso Fechamento * Fechamento da base) + (Peso Abertura * Abertura da base) + Peso do volume

A diferença obtida do valor real e o valor predito é relativamente pequena, demonstrando um erro dos valores em porcentagem abaixo dos 0,30%. Essas estatísticas demonstram a qualidade do modelo gerado em retratar a realidade da amostra, apresentando níveis baixíssimos de discrepância e um alto nível de acurácia.

6.1.2.2 Multilayer Perceptron

A partir dos resultados gerados sem o pré-processamento da base e utilizando 80% para o treinamento do modelo e 20% para o teste, foi obtido como resultado os seguintes parâmetros de validação para o MLP.

Validação do modelo para Fechamento

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0548
Erro quadrático médio	0,0757
Erro relativo	1,61%
Erro quadrático médio	1,96%
Total de instancias	482

Validação do modelo para máxima

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,05
Erro quadrático médio	0,0697
Erro absoluto relativo	1,46%
Erro quadrático médio	1,80%
Total de instancias	482

Validação do modelo para Mínima

Indicadores	Valores
Coeficiente de correlação:	0,9999
Erro absoluto médio	0,0594
Erro quadrático médio	0,0797
Erro absoluto relativo	1,76%
Erro quadrático médio	2,06%
Total de instancias	482

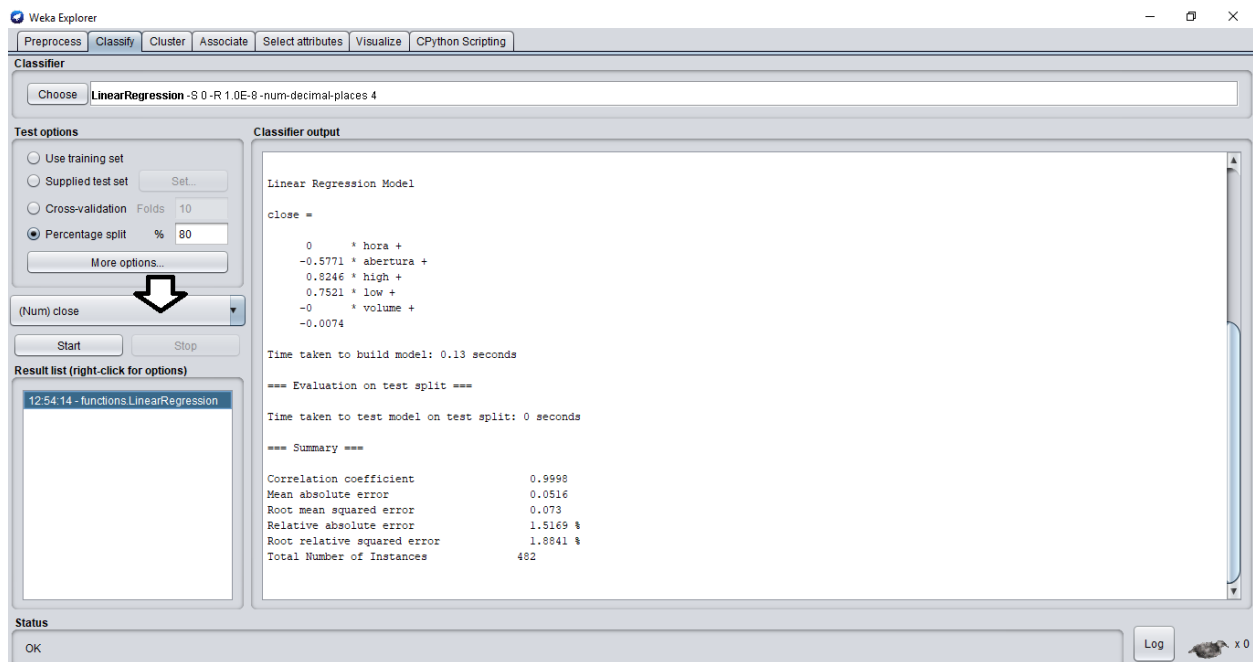
Observando as diferenças demonstradas pelos indicadores entre os valores da validação obtidos pela Regressão Linear e o uso do MLP realizamos um comparativo para se tornar mais palatável a diferença dos erros, utilizaremos o modelo de mínima de ambos os métodos, o erro relativo da Regressão Linear é de 1,39% enquanto que do MLP é de 1,76% havendo uma diferença significativa de 0,36%, sendo um demonstrativo importante perante a qualidade dos modelos.

A partir dessa análise dos indicadores podemos intuir que um modelo preditivo gerado pelo MLP, apesar de ser parecido, é relativamente pior do que o da Regressão Linear. Vale ressaltar que essa diferença do erro relativo se mantém nos outros modelos preditivos.

6.2 Execução com pré-processamento

Para a execução utilizando o pré-processamento com exclusão das tuplas que possuem o valor 0 no volume e separando para ser utilizado 80% da base no treinamento dos modelos e 20% para validação dos mesmos.

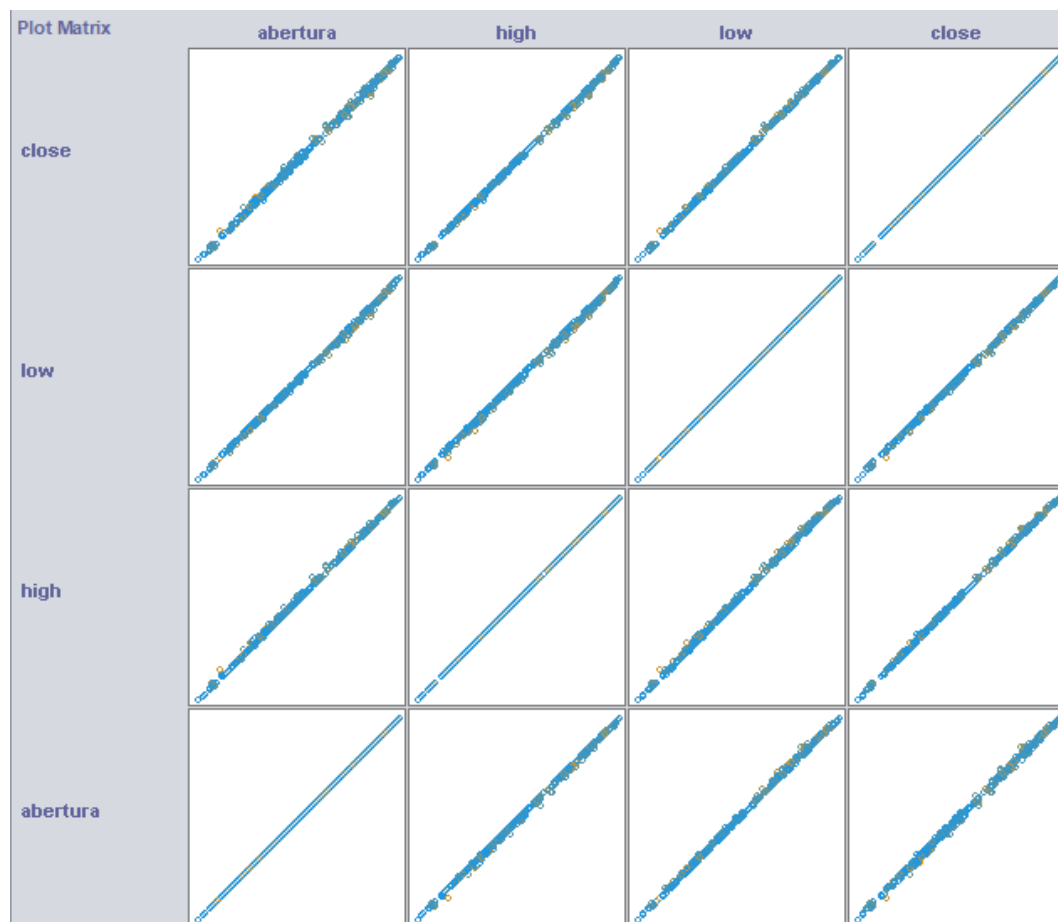
Na imagem abaixo está a tela dos resultados gerados a partir do programa WEKA para o primeiro tópico apresentado sobre os modelo preditivos obtidos, após o carregamento da base realizamos o pré processamento a partir disso temos uma base com 2389 tuplas, a partir disso é configurado na aba de classificação o algoritmo a ser aplicado na base e os parâmetros necessários para essa execução, a seta indica o atributo objetivo da predição, essa é a única configuração que é alterada para os demais modelos gerados neste tópico com a realização de limpeza da base.



6.2.1 Regressão Linear

Nessa sessão discutiremos a partir do método de regressão linear os modelos e previsões realizadas para os atributos de Fechamento, Máxima e Mínima assim respectivamente. Após a explicitação do modelo iremos analisar as medidas de precisão das previsões a fim de comparar sua eficácia em representar a realidade dos dados.

O gráfico a seguir possui pré processamento e mostra como a correlação dos atributos é alta e como os valores respeitam a linha ideal da regressão (diagonal dos gráficos) isso é perceptível pela distância dos pontos à linha principal. Comparativamente com o mesmo gráfico relativo há sessão sem pré processamento, é nítido que são muito parecidos, isso se dá pela diferença mínima de 0,0099% de uma base para a outra.

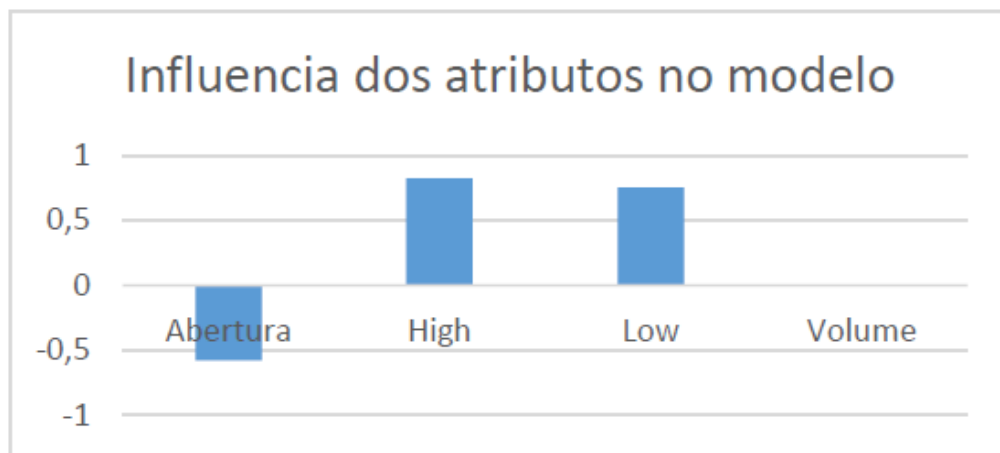


6.2.1.1 Modelo Preditivo para Fechamento

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Fechamento
Dia	Não relevante
Hora	Não relevante
Abertura	-0,5771
Máxima	0,8246
Mínima	0,7521
Volume	-0,0082

Os atributos dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos serem referências de tempo, enquanto os demais são valores reais precificados pelo mercado, caso fosse realizada a conversão dos valores de tempo para números inteiros iria distorcer o modelo tendo em vista que são métricas diferentes de medição.



Podemos observar que a influência do volume nesse modelo preditivo é praticamente nula, isso se dá ao fato de que no fechamento do mercado a quantidade de transações está naturalmente mais baixa, devido ao gerenciamento dos investidores da exposição ao risco, pelo receio de não completar a transação e ela ter que ser feita no dia seguinte, na qual o valor alvo do ativo não é certo.

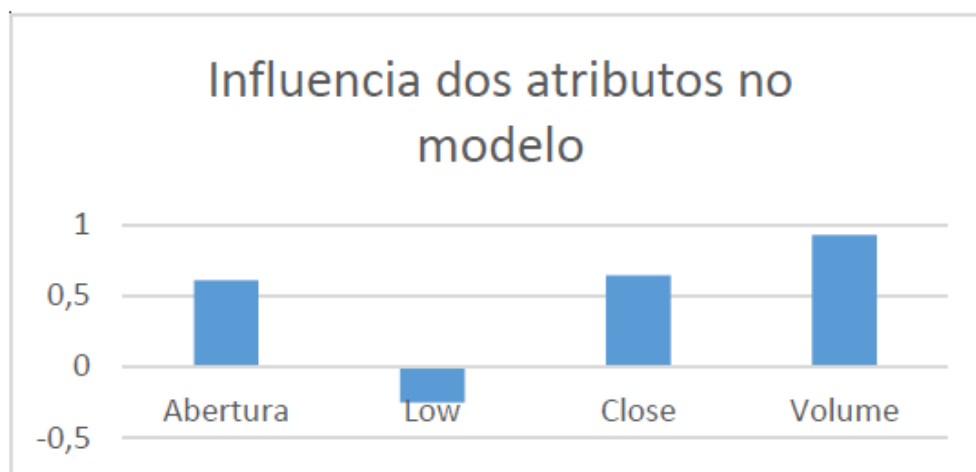
Comparativamente com o modelo respectivo sem o pré processamento a diferença entre os coeficientes obtidos nos métodos é praticamente nula, tendo em vista que a base de dados utilizada para um e para o outro difere apenas 21 linhas perante as mais .

6.2.1.2 Modelo Preditivo para Máxima

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Máxima
Dia	Não relevante
Hora	Não relevante
Abertura	0,6089
Mínima	-0,2515
Fechamento	0,6444
Volume	0,9371

Vale reiterar que os atributos dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos, se fosse realizada a conversão deles para números inteiros como o restante dos atributos iria distorcer o modelo.



Vale ressaltar que para o modelo preditivo da máxima o volume possui uma influência positiva maior do que para os outros modelos, isso é um reflexo sobre como o mercado funciona, é um fator essencial para valores máximos o aumento do volume, se o valor se consolidar ou se houver uma correção do preço o volume de transações tende a estar mais alto devido a atenção e emoção dos investidores, tendo em vista que o risco que a volatilidade traz como influência aos ativos.

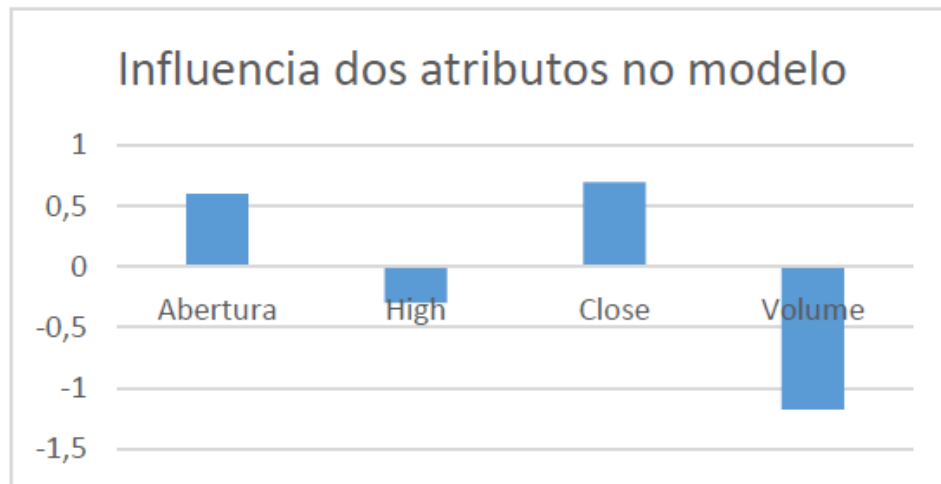
Comparativamente com o modelo respectivo sem o pré processamento a diferença entre os coeficientes obtidos nos métodos é mínima.

6.2.1.3 Modelo Preditivo para Mínima

Os coeficientes obtidos para cada atributo no modelo são os seguintes:

Atributos	Predição para Mínima
Dia	Não relevante
Hora	Não relevante
Abertura	0,6023
Máxima	-0,2978
Fechamento	0,6933
Volume	-1,1933

Como dito acima o dia e a hora foram considerados irrelevantes para a geração do modelo, devido aos tipos de dados dos atributos, se fosse realizada a conversão deles para números inteiros como o restante dos atributos iria distorcer o modelo.



Vale reiterar que para o modelo preditivo da mínima o volume possui uma influência negativa maior do que para os outros modelos, isso é um reflexo do mercado, sendo um fator importante a diminuição do volume quando o mercado está em queda, estabilizando naturalmente o preço do mercado e diminuindo a volatilidade.

Comparando com o modelo respectivo sem o pré processamento a diferença entre os coeficientes obtidos nos métodos é praticamente nula.

6.2.1.4 Medidas de precisão da previsão

Neste tópico novamente abordaremos as medidas de avaliação dos modelos obtidos para os atributos de Fechamento, Máxima e Mínima, os valores não variam muito de um modelo para outro devido a acurácia alta obtida em todos. Ao final das tabelas iremos fazer comentários gerais sobre o que significa cada métrica de previsão e o valor que ela retorna corresponde com os objetivos do trabalho

Validação do modelo para Fechamento	
Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0506
Erro quadrático médio	0,0715
Erro relativo	1,49%
Erro quadrático relativo	1,84%
Total de instancias testadas	478

Validação do modelo para Máxima

Indicadores	Valores
Coeficiente de correlação:	0,9999
Erro absoluto médio	0,0454
Erro quadrático médio	0,0625
Erro relativo	1,33%
Erro quadrático relativo	1,60%
Total de instancias testadas	478

Validação do modelo para Mínima

Indicadores	Valores
Coeficiente de correlação:	0,9999
Erro absoluto médio	0,0482
Erro quadrático médio	0,0679
Erro relativo	1,42%
Erro quadrático relativo	1,75%
Total de instancias testadas	478

Nessa sessão apontamos os indicadores relativos aos modelos que houveram pré processamento, a diferença entre os valores com e sem essa etapa foram significativamente pequenas devido a alteração de 21 linhas de toda a base de dados.

Coeficiente de correlação: Esse indicador mede o grau de correlação entre as variáveis de escala métrica. O resultado muito próximo a 1 indica que os valores das variáveis do modelo são fortemente e positivamente relacionados.

Erro absoluto médio: É a medida da diferença média entre duas variáveis contínuas. O resultado próximo a 0 demonstra que o erro do modelo é extremamente baixo.

Erro quadrático médio: É uma das medidas usadas para medir a diferença entre os valores preditos por um modelo e os valores observados. O resultado próximo a 0 demonstra que o modelo possui uma acurácia alta e está muito próximo a linha ideal.

Erro relativo: É um demonstrativo do quão impreciso o modelo se torna em relação ao valor real buscado na análise, trata-se da diferença entre um valor real e um valor aproximado pode ser afetado por arredondamentos do software

Erro quadrático relativo: É usado para expressar a acurácia dos resultados numéricos, com sua utilidade ajudará na comparação dos estimadores com o parâmetro ao quadrado sua vantagem de que apresenta valores do erro nas mesmas dimensões da variável analisada

.....

Total de instâncias testadas: 478

6.2.1 Previsão dos modelos

Nessa sessão abordaremos os modelos nos quais não foram aplicados pré-processamento aplicados a dados aleatórios da base, apresentando a diferença entre o valor real e o predito, assim como seus erros percentuais.

Nessa sessão abordaremos os modelos nos quais foram aplicados o pré-processamento de exclusão de tuplas nulas, aplicando os coeficientes em equações lineares para que seja um valor predito para os atributos de Máxima, Mínima e Fechamento, mostrando a diferença entre o valor real e o predito, assim como seus erros percentuais.

6.2.1.1 Regressão Linear

A seguir apresentamos uma tabela que compila os modelos gerados e aplicá eles a um exemplo aleatório da base de dados utilizada.

Dados	Valores	Classe predita	Diferenças	Erro
Index	55			
Data	2018.01.10			
Hora	00/01/1900			
Abertura	R\$1.634.000			
High	R\$1.638.000	R\$1.634.585	R\$3.415	0,21%
Low	R\$1.628.000	R\$1.625.053	R\$2.947	0,18%
Close	R\$1.628.000	R\$1.632.132	-R\$4.132	-0,25%
Volume	4143			

A classe predita é baseada na multiplicação do coeficiente obtido e o respectivo valor do atributo.

Fórmula:

$\text{Previsão} = (\text{Peso 1} * \text{Valor 1}) + (\text{Peso 2} * \text{Valor 2}) + (\text{Peso 3} * \text{Valor 3}) + \text{Peso 4}$

Exemplo:

$\text{Previsão da Máxima} = (\text{Peso Mínima} * \text{Mínima da base}) + (\text{Peso Fechamento} * \text{Fechamento da base}) + (\text{Peso Abertura} * \text{Abertura da base}) + \text{Peso do volume}$

Podemos observar nesse caso da previsão do valor da máxima que o erro variou 0,01%, considerado insuficiente para tirar alguma conclusão devido a porcentagem extremamente baixa para modelos que preveem valores, porém é esperado que para a maioria dos exemplos o erro do modelo com a realização do pré processamento seja maior devido ao erro demonstrando nos indicadores serem maior para.

6.2.1.2 Multilayer Perceptron

A partir dos resultados gerados sem o pré-processamento da base e utilizando 80% para o treinamento do modelo e 20% para o teste, foi obtido como resultado os seguintes parâmetros de validação para o MLP.

Validação do modelo para Fechamento

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0887
Erro quadrático médio	0,1069
Erro relativo	2,61%
Erro quadrático relativo	2,75%
Total de instancias testadas	478

Validação do modelo para Máxima

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0635
Erro quadrático médio	0,0859
Erro relativo	1,85%
Erro quadrático relativo	2,20%
Total de instancias testadas	478

Validação do modelo para Mínima

Indicadores	Valores
Coeficiente de correlação:	0,9998
Erro absoluto médio	0,0812
Erro quadrático médio	0,0948
Erro relativo	2,39%
Erro quadrático relativo	2,44%
Total de instancias testadas	478

Novamente agora com o pré processamento realizado nos dados podemos observar as diferenças demonstradas pelos indicadores entre os valores da validação obtidos pela Regressão Linear e o uso do MLP realizamos um comparativo para se tornar mais palatável a diferença dos erros, utilizaremos o modelo de mínima de ambos os métodos, o erro relativo da Regressão Linear é de 1,42% enquanto que do MLP é de 2,39% havendo uma diferença

significativa de 0,94%, sendo nesse caso 3x maior do que sem o pré processamento um demonstrativo importante perante a qualidade dos dados e dos modelos obtidos.

Esse erro maior na aplicação da base de dados com um algoritmo realizado antes da geração dos modelos se deve provavelmente a quantidade menor de tuplas utilizadas para executar a geração e validação do teste.

A partir dessa análise dos indicadores podemos intuir que um modelo preditivo gerado pelo MLP, apesar de ser parecido, é relativamente pior do que o da Regressão Linear. Vale ressaltar que essa diferença do erro relativo se mantém nos outros modelos preditivos.

A seguir um compilado geral das estatísticas de validação do modelo de previsão, os valores entre as tabelas são as diferenças dos erros.

Sem pré processamento				Com pré processamento			
Regressão linear		MLP		Regressão linear		MLP	
Validação do modelo para Fechamento		Validação do modelo para Fechamento		Validação do modelo para Fechamento		Validação do modelo para Fechamento	
Indicadores	Valores	Indicadores	Valores	Indicadores	Valores	Indicadores	Valores
Coefficiente de correlação:	0,9998	Coefficiente de correlação:	0,9998	Coefficiente de correlação:	0,9998	Coefficiente de correlação:	0,9998
Erro absoluto médio	0,0516	Erro absoluto médio	0,0548	Erro absoluto médio	0,0506	Erro absoluto médio	0,0887
Erro quadrático médio	0,073	Erro quadrático médio	0,0757	Erro quadrático médio	0,0715	Erro quadrático médio	0,1069
Erro relativo	1,53%	0,08% Erro relativo	1,61%	Erro relativo	1,49%	1,12% Erro relativo	2,61%
Erro quadrático relativo	1,88%	0,07% Erro quadrático médio	1,96%	Erro quadrático relativo	1,84%	0,91% Erro quadrático relativo	2,75%
Total de instancias testadas	482	Total de instancias	482	Total de instancias testadas	478	Total de instancias testadas	478
Validação do modelo para Máxima		Validação do modelo para máxima		Validação do modelo para Máxima		Validação do modelo para Máxima	
Indicadores	Valores	Indicadores	Valores	Indicadores	Valores	Indicadores	Valores
Coefficiente de correlação:	0,9999	Coefficiente de correlação:	0,9998	Coefficiente de correlação:	0,9999	Coefficiente de correlação:	0,9998
Erro absoluto médio	0,046	Erro absoluto médio	0,05	Erro absoluto médio	0,0454	Erro absoluto médio	0,0635
Erro quadrático médio	0,0669	Erro quadrático médio	0,0697	Erro quadrático médio	0,0625	Erro quadrático médio	0,0859
Erro relativo	1,35%	0,11% Erro absoluto relativo	1,46%	Erro relativo	1,33%	0,53% Erro relativo	1,85%
Erro quadrático relativo	1,72%	0,07% Erro quadrático médio	1,80%	Erro quadrático relativo	1,60%	0,60% Erro quadrático relativo	2,20%
Total de instancias testadas	482	Total de instancias	482	Total de instancias testadas	478	Total de instancias testadas	478
Validação do modelo para Mínima		Validação do modelo para Mínima		Validação do modelo para Mínima		Validação do modelo para Mínima	
Indicadores	Valores	Indicadores	Valores	Indicadores	Valores	Indicadores	Valores
Coefficiente de correlação:	0,9999	Coefficiente de correlação:	0,9999	Coefficiente de correlação:	0,9999	Coefficiente de correlação:	0,9998
Erro absoluto médio	0,0472	Erro absoluto médio	0,0594	Erro absoluto médio	0,0482	Erro absoluto médio	0,0812
Erro quadrático médio	0,0648	Erro quadrático médio	0,0797	Erro quadrático médio	0,0679	Erro quadrático médio	0,0948
Erro relativo	1,39%	0,36% Erro absoluto relativo	1,76%	Erro relativo	1,42%	0,97% Erro relativo	2,39%
Erro quadrático relativo	1,68%	0,38% Erro quadrático médio	2,06%	Erro quadrático relativo	1,75%	0,69% Erro quadrático relativo	2,44%
Total de instancias testadas	482	Total de instancias	482	Total de instancias testadas	478	Total de instancias testadas	478

7. Conclusão

As conclusões acerca do desenvolvimento do trabalho apontam que, para o objetivo proposto pelo grupo de aplicar a tarefa de predição, os resultados mostraram que o algoritmo de Regressão Linear apresenta um modelo que representa melhor a realidade dos dados em comparação ao MLP (Multilayer Perceptron).

Os resultados obtidos nos processos de precisão da predição mostram que a Regressão Linear obteve valores mais próximos do ideal e com um coeficiente de erros menor, em comparação ao MLP, sendo esse pior de 2 a 14 vezes do que o algoritmo de Regressão Linear. Os valores respectivos dessas diferenças são : 0,08% para regressão e 1,46% para o MLP.

Assim, podemos concluir que ambos os métodos utilizados foram de extrema utilidade, com seus resultados similares conseguimos chegar em objetivos previstos. No entanto, o Multilayer Perceptron, utilizado na plataforma Weka, não nos trouxe valores em sua saída,

assim o grupo se propôs fazer comparações entre os erros dos dois métodos abordados. Conseguindo chegar nos resultados obtidos até então.

8. Referências Bibliográficas

- [1] Cetax. **Data mining: O que é, conceito e definição.** Disponível em <<https://www.cetax.com.br/blog/data-mining/>> Acessado em: 03 de junho de 2019.
- [2] AGGARWAL, Charu C. **Neural Networks and Deep Learning: A Textbook.** Springer International Publishing AG. Edição: 1st ed. 2018.
- [3] LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley and Sons, Inc, 2005.
- [4] OLIVEIRA, R. R; CARVALHO, C. L. **Algoritmos de agrupamento e suas aplicações.** Technical report, Universidade Federal de Goiás, 2008.
- [5] CASTRO, Leandro Nunes. FERRARI, Daniel Gomes. **Introdução à mineração de dados.** Saraiva Educação 2016, 1ª Edição
- [6] FAYYAD, U; PIATETSKY-SHAPIO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.