UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JOZUA HENRIQUE SCHUSTER SCARAVONATTI

APLICAÇÃO DA TÉCNICA DE REGRESSÃO PARA ANÁLISE DE DADOS CLIMÁTICOS E PREVISÃO DE SAFRA

TRABALHO DE DIPLOMAÇÃO

MEDIANEIRA
2015

JOZUA HENRIQUE SCHUSTER SCARAVONATTI

APLICAÇÃO DA TÉCNICA DE REGRESSÃO PARA ANÁLISE DE DADOS CLIMÁTICOS E PREVISÃO DE SAFRA

Trabalho de Diplomação apresentado à disciplina de Trabalho de Diplomação, do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas – COADS – da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para obtenção do título de Tecnólogo.

Orientador: Prof. Dr. Evando Carlos Pessini

Coorientador: Prof. Dr. Arnaldo Candido Junior

MEDIANEIRA

2015

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

TERMO DE APROVAÇÃO

APLICAÇÃO DA TÉCNICA DE REGRESSÃO PARA ANÁLISE DE DADOS CLIMÁTICOS E PREVISÃO DE SAFRA

Por

JOZUA HENRIQUE SCHUSTER SCARAVONATTI

Este Trabalho de Diplomação (TD) foi apresentado às 08:20 h do dia 10 de junho de 2015 como requisito parcial para a obtenção do título de Tecnólogo no Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. Os acadêmicos foram argüidos pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Evando Carlos Pessini
UTFPR – Câmpus Medianeira
(Orientador)

Prof. Dr. Claudio Leones Bazzi
UTFPR – Câmpus Medianeira
(Convidado)

Prof. Dr. Everton Coimbra de Araújo
UTFPR – Câmpus Medianeira
(Convidado)

Prof. Me. Juliano Rodrigo Lamb
UTFPR – Câmpus Medianeira
(Responsável pelas atividades de TCC)

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

SCARAVONATTI, S. H. Jozua. Aplicação da Técnica de Regressão para Análise de Dados Climáticos e Previsão de Safra. Trabalho de conclusão de curso (Tecnologia em Análise e Desenvolvimento de Sistemas), Universidade Tecnológica Federal do Paraná. Medianeira 2014.

Este trabalho tem como o objetivo apresentar um modelo de previsão de safra com base no histórico de produção de soja e dados climáticos do estado do Paraná. O estudo foi realizado com os dados das safras de soja dos anos de 2003 a 2012 de 8 cidades do estado, dados estes coletados junto à Secretária da Agricultura e do Abastecimento do Paraná(SEAB). Os dados climáticos foram obtidos através do Instituto Nacional de Meteorologia. Foi utilizado o *software* de mineração de dados Weka, responsável por cruzar os dados e fazer a previsão de safra. Com a intenção de gerar modelos de previsão de safra foi empregado algoritmos do Weka, *Linear Regression, Pace Regression* e *LeastMedSq Regression.* São algoritmos que trabalham para a formulação desses moldes. Os dados climáticos a serem minerados foram agrupados por medias mensais, bimestrais e trimestrais a fim de verificar qual deles teriam os melhores resultados. Após a geração dos modelos de previsão, os mesmos foram validados por meio de amostras, com a intenção de encontrar um modelo que o resultado mais aproxime a estimativa alcançada e o valor real.

Palavras – chaves: Técnica de Regressão, Modelos de Previsão de Safra de Soja, Weka.

ABSTRACT

SCARAVONATTI, S. H. Jozua. Applying Regression Technique of for Climatic Data Analysis and Crop Forecast. Term Paper (Technology Analysis and Systems Development), Federal Technological University of Paraná. Medianeira 2014.

This work's objective is to present a crop forecast model based on the soy production history and the weather data for the state of Paraná. The research was made with the soy crop data of the years from 2003 to 2012 from 8 cities of the state, these data were collected together with the Secretary of Agriculture and Supply of Paraná(SEAB). The weather data were obtained through the National Institute of Meteorology. The software responsible for crossing data and making the crop forecast is Weka. Willing to generate crop forecast models some algorithms were used from Weka, *Linear Refression, Pace Regression* e *LeastMedSq Regression* are algorithms that work for these models. The weather data to be mined were grouped by monthly, bimonthly and quarterly means aiming to check which one of them would have the best result. After the forecast models were generated, they have been validated through samples, willing to find a model having the result closest to the predicted one and the real value.

Keywords: Regression technique, Soy Crop Forecast Models, Weka.

LISTA DE SIGLAS

AM	Aprendizado de Máquina
BDMEP	Banco de Dados Meteorológicos para Ensino e Pesquisa
DERAL	Departamento de Economia Rural
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto de Nacional de Meteorologia
KDD	*Knowledge Discovery in Databases*
KG	Kilos
OCDE	Cooperação e Desenvolvimento Econômico
PIB	Produto Interno Bruto
SEAB	Secretaria da Agricultura e Abastecimento
SGDB	Sistema Gerenciador de Banco de Dados
WEKA	*Waikato Environment for Knowledge Analysus*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A mineração de dados é o processo de descoberta de informações acionáveis em grandes conjuntos de dados, usando análises matemáticas para derivar padrões e tendências existentes nesses. Normalmente, esses padrões não podem ser descobertos com a exploração de dados tradicional, pelo fato das relações serem muito complexas ou por haver muitos dados.

São esses padrões e tendências que podem ser coletados e definidos como um modelo de mineração de dados. Os modelos de mineração podem ser aplicados à cenários específicos, como previsão, risco e probabilidade, recomendações, localizando sequências e agrupamentos.

De acordo com Luchini e Bianchi (2014) a produção de grãos é muito importante para o estado do Paraná, onde ocupa um lugar de destaque no cenário agrícola nacional, sendo um dos maiores produtores de soja do país. Basicamente, o clima é o recurso que mais influência na produção agrícola, tem-se em vista que investimentos em insumos e acompanhamento acabam por ser inúteis quando o clima não é favorável. Esta situação problemática até pode ser amenizada com alternativas de irrigação ou mesmo adequação de cultivares precoce e mais resistente à falta de água. Assim, verifica-se que compreender o comportamento das precipitações pluviométricas, é algo complexo, porém necessário, considerando que seguem certos padrões em cada ano.

Dentro dessa proposta, apresenta-se um projeto de pesquisa, com a finalidade de avaliar a aplicação de tarefas de mineração de dados no desenvolvimento de sistemas de previsão de safra da cultura do soja, para as cidades de Campo Mourão, Castro, Curitiba, Irati, Ivaí, Londrina, Maringá e Paranaguá do estado do Paraná. Por meio do histórico de safra combinados com o histórico de dados climáticos dos últimos 10 anos, utilizando a técnica de Data-Mining, objetivando futura estimativa de safra (Previsão).

1.1 OBJETIVO GERAL

Pesquisar um tipo de modelo matemático que seja capaz de realizar uma previsão de rendimento das culturas de soja e milho fazendo uso de dados climáticos no estado do Paraná, por meio de técnicas e ferramentas de mineração de dados.

1.2 OBJETIVOS ESPECÍFICOS

a. Desenvolver um referencial teórico sobre mineração de dados na produção de grãos (soja) e sua importância no estado do Paraná;
b. Realizar levantamento de dados climáticos e produtivos, necessários para definição do modelo a ser criado;
c. Aplicar técnicas de mineração de dados nos dados observados, com intuito de criação de um modelo que possa estimar a produção da cultura do soja no estado do Paraná;
d. O modelo matemático será validado com base em experimentos.

1.3 JUSTIFICATIVA

O Brasil apresenta índices de desenvolvimento agrícola acima da média mundial, de acordo com o estudo da Organização para Cooperação e Desenvolvimento Econômico (OCDE). O país também lidera a produtividade agrícola na América Latina (MINISTÉRIO DA AGRICULTURA, 2014).

O agronegócio representa percentual relevante do Produto Interno Bruto (PIB) brasileiro, que representa a soma de todas as riquezas produzidas no País (MINISTÉRIO DA AGRICULTURA, 2014).

Essas atividades carecem cada vez mais de altos investimentos para melhorar o retorno de produtividade, e tendo isso em vista, planejamento financeiro e previsão de clima, mercado futuro e safras, são práticas que exigem a atenção do produtor, que precisa estar diariamente informado sobre o clima e o mercado, a fim de amenizar os riscos à sua produção e melhorar a gestão agrícola (GONÇALVES, 2000).

Considerando que a chuva na quantidade certa e o calor no período correto são fundamentais para o crescimento, desenvolvimento e a produtividade da safra (MINISTÉRIO DA AGRICULTURA, 2014), o presente projeto determina a precipitação pluviométrica, insolação, umidade relativa do ar, temperatura máxima e temperatura mínima como variáveis de suma importância. Juntando essa variável com o histórico de dados de produtividade e aplicando os mesmos em uma ferramenta de mineração de dados (WEKA), os resultados obtidos (Previsão de Safra) podem auxiliar as safras futuras, melhorando a organização interna do setor agroindustrial, ou seja, englobando toda a logística da produção desde a data da colheita até a chegada do produto ao destino final.

A exemplo de utilização dessa ferramenta de mineração de dados, Boschi *et al* (2011) apresenta um estudo que comprova a eficácia do mesmo, neste constatou-se que os níveis de chuva podem representar limitação, ou aumento nos rendimentos das culturas agrícolas. Temperaturas abaixo de 0° pode acarretar em geadas que por sua vez destrói as células da planta, e na maioria das vezes, destrói as lavouras ocasionando perdas inestimáveis na produtividade da cultura.

A partir disso, o projeto em questão objetiva o melhoramento da gestão agrícola, a fim de oferecer uma ferramenta de previsão de safra, essa ferramenta vem estimular esses produtores com o intuito de registro de dados climáticos de suas propriedades.

1.4 ESTRUTURA DO TRABALHO

O trabalho foi organizado em 6 capítulos:

- O primeiro capítulo apresenta a introdução, objetivos (geral e específicos) e justificativa do seu trabalho;
- O segundo capítulo aborda conceitos de agricultura (revisão de literatura) correlacionada com a importância do clima. Apresenta o estado da arte do negócio, com foco na produção da cultura do soja;
- O terceiro capítulo aborda conceitos de Mineração de Dados e suas técnicas - apresenta o estado da arte do negócio;
- O quarto capitulo consiste na definição dos materiais e métodos da aplicação;
- O quinto capitulo corresponde a etapa de resultados e discussão;
- O sexto capítulo contém as considerações finais e propostas para trabalhos futuros.

2 REVISÃO DE LITERATURA

2.1 INTRODUÇÃO

A agricultura é um conjunto de métodos e técnicas utilizadas para o cultivo de plantas com o objetivo de gerar vários tipos de produtos (vegetais, legumes frutas, verduras) pra diversos fins, dentre eles pode-se citar a alimentação humana e animal, para gerar bebida, energia, matéria prima para obtenção de ferramentas, medicação, roupas, produtos para construção dentre outros. Desde os primórdios o homem se beneficia da agricultura e na atualidade se tornou uma atividade indispensável para a sobrevivência humana.

O clima e o solo são fatores naturais fundamentais para ter uma boa colheita, pouca e muita chuva podem prejudicar a produção, assim como variâncias climáticas. Já o solo é um recurso mineral renovável essencial, é nele que a planta se fixa, retira seus nutrientes e água para o seu desenvolvimento. Outro fator diretamente ligado a agricultura é o fator humano, responsável por preparar a terra, plantar, irrigar e adubar quando necessário e colher. Segundo Freitas (2014), quanto maior for a área de plantio maior vai ser o número de mão de obra qualificada necessária para cuidar do plantio e colheita. Do mesmo modo, quanto mais mecanizada for a propriedade, menor será a necessidade de mão de obra humana, pois com a utilização de máquinas agrícolas o plantio e a colheita são realizados de forma mais rápida e eficiente.

A agricultura moderna pode ser enquadrada em dois tipos distintos, a agricultura intensiva e extensiva. Agricultura extensiva também conhecida como agricultura familiar, é realizada de forma tradicional, é desenvolvida em pequenas propriedades, de pouca produtividade, com baixa ou nenhuma tecnologia, onde a mão de obra é composta por pessoas da mesma família. Já a agricultura intensiva, também conhecida como agricultura de precisão, é usada de forma comercial, a produção é destinada à exportação e para abastecimento interno, no qual é usada muita tecnologia avançada e pouca mão de obra, aplicada em grandes propriedades com isso consegue um bom índice de produtividade (FREITAS, 2014)

2.2 AGRICULTURA NO BRASIL

Com terras produtivas, vasta extensão territorial, investimento em tecnologia e um bom clima para a agricultura, o Brasil é um dos principais produtores e fornecedores mundiais de alimentos, sendo o quinto maior produtor agrícola do mundo. A crescente participação do país no mercado internacional é resultado da combinação desses fatores. O mercado agrícola brasileiro exporta para mais de 180 países, tendo como principais compradores a China, União Europeia e Estados Unidos, além dos países do Mercosul.

De acordo com os dados divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2014), a área de cultivo no país chegou a 56,3 milhões de hectares, os dados do levantamento mostram que o valor da produção do ano de 2003 foram de R$ 232,5 bilhões, com um aumento de 14% em relação ao ano anterior. As mais importantes culturas da produção agrícola nacional são a soja, a cana de açúcar e o milho, que juntos representaram, 59,7% do valor total da produção agrícola do país. Esses números refletem os esforços dos agricultores em aumentar a produtividade ao invés da área de cultivos, investindo cada vez mais em tecnologia e aperfeiçoamento do plantio direto.

O Ministério da Agricultura(2014), estima que o Brasil consiga a liderança mundial na safra de 2020/2021. Essa expansão deverá ser impulsionada, especialmente, pela modernização do campo, pela melhoria nas condições da propriedade familiar e pelo aumento do volume de exportações.

2.3 CLIMA E AGRICULTURA

A chuva e a temperatura na quantidade certa são fundamentais para o crescimento, desenvolvimento e a produtividade da safra. No entanto, devido sua

ampla extensão territorial, é comum que ocorra no país adversidades climáticas que podem afetar direta ou indiretamente a produção agrícola dos diversos produtos produzidos, tais como seca, vendaval, chuvas em excesso, granizo e geadas quando a temperatura do ar atinge 0°C.

Previsões Climáticas precisas e com antecedência de até 12 meses podem potencialmente permitir que os agricultores tomem decisões que possam reduzir os impactos indesejados e proveito de um tempo favorável. Para que o aumento da produtividade continue subindo, é indispensável o controle do clima no uso das lavouras.

Diante disso, passou a ser desenvolvido no Brasil estudos que permitissem indicar, com maior margem de segurança, o local e a data mais apropriada para plantar determinada cultura. Por exemplo a região Nordeste do país apresenta clima semi-árido, as principais características desse clima são altas temperaturas e baixa umidade durante o ano todo, com índices de pluviosidade bastante baixa (de 250 a 1000 mm por ano) tornando propícia a produção de cana de açúcar e algodão. Já a região Sul encontra-se o clima subtropical, com índices de umidade moderado, com pluviosidade alta (de 1200 mm a 2000 mm anuais) variando essa quantidade conforme a estação do ano, por esses motivos fica recomendado o plantio de soja e milho (EMBRAPA, 2014).

2.4 ASPECTOS BOTÂNICOS E PRODUTIVOS DA SOJA

2.4.1 Introdução

A soja (*Glycine max* (L) Merrill), é uma planta herbácea, incluída na classe *Dicotyledoneae*, ordem Rosales, família *Leguminosae*, subfamília das *Papilionoideae*, gênero *Glycine L*. (GOMES, 1990).

Proveniente do continente Asiático e conhecida como uma das plantas mais antigas do planeta, a soja não era muito utilizada para o consumo humano, sua planta

era do tipo forragem, e era destinada a alimentação do gado, cavalo e outros tipos de animais.

Foi a partir da década de 1940 que o mundo descobria a importância da soja para matérias primas que são destinadas a produção de alimentos, e a área de cultivo para produção de grãos já ultrapassava a de forragem, e essa produção começou a crescer exponencialmente até os dias de hoje onde é uma das culturas mais cultivadas e comercializadas do mundo é o que afirma a Empresa Brasileira de Pesquisa e Agropecuária (EMBRAPA, 2009).

No Brasil, a soja é a cultura que mais é aplicada no plantio, cerca de 49% da área plantada no pais. Cultivada principalmente nas regiões Centro Oeste e Sul do país. Os estados de Mato Grosso e Paraná são os que mais produzem essa cultura, juntos os estados produziram 39,44 toneladas de soja, responsáveis por 48% da safra (2012/13) da leguminosa no Brasil. Bom nível de investimento tecnológico e ciclo pluviométrico da região ser muito favorável estão entre os principais fatores pra que isso aconteça, é o que afirma o Departamento de Economia Rural do estado do Paraná (DERAL, 2013).

2.4.2 Aspectos Botânicos

A soja possui uma grande diversidade no seu ciclo, seus cultivares tem ciclos de 60 a 120 dias, no Brasil esse número varia de 100 e 160 dias. A soja pode ser classificada em vários grupos (maturação precoce, semiprecoce, médio, semitardio e tardio) e apresentar 3 tipos de crescimento (indeterminado, semideterminado e determinado) dependendo da região onde se encontra.

Suas folhas na maioria das vezes é da cor verde pálida, em outras, verde escura, seu caule pode variar de 80 a 150 cm, dependendo da quantidade de luz e chuva que a planta fica exposta. O comprimento das raízes pode chegar a até 1,80 m. Na maior parte encontrando-se a 15 cm de profundidade do solo. Suas sementes são lisas e ovais, podendo ser deparadas nas cores amarela, preta ou verde (GOMES, 1990).

O legume da soja é amenamente curvado, formado por duas valvas de um carpelo simples, medindo de 2 até 7cm, onde aloja de 1 até 5 sementes. A cor da vagem da soja varia entre amarela-palha, cinza e preta, dependendo do estágio de desenvolvimento da planta (NUNES, 2011).

2.4.3 Exigências Hídricas

A quantidade certa de água é muito importante para uma boa produção da soja, principalmente na fase de germinação da planta e floração-enchimento dos grãos. Na etapa de desenvolvimento tanto o excesso como a falta de água são prejudiciais para a boa formação da planta. A semente de soja precisa absorver no mínimo 50% de seu peso em água para ter um bom período de germinação. Nesta fase, a quantidade de água não pode passar de 85% do seu total de água disponível e não pode ser menor que 50% (EMBRAPA, 2009).

Conforme o desenvolvimento da planta cresce a quantidade de água necessária vai aumentando, atingindo o auge durante a floração-enchimento dos grãos (7 a 8 mm/dia), caindo a necessidade de água após esse período. Falta de água nessa fase, pode causar alterações fisiológicas na planta, como por exemplo o fechamento do estômato e o enrolamento da folha. Como consequência ocorre as quedas prematuras das folhas, flores e vagens, resultando na redução do número de grãos. O ciclo de água perfeito para uma boa produtividade de soja varia entre 450 a 800 mm (EMBRAPA, 2009).

2.4.4 Exigências Térmicas

A cultura da soja se adapta melhor a temperaturas do ar variando de 20°C a 30°C, sendo 25°C a temperatura ideal para um crescimento rápido e uniforme. Na

época do plantio da soja deve se evitar temperaturas muito baixas, temperaturas abaixo de 20°C pode prejudicar a germinação e emergência da planta. O desenvolvimento é pequeno ou nulo a temperaturas menores ou iguais a 10°C.

Entretanto temperaturas acima de 40°C também implicam na fase de crescimento da planta, provocando distúrbios na floração e diminuição da capacidade de retenção de vagens (EMBRAPA, 2009).

2.4.5 Aspectos de Plantio da Soja

Alguns detalhes no sistema de produção deverão ser levados em consideração, para que a lavoura se torne mais competitiva. De acordo com Nunes (2011), o cuidado com a distribuição de sementes nas fileiras, a profundidade de plantio e o espaçamento entre fileiras são fatores importantes para a obtenção da máxima qualidade de plantio.

A soja tem preferência por solos com teores de argila que vão desde 15% a 35% até teores maiores que 35%, e apresentar um bom sistema de drenagem e retenção de água para poder conceder nutrientes indispensáveis para as plantas. Mas se o solo tiver dificuldade em suprir a demanda de nutrientes para a planta, devem ser aplicados aos adubos e fertilizantes (FARIAS, 2007).

Nas épocas indicadas de semeadura, devem ser utilizados espaçamentos de 20 a 50 cm entre as fileiras. A profundidade de cada cova deve ficar a uma profundidade de 3 a 5 cm, em profundidades superiores dificulta o nascimento da planta, principalmente em solos arenosos.

A época da colheita deve ser feita quando as sementes atingirem a maturidade completa (Estádio R8), normalmente quando o grau de umidade das sementes se encontra abaixo dos 18%, durante o processo natural de secagem no campo. Por outro lado se o grau de umidade dos grãos cair muito pode prejudicar as sementes, diminuindo o peso dos grãos para a comercialização (FARIAS, 2007).

3 MINERAÇÃO DE DADOS

3.1 INTRODUÇÃO

O aumento do volume de dados gerados pelas áreas governamentais, cientificas e corporativas cresce mais a cada dia, mas apenas disponibilizar estes dados ainda não é o suficiente.

Muitos desses dados gerados, podem ter alguma relação. De modo simplório, um bom exemplo de uso do data mining pode ser empregado num mercado, onde foi descoberto que os clientes que compram o produto X teriam mais chances de acabar comprando o produto Y junto, sendo assim colocaram os dois produtos na mesma gondola, e as vendas dos produtos aumentaram. Esse caso não seria tão hipotético se não fosse o que aconteceu em uma das lojas do Wal-Mart, onde descobriu-se que o perfil do consumidor de cervejas era semelhante ao de fraldas nas sextas-feiras, usando um dos algoritmos de data mining, resultado aproximaram as prateleiras dos dois produtos e as vendas subiram 30% as sextas-feiras. Outro exemplo é a companhia MasterCard que processa mais de 30 milhões de transações por dia e utiliza Mineração de Dados para extrair vários tipos de estatísticas possíveis sobre o perfil de seus consumidores.

Para um melhor aproveitamento dessas informações foi criada a Mineração de dados, do inglês *Data Mining*, que tem como objetivo, descobrir e analisar relacionamentos entre esses dados, a fim de fornecer informações valiosas, para que se possa ser desenvolvida uma previsão de tendências futuras, baseada no passado (GONÇALVES, 2014).

A parte de análise e processo dos dados baseia-se através de várias técnicas implementadas em *software*s inteligentes, que são capazes de vasculhar esses dados oriundos de SGDB's e informar ao usuário informações que sejam consideradas como potencialmente úteis, em geral é retirado informações dos dados e transformam em algo significativo.

Segundo Albernethy (2010), empresas como Facebook, Google, Yahoo dentre outras já estão usando mineração de dados, de forma a aumentar a renda e manter-

se competitivas em relação à concorrência, e quem não estiver logo estará em desvantagem no mercado atual.

3.2 PROCESSO DE MINERAÇÃO DE DADOS

Muitas pesquisas têm sido voltadas para o desenvolvimento de técnicas com objetivo de colher informações a partir de um grande volume de dados e transformar estas informações em conhecimento útil. O problema é que na maioria das vezes estes registros representam apenas dados e não conhecimento. Buscando transformar estes dados em informação, foi desenvolvido no ano de 1989 o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases – KDD*).

O processo de KDD surgiu com o intuito de descobrir informações em dados, que se tornam necessárias para a obtenção do resultado desejado, apresentando algoritmos de data mining para extrair padrões classificados como conhecimento. Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados. Entretanto, os sistemas que empregam o método KDD precisam ser vistos como uma ferramenta interativa, não como um sistema de análise automática.

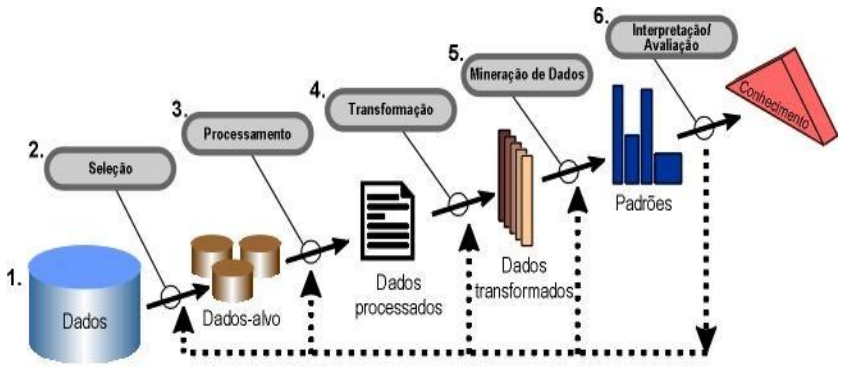


Figura 1 - Etapas correspondentes ao método KDD

Fonte: Fayyad et al.(1996)

A utilização do processo de KDD, bem como a mineração de dados, é caracterizado por várias etapas, que se tornam necessárias para a obtenção do resultado desejado. De acordo com Fayyad et al. (1996), o processo de busca de conhecimento contém uma série de passos: seleção, pré-processamento e limpeza, transformação, mineração de dados (data mining) e interpretação/avaliação. Para ter uma descrição mais detalhada sobre essas operações, é usado como exemplo a representação da Figura 1:

De acordo com Camilo et al. (2009), as representações das etapas do KDD são as seguintes:

1) Definição de Metas: Nessa primeira etapa, é feito a definição e o entendimento do problema a ser resolvido através do processo de KDD. Conhecer o tipo dos dados com o qual se irá trabalhar também é fundamental para a escolha do método mais adequado. Os dados a serem minerados podem ser divididos em dois grupos: quantitativos e qualitativos. Os dados quantitativos são representados por valores numéricos. Já os dados qualitativos contêm os valores nominais e ordinais.
2) Seleção: Corresponde a escolha do conjunto de dados apropriados para análise. As fontes fornecedoras dos dados podem ser de vários formatos e proveniente de várias fontes (planilhas, *data warehouses*, SGDB's, dentre outras);
3) Processamento: Essa etapa é fundamental, pois normalmente os dados disponíveis nem sempre se encontram na forma adequada: valores errados, dados inconsistentes, registros incompletos, sendo necessário fazer a "limpeza" dos dados, eliminação de ruídos e erros, e procedimentos para verificação da falta de dados.
4) Transformação: Essa etapa corresponde a adequação dos dados aos algoritmos, sendo que alguns deles são aplicados somente com valores numéricos e outros apenas com valores categóricos, sendo assim necessário transformar os dados conforme a necessidade. Nesta fase também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo, visando com isto melhorar o

desempenho do algoritmo de análise, dentre os principais métodos estão a suavização dos dados, normalização e agrupamento.

5) Mineração de Dados: Etapa onde é feita a aplicação dos algoritmos para descoberta de padrões nos dados. Envolve a seleção de técnicas adequadas para poder cumprir as metas.

6) Interpretação/Avaliação: Os padrões identificados pelo sistema são interpretados em conhecimento, que pode então ser utilizado para suportar a tomada de decisão humana.

O processo KDD é uma sequência de várias etapas, onde cada etapa esta correlativamente ligada a fase antecedente e a fase subsequente, sendo muito importante para se gerar um bom resultado não ignorar nenhuma etapa. Muitos autores agrupam o processo do KDD em 3 passos relativamente gerais, a etapa de pré-processamento (Definição de Metas, Seleção, Processamento e Transformação) que uma vez realizada permite que o procedimento da etapa de Mineração de Dados seja muito mais eficaz, e em consequência o resultado possa ser mostrado e interpretado pelo usuário de maneira que possa usar o conhecimento abstraído em tomadas de decisões.

3.3 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (AM) é uma subárea da Mineração de Dados, concentrada em desenvolver modelos que se possa “aprender” por meio de experiências. No aprendizado é introduzido algoritmos dedutivos que baseados em estatística que extraem regras e padrões em grandes quantidade de dados (REZENDE, 2003).

As técnicas de AM aplicam um princípio de inferência denominado indução. A indução é caracterizada pelo raciocínio originado em um conceito específico e generalizado, ou seja, adquire conhecimento a partir de um conjunto de exemplos. Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos

apresentados. Entretanto, as suposições geradas mediante a deduções indutivas podem ou não conter a verdade.

O aprendizado indutivo pode ser dividido em duas das principais categorias, sendo elas definidas como aprendizado supervisionado e não supervisionado. A Figura 2 representa a hierarquia do aprendizado indutivo.

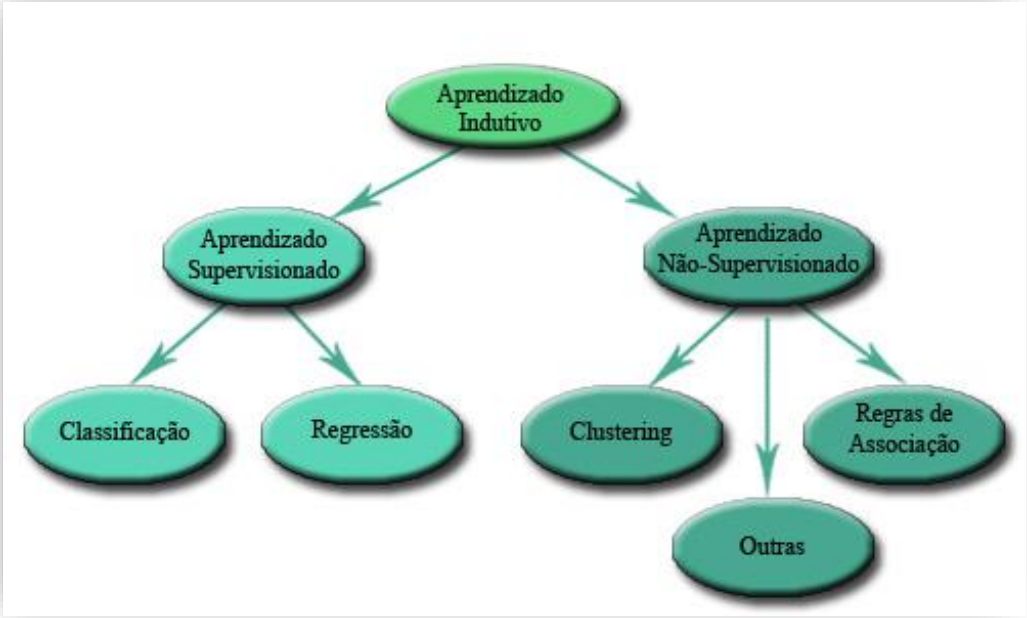


Figura 2 - Hierarquia do aprendizado indutivo

Fonte: REZENDE (2003)

O tipo de aprendizado abordado neste trabalho é o supervisionado. O objetivo do aprendizado supervisionado é construir um classificador (indutor) que possa determinar a classe de novos exemplos a partir de exemplos treinamento com classe rotulada. No próximo subcapitulo é abordado as definições de cada tipo de aprendizado (tarefas).

3.4 TAREFAS DE MINERAÇÃO DE DADOS

A técnica de mineração usada no processo de KDD está ligada a forma com que os dados foram pré-processados. Determinados algoritmos possuem restrições quanto aos tipos de variáveis envolvidas no problema. Segundo Matos (2012), o uso de vários algoritmos diferentes para executar a mesma técnica, podem gerar resultados distintos. A escolha da técnica a ser utilizada no processo de mineração de dados, depende somente do tipo de tarefa de KDD a ser efetivamente realizada. O que torna necessário distinguir o que é uma tarefa e o que é uma técnica de mineração.

A tarefa consiste na objetividade na busca dos dados, que tipo de regularidades ou conjunto de padrões pode ser interessante encontrar através do aprendizado de máquina (MATOS, 2012). Neste subcapítulo são apresentadas as principais tarefas e técnicas de aprendizado de máquina, as principais são.

- Classificação (*Classification*): Tarefa que consiste na busca por padrões que classificam elementos rotulados visando identificar a qual classe um determinado registro pertence, descobrindo algum tipo de relacionamento entre os atributos preditivos a partir de um conjunto de dados. Por exemplo, a tarefa de classificação pode ser usada para determinar quando uma transação de cartão de crédito pode ser uma fraude ou até mesmo para diagnosticar onde uma determinada doença pode estar presente.
- Associação (*Association*): Essa tarefa tem como objetivo encontrar toda/qualquer tipo de associações em que a presença de um conjunto de itens em uma transação implica na presença de outros itens nessa mesma associação. Devido a sua grande eficácia nos resultados, esse tipo de técnica é muito utilizada e solicitada pelas equipes de marketing das empresas, principalmente nas análises da "Cestas de Compras" (*Market Basket*), onde é identificado quais produtos são levados juntos pelos consumidores.
- Agrupamento (*Clustering*): Essa tarefa visa identificar e aproximar os registros similares. Um agrupamento (cluster) é uma coleção de registros idênticos entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa

difere da classificação, pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado).

Neste trabalho será empregada a técnica de regressão, devido ao seu objetivo de encontrar um modelo para predição de um atributo contínuo como função dos outros atributos. A técnica de regressão são explorada na próxima seção.

3.4.1 Regressão

A tarefa utilizada nesse trabalho é a Regressão (*Regression*). Essa tarefa consiste na procura por uma função que mapeia os registros de um banco de dados em valores reais. Essas funções podem ser lineares ou não, e se restringem a dados numéricos. Um exemplo dessa técnica é o mapeamento de informações de tempo de experiência (variável X) em ano e salário anual dos funcionários de uma empresa (variável Y). Aplicando-se o modelo de regressão linear pode-se chegar a uma função linear. Dessa maneira, baseado nos dados armazenados, obtêm-se uma função linear que pode ser utilizada para predizer o valor de uma variável em função da outra (GOLDSCHIMDT e PASSOS, 2005).

Na aplicabilidade de mineração de dados, a regressão não é utilizada precisamente para gerar um número absoluto, mas sim para criar um modelo que permite detectar padrões, prever a saída, e tirar conclusões baseadas em dados. Entretanto, não é fácil medir o desempenho de predição do modelo, pois como o valor a ser predito assume valores numéricos, não se pode afirmar se o valor predito está correto ou não. Por isso, a maioria das medidas de precisão utilizadas em problemas de regressão são baseadas na diferença entre o valor predito pelo algoritmo e o valor a ser encontrado.

4 MATERIAIS E MÉTODOS

Serão descritas as ferramentas e técnicas utilizadas para o desenvolvimento desse trabalho e efetuadas experiências em tarefas de regressão, com o objetivo de fazer comparações a fim de encontrar os melhores resultados.

4.1 *SOFTWARE* DE MINERAÇÃO DE DADOS WEKA

Com o início da mineração de dados, os primeiros *software*s de data mining começaram a ser desenvolvidos em meados da década de 1990, ainda em ambiente acadêmico. Existem várias ferramentas comerciais para data mining, desenvolvidas por empresas como SAS (Enterprise Miner), Oracle (ODM) e SPSS (Clementine) e Weka, ferramenta escolhida para o desenvolvimento desse trabalho.

O Weka (*Waikato Environment for Knowledge Analysis*) é um *software open-source* de mineração de dados desenvolvido pela Universidade de Waikato na Nova Zelândia. Escrito na linguagem Java, o Weka disponibiliza vários algoritmos para as tarefas de mineração de dados, oferecendo suporte para classificação, regressão, agrupamento, regras de associação e visualização (tabelas e gráficos). Também possui um conjunto de bibliotecas podendo ser implementado por outras aplicações.

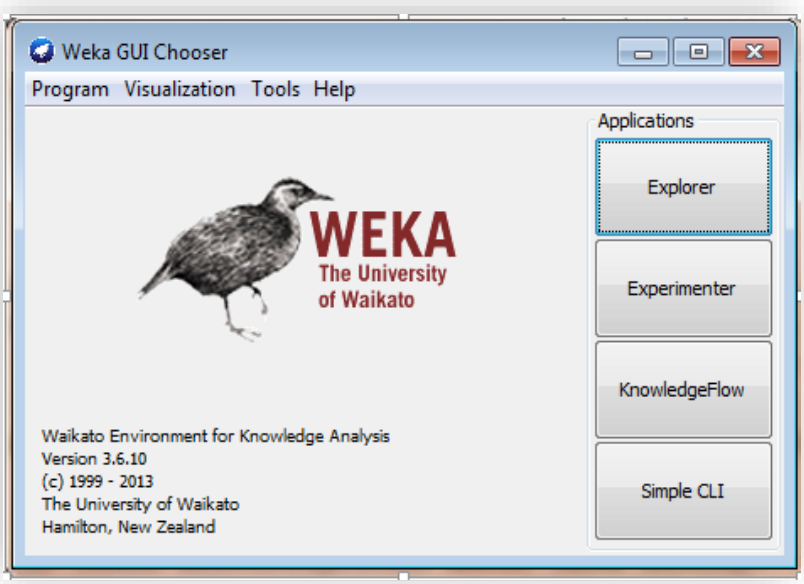


Figura 3 - Interface inicial do software de mineração de dados Weka

Fonte: Autoria Própria

A interface inicial da ferramenta Weka (Figura 3) oferece quatro opções para serem aplicadas na mineração, e de acordo com Cruz (2007), cada uma delas com suas características específicas:

- *Explorer*: É a interface a ser utilizada nesse trabalho, proporcionando ao usuário um ambiente gráfico para manipulação de dados e a utilização de diversos tipos de algoritmos de mineração de dados de forma interativa.
- *Experimenter:* Com essa opção é possível programar experimentos automatizados, podendo modificar sistematicamente os parâmetros de uma dada técnica sobre um dado conjunto de dados, podendo os resultados serem salvos em ficheiros para uma futura análise.
- *Knowledge Flow:* Possui interface gráfica e permite o desenvolvimento de projetos de Data Mining para o processamento de fluxos de dados relacionando funcionalidades do Weka de forma modular (steps).
- *Simple CLI* (*Command Line Interface*) Permite aplicar as técnicas do Weka por meio de linhas de comando.

4.1.1 Ambiente de Mineração de Dados Explorer

Para um melhor entendimento dos resultados, disponibilizando de um ambiente gráfico, é necessário optar pela opção *Explorer,* localizada na tela inicial. Ao selecionar o modo *Explorer*, é fornecida várias opções para serem utilizadas na manipulação dos dados, e são constituídas na ordem pelas abas de *Preprocess*, *Classify*, *Cluster*, *Associate*, *Selected Atributes* e *Visualize*. Para cada uma das abas da opção *Explorer* tem as suas devidas aplicabilidades e ficam ativas após a escolha de um conjunto de dados na aba do pré-processamento.

Na aba de pré-processamento (*Preprocess*), nesta tela, o WEKA permite revisar os dados com os quais será trabalhado. A divisão esquerda da janela do *Explorer* representa todas as colunas dos dados a serem utilizadas (Atributos), e o número de linhas de dados fornecidas (Instâncias). Ao selecionar cada coluna, a divisão direita da janela do *Explorer* exibe as informações sobre os dados daquela coluna com o seu conjunto de dados. Nessa janela também é feita a conexão com a base de dados e a ferramenta Weka, seleção de filtros de atributos, seleção e exclusão de instâncias de dados, e também pode ser utilizada para carregar um arquivo de dados cujo o formato é nomeado por “ARFF”, reconhecido pela ferramenta WEKA. O arquivo com a extensão ARFF é descrito na subseção 4.3. A Figura 4 apresenta a tela do Explorer da ferramenta Weka, com a aba de Processamento selecionada.

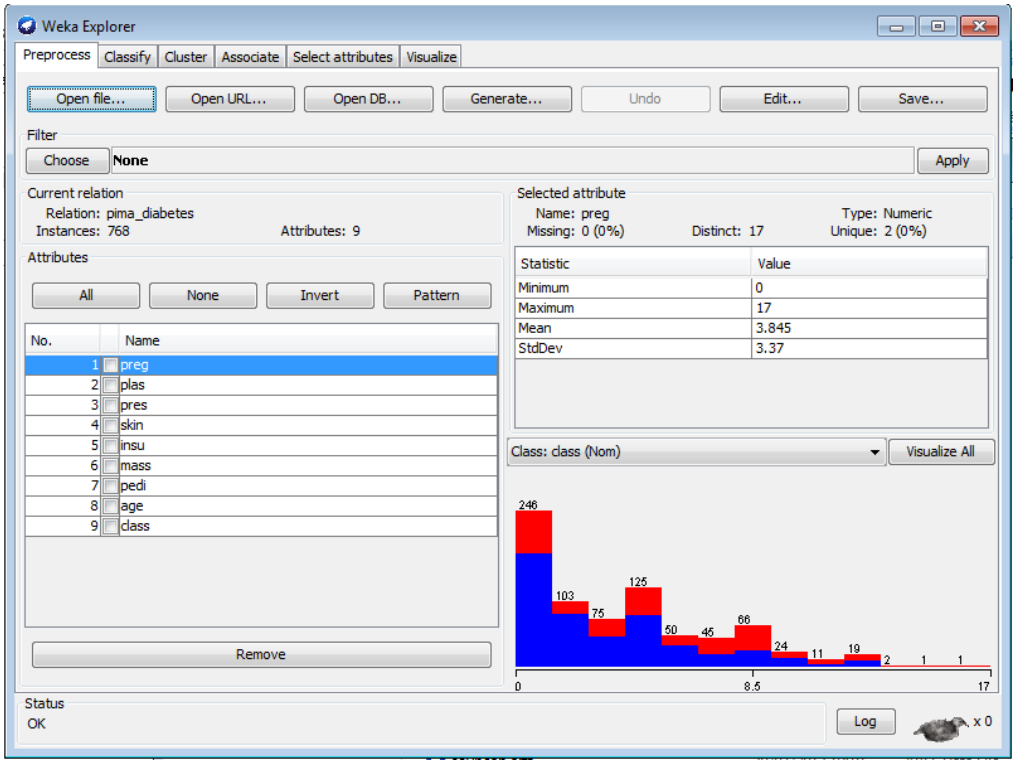


Figura 4 - Interface para utilização do modulo Explorer da ferramenta Weka

Fonte: Autoria Própria

A partir dos dados carregados na aba de pré-processamento, já é possível realizar a tarefa de mineração, com os algoritmos da aba de Classificação de Atributos.

As demais abas do modulo *Explorer* não tem contribuição para os algoritmos de regressão, mas apresentam resultados para outros tipos de tarefas. A aba d*e Cluster* (Agrupamento), permite utilizar algoritmos para encontrar dados que tenham algum tipo de semelhança. A aba de *Associate (*Associação*)* serve para descobrir regras de associação a partir de algoritmos específicos para a utilização dessa tarefa. A aba de Visualização de Resultados (*Visualize*) apresenta instâncias do arquivo ARFF, e representadas por gráficos de dispersão de duas dimensões por meio de coordenadas “x” e “y”.

E, por fim, a aba de *Select AtributesI* (Seleção de Atributos) seleciona e define uma relevância dentre os atributos selecionados.

A aba do *Classify* (Classificação dos atributos) terá maior ênfase no desenvolvimento deste trabalho, porque esta opção é destinada a fazer o comparativo dos algoritmos de regressão e é descrita na próxima seção.

4.1.2 Modo de Classificação (*Classify*)

A tela principal do modo de classificação, apresentado na Figura 5, permite que o usuário possa treinar e testar sistemas de aprendizagem (conjunto de treinamento) que classificam ou realizam uma regressão dos dados selecionados por meio do campo *Classifier* (Classificador), onde pode ser escolhido o algoritmo de regressão que será empregado.

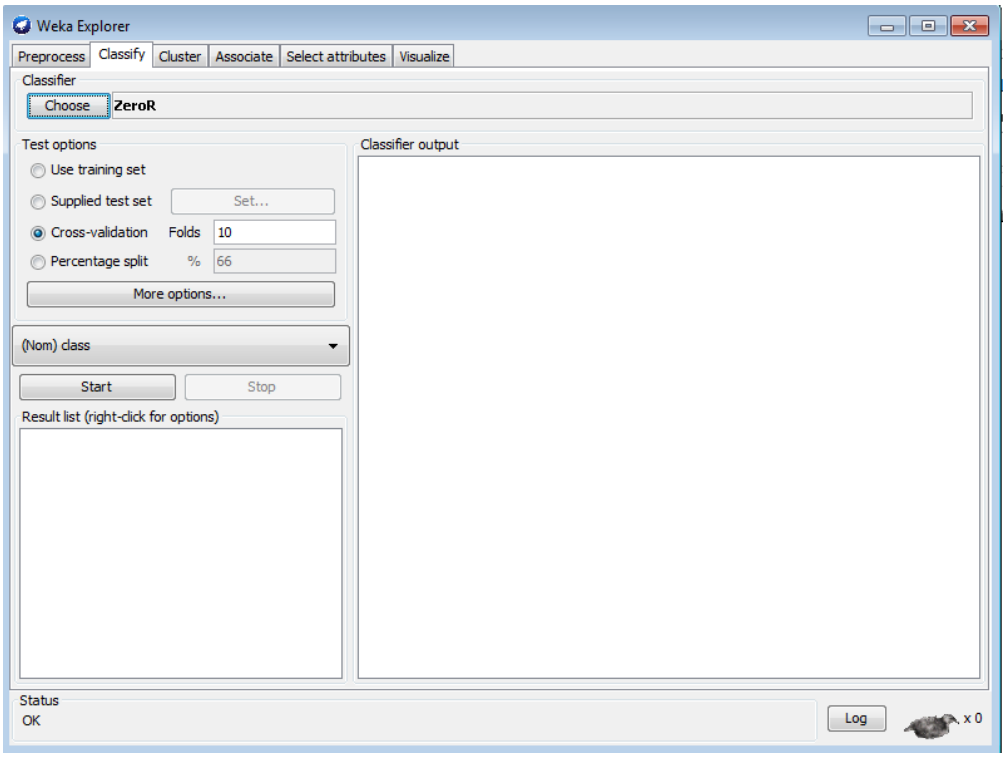


Figura 5 - Interface de Classificação da ferramenta Weka

Fonte: Autoria Própria

As opções de *Test Options* (Opções de Teste) apresenta os principais recursos de como o modelo de regressão será realizado e qual será o tipo de saída após a etapa de mineração de dados. O usuário tem várias escolhas a serem determinantes para a mineração de dados, essas alternativas são correspondentes às opções:

- *Use Training Set* (Usar Conjunto de Treinamento): esta seleção indica ao Weka que toda a base de dados marcada na aba de Processamento será utilizada no processo de mineração;

- *Supplied Test* (Teste Fornecido): seleção que permite escolher um outro conjunto de dados para ser testado e realizar avaliações em conjunto;
- *Cross-validation* (Validação Cruzada): seleção que deixa o WEKA estabelecer um modelo baseado em subconjuntos dos dados fornecidos e então calcular sua média para criar um modelo final. Basicamente a técnica de Validação Cruzada consiste em dividir a base de dados em x partes (*folds*). Destas, x-1 partes são aproveitadas para o treinamento e uma serve como base para os testes. O processo é repetido x vezes, de forma que cada parte seja usada uma vez como conjunto de testes. Ao final, a correção total é calculada pela média dos resultados adquiridos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas.
- *Percentage Split* (Divisão de porcentagens): seleção que permite a ferramenta WEKA a tomar um subconjunto percentual dos dados fornecidos para construir um modelo final.

No botão *More Options* (Mais Opções) são exibidas as configurações para a apresentação dos dados que foram minerados. Logo abaixo das opções de teste, há uma caixa (*combobox*) que permite selecionar a variável dependente (coluna a ser prevista) sendo o último passo para criar o conjunto de treinamento.

Por fim, o botão *Start* que realiza a execução da etapa de mineração de dados. O resultado da classificação pode ser visualizado no quadro *Classifier Output* (Saída do Classificador).

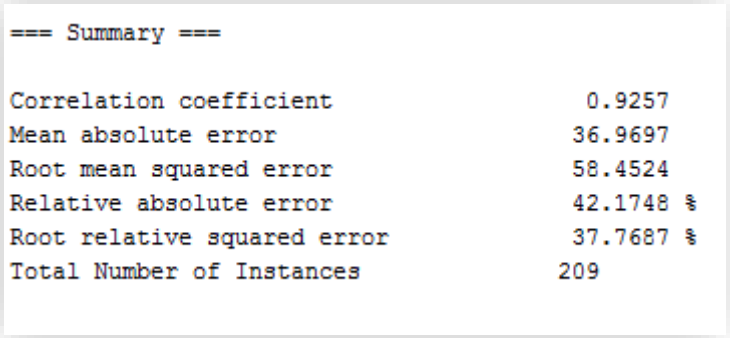
4.1.3 Medidas de Avaliação da Previsão

Em um ambiente cuja a utilização é a regressão, após o treinamento o simulador Weka traz no resultado da mineração uma série de informações, as quais

são importantes para serem interpretadas e saber o potencial de acertos do modelo que foi construído.

- *Run information* (Informações sobre a Execução): traz informações sobre a execução.
- *Classifier mode* (Informações sobre o Modelo de Classificação): traz informações sobre o classificador
- *Summary* (Sumário):traz uma lista de estatísticas, como acurácia geral do experimento de acordo com opção de validação.

As informações contidas no Sumário podem validar o modelo de regressão por intermédio de vários parâmetros estatísticos, como apresenta a Figura 6.



```
=== Summary ===

Correlation coefficient                  0.9257
Mean absolute error                     36.9697
Root mean squared error                 58.4524
Relative absolute error                 42.1748 %
Root relative squared error             37.7687 %
Total Number of Instances              209
```

Figura 6 - Parâmetros estatísticos da regressão

Fonte: Autoria Própria

A medida *Correlation Coefficient* (Coeficiente de Correlação) indica a força e a direção do relacionamento linear entre as duas ou várias variáveis a serem testadas. Seu valor varia de -1 a 1, e quanto mais próximo de 1 ou -1, maior o grau de associação entre as variáveis. Com base nos resultados dos cálculos de correlação entre atributos, concorda-se nomear os atributos fortemente correlacionados àqueles que possuíam coeficientes de correlação maiores ou iguais a 0,8.

Além do Coeficiente de correlação, o sumário também apresenta o número de instâncias calculadas e a quantidade de erros na previsão, que são calculados pelo programa e mostrados ao usuário por meio das seguintes medidas (Witen et al, 2011*)*:

- *Mean absolute error – MAE (Erro médio absoluto):* Calcula a diferença entre os valores raiz e os preditos, é a média do erro da predição.
- *Root mean squared error* - RMSE (Erro médio da raiz quadrada): Medida calculada pela média da raiz quadrada da diferença entre o valor calculado e o valor correto. É a raiz quadrada do Erro médio absoluto.
- *Relative absolute error* – ERA (Erro relativo absoluto): É a medida em porcentagem que corresponde ao erro total absoluto. Assim como o coeficiente de correlação, essa medida implica diretamente no resultado da previsão. Quanto mais alto o coeficiente de correlação, mais baixa será a média de erro absoluto, e vice e versa.
- *Root relative squared error* – RRSE (Erro relativo da raiz quadrada): Medida que reduz o quadrado do erro relativo na mesma dimensão na quantidade sendo predita, incluindo raiz quadrada. Assim como a raiz quadrada do erro significativo (RMSE), este exagera nos casos em que o erro da predição foi significativamente maior do que o erro significativo.

4.1.4 Modelos de Regressão – Funções de Saída

Após as configurações do Modo de Classificação, utilizando algoritmos de regressão (subcapitulo 4.2), a ferramenta Weka dispõe de um modelo de regressão que pode ser observado no *Classifier Output.*

O modelo é gerado e cabe ao usuário fazer as substituições dos atributos pelos valores correspondentes. Nem sempre o modelo irá apresentar todos os atributos de entrada que foram introduzidos no Weka, pois alguns algoritmos só utilizam as colunas que contribuem estatisticamente para a precisão da regressão, podendo descartar e ignorar as colunas que não ajudam a criar um bom modelo (ABERNETHY, 2010).

Para debater a metodologia do modelo de Regressão é utilizado um caso hipotético de uma imobiliária, onde se deseja prever o preço de uma casa (variável dependente) baseado em várias variáveis independentes (metragem da casa, tamanho do lote, quantidade de banheiros). O modelo é criado com base em outras casas comparáveis da região e no preço pelo qual elas foram vendidas, e então

colocando os valores de sua própria casa neste modelo, para produzir o preço esperado (ABERNETHY, 2010). A Figura 7 apresenta um modelo de regressão utilizado na ferramenta WEKA.

```
preçoVenda =
(-26.6882 * metragemCasa) +
(7.0551 * tamanhoLote) +
(43166.0767 * quantidadeQuartos) +
(42292.0901 * quantidadeBanheiros)
- 21661.1208
```

Figura 7 - Modelo de Regressão Weka

Fonte: ABERNETHY (2010)

Após gerado o modelo de regressão cabe as substituições dos atributos pelos valores independentes para ser calculado um valor estimado para ser atribuído ao preço do imóvel. A Figura 8 representa o modelo com os valores atribuídos aos atributos, ao realizar o cálculo é estimado o valor de R$ 219.328,00 ao imóvel.

```
preçoVenda =
(-26.6882 * 3198 ) +
(7.0551 * 9669 ) +
(43166.0767 * 5 ) +
(42292.0901 * 1 )
- 21661.1208
preçoVenda = R$219.328
```

Figura 8 - Modelo de Regressão Weka com atribuição de valores

Fonte: ABERNETHY (2010)

4.2 ALGORITMOS DE REGRESSÃO WEKA

Os métodos para gerar modelos de previsão na ferramenta Weka estão implementados em diversos algoritmos de regressão, podendo ser encontrados por meio do botão *Choose* (*Escolha*), no pacote *Functions* (Funções). Os algoritmos de regressão Weka aplicam o conceito de Aprendizado de Máquina e os modelos gerados mediante a esta definição são métodos indutivos, podendo ou não conter verdade.

Dos algoritmos disponíveis nesta categoria, quatro deles implementam a regressão linear múltipla, e geram modelos de predição (funções), os algoritmos são: *Simple Linear Regression*, *Linear Regression*, *Least Med Sq* e *Pace Regression* (Witen et al, 2011*)*.

Como o presente trabalho desenvolve um modelo de regressão de múltiplas variáveis, é dispensável aplicar o algoritmo *Simple Linear Regression (*Regressão Linear Simples*),* pois o mesmo algoritmo trabalha com apenas duas variáveis a serem empregadas na geração do modelo, o que pode acabar acarretando em falhas na hora de calcular a estimativa.

Dentre os algoritmos de regressão que trabalham com várias variáveis e propõem um modelo de regressão estão o *Linear Regression, Least Med Sq* e *Pace Regression.*

4.2.1 LeastMedSq Regression (*Least Med Square*)

A Regressão *Least Med Sq* ou regressão dos mínimos quadrados, é um método de regressão linear múltipla que minimiza a mediana dos erros quadrados a partir da linha de regressão. O algoritmo requer atributos de entrada e saída contínuos, e não permite que hajam valores faltantes nos atributos (Witen et al, 2011*)*.

Ao utilizar o método dos mínimos quadrados, o algoritmo possibilita um melhor ajuste de dados (relação entre variáveis), e que se obtenha a reta que melhor o representa mediante a soma dos quadrados dos desvios ou resíduos.

4.2.2 Linear Regression (Regressão Linear)

O algoritmo de regressão linear desempenha o método de regressão linear múltipla e assim como o algoritmo *Least Med Sq* implementa a técnica de mínimos quadrados e pode desempenhar opcionalmente uma seleção de atributos.

O algoritmo age como uma função, diminuindo a ordem dos seus coeficientes padronizados até que um critério de parada seja alcançado. (Witen et al, 2011*)*.

4.2.3 Pace Regression

Basicamente esse algoritmo de regressão constrói modelos de regressão lineares. Quando há muitos atributos a Regressão Pace é particularmente boa em determinar quais atributos devem ser descartados. Certamente em certas condições de regularidade é provavelmente ótima quando o número de atributos tende ao infinito (Wang and Witten, 2002).

4.3 ARQUIVO ARFF

O *software* de mineração de dados Weka, reconhece arquivos em formato ARFF, sendo o formato próprio da ferramenta. Para formatar os dados como ARFF, pode utilizar-se um simples editor de texto como bloco de notas (Notepad).

O arquivo deve ser organizado por cabeçalho e base de dados. No cabeçalho é definido o nome do *dataset,* e corresponde pela anotação `@relation`, as linhas onde irão ser declaradas os tipos de atributos a serem minerados por meio da anotação `@attribute`. Já a base de dados, é assinada pela anotação @data e consiste na lista de todas as instâncias (linhas) usadas no *dataset* com os valores dos atributos para cada instância, separados por vírgulas, cada instância é representada

em uma única linha. Os atributos precisam aparecer na ordem em que foram declarados no cabeçalho conforme é visualizado na Figura 9.

O arquivo ARFF suporta atributos numéricos e categóricos. Os atributos numéricos devem ser indicados pela palavra-chave *numeric,* quando números inteiros ou real, para números que possuem casas decimais. Os atributos categóricos precisam ser definidos por uma lista contendo todos os valores do atributo, apresentados dentro de chaves, exemplo `@attribute Clima_Tempo{chuvoso, nublado, ensolarado}`. No arquivo também podem ser empregados os tipos primitivos *date* e *string*. O arquivo ARFF também irá reconhecer comentários dentro do arquivo, sendo necessário os valores estarem entre os símbolos de porcentagem (%).

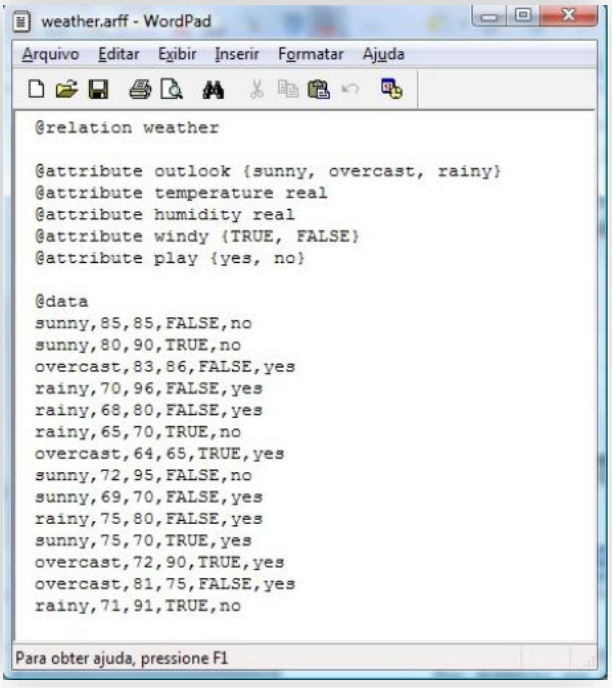


Figura 9 – Organização arquivo formato .ARFF

Fonte: WEKA (2015)

5 RESULTADO E DISCUSSÃO

Neste capítulo serão utilizadas as técnicas de mineração de dados aplicadas em histórico de dados climáticos e de produção da cultura da soja dos últimos 10 anos do estado do Paraná, a fim de desenvolver um modelo de previsão de safra. Em seguida, serão avaliados os modelos de regressão, constituirá nessa etapa uma apresentação dos resultados obtidos e desenvolvidas comparações sobre os desempenhos dos algoritmos para os diferentes tipos de testes realizados.

5.1 COLETA DOS DADOS

5.1.1 Dados Climáticos

O histórico de dados climáticos obtidos para esse trabalho foram fornecidos pelo Instituto Nacional de Meteorologia (INMET), e são encontrados por meio do domínio www.inmet.gov.br. O instituto conta com o sistema de BDMEP – Banco de Dados Meteorológicos para Ensino e Pesquisa, e abriga dados meteorológicos diários em forma digital a partir do ano de 1961 das estações que estão espalhadas pelo Brasil. Para o estado do Paraná estão disponíveis os dados das estações meteorológicas que se encontram nas cidades de Campo Mourão, Castro, Curitiba, Irati, Ivaí, Londrina, Maringá e Paranaguá. A Figura 10 representa as cidades utilizadas para a coleta de dados climáticos e produtivos.

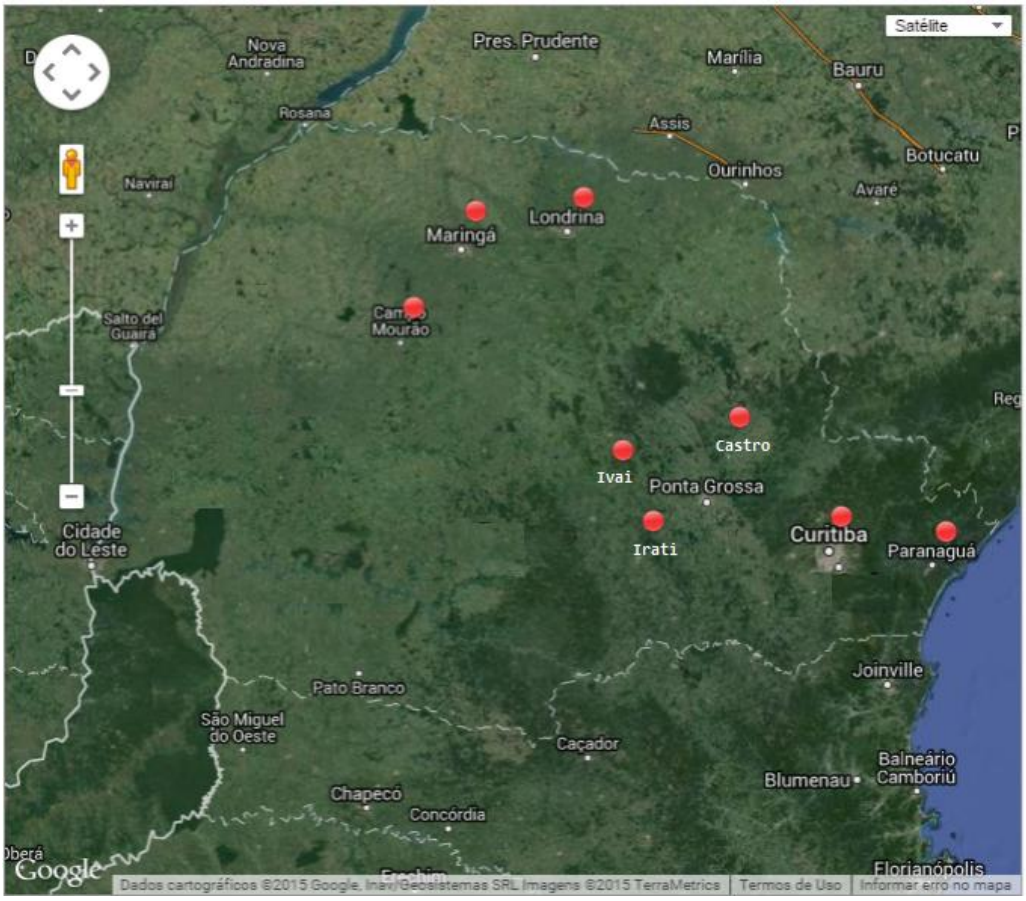


Figura 10 - Representação das cidades utilizadas para o levantamento de dados.

Fonte: INMET (2015).

Para o levantamento dos dados foi levada em consideração os tipos de dados que são importantes para o desenvolvimento das culturas do soja e do milho e que estão disponíveis no BDMEP, são elas:

- Histórico de chuva: Trata da quantidade de chuva que caiu durante os meses de um determinado ano, sua medida está colocada em milímetros
- Histórico de temperatura máxima: É a média de temperatura máxima correspondente a determinado mês, sua medida está em grau Celsius (C°).
- Histórico de temperatura mínima: É a média de temperatura mínima correspondente a determinado mês, sua medida está em grau Celsius (C°).

- Histórico de insolação: Esse histórico representa a quantidade de energia solar que atingiu determinada cidade, sua medida está em horas somadas de um determinado mês.
- Histórico de umidade relativa do ar: Corresponde a porcentagem média de umidade relativa de determinado mês.

Os dados foram coletados dos anos de 2003 a 2013, nos meses de Outubro a Março, época em que ocorre a safras de soja no estado do Paraná.

5.1.2 Dados de produção do soja

Os dados de produção das culturas de soja e de milhos foram coletados por intermédio do SEAB - Secretária de Agricultura e Abastecimento do Estado do Paraná. É necessário realizar um cadastro no site www.agricultura.pr.gov.br, após o cadastro, os registros são enviados para o e-mail do solicitante. Os dados estão em planilhas eletrônicas (Excel) e apresentam registros de várias culturas (soja, milho, arroz, trigo) dos anos 2003 a 2013, sendo organizados por núcleo regional ou município. Os dados de produção aplicados nesse trabalho foram associados com as cidades que foram coletadas os dados climáticos.

Os registros de produção são informados por meio de três variáveis, são elas:

- Área: Corresponde a área total de plantio de determinada cultura em determinada região/cidade. Sua medida está em hectare (ha).
- Produção: Corresponde a quantidade em toneladas (t) que foi colhida em determinada região/cidade.
- Produtividade: Corresponde a quantidade média de quilos (kg) colhidos de determinada cultura por hectare (Produtividade = Área / Produção).

5.2 MODELOS DE PREVISÃO DE SAFRA

Os modelos foram divididos em 3 etapas, dados climáticos aplicados por mês, bimestre e trimestre, dos meses de Outubro a Março, período respectivo as safras do soja, os quais são descritos com mais detalhes nas seções a seguir, bem como os resultados obtidos. Cada etapa procurou aperfeiçoar a anterior, com base nos melhores valores obtidos em cada uma, sendo que para cada etapa foram aplicados os algoritmos *Linear Regression*, *Pace Regression e Least Med Sq,* e para cada etapa adicionar um modelo de regressão. Em seguida, propõe-se juntar as medições realizadas 3 bases de dados, criando assim um único modelo de predição, ou seja, que apresente a maior taxa de coeficiente de correlação.

Para gerar os modelos não foram levados em consideração dados faltantes, aplicando a tarefa de Processamento do modelo KDD. Seguindo esse pensamento foi necessário descartar os registros obtidos da cidade de Paranaguá, pois a mesma não apresentava no BDMEP dados de Insolação mensal dos anos de 2004 a 2013. Também foram recusados os registros dos anos de 2003 e 2004 da cidade de Castro, pois nesse período o BDMEP não apresentava os dados climáticos respectivos dessa cidade. Então para cada modelo realizado apresentou a mesma quantia de instâncias (linhas) 69, mudando então a quantidade de atributos (colunas) de um modelo para outro.

No início dos testes foram desenvolvidos modelos de regressão com variáveis dependentes do tipo Produtividade (Produtividade = Área / Produção), Produção (total colhido) e Área do Plantio (Área correspondente ao plantio da soja), a fim de avaliar o comportamento das mesmas. Entretanto os testes do tipo Produtividade como variável independente teve como retorno índices de coeficiente de correlação abaixo de 0,6, ou seja, bem abaixo de ser um modelo confiável. Portanto foi decido trabalhar com variáveis dependentes somente do tipo Produção Total e Área do Plantio.

Quanto as configurações da ferramenta Weka, foi utilizado o mesmo padrão entre os modelos para evitar dados inconsistentes. Foi adotado o padrão Validação-Cruzada (*Cross-Validation*) e atribuído o valor 10 ao seu campo (*Folds*), isso resulta

que será testado o modelo 10 vezes, e na sequência irá retornar o resultado em base dos modelos testados.

5.2.1 Modelo 1 – Baseado em dados climáticos com médias mensais

O primeiro modelo foi construído com base em um arquivo ARFF com atributos referentes ao clima, cujos valores correspondem à media para cada mês do período de uma safra, por exemplo, quantidade média de chuva do mês de Outubro representado pelo atributo chuvaOUT, do mês de Novembro pelo atributo chuvaNOV, e assim para cada mês, de Outubro a Março para cada variável (definidas no subcapitulo 5.1.1).

A Figura 11 representa a organização dos atributos no arquivo ARFF para o primeiro modelo:

```
@relation  DadosProducaoSOJA

@attribute regiao numeric
@attribute chuvaOUT numeric @attribute chuvaNOV numeric @attribute chuvaDEZ numeric
@attribute chuvaJAN numeric @attribute chuvaFEV numeric @attribute chuvaMARC numeric
@attribute tempMaxOUT numeric @attribute tempMaxNOV numeric @attribute tempMaxDEZ numeric
@attribute tempMaxJAN numeric @attribute tempMaxFEV numeric @attribute tempMaxMARC numeric
@attribute tempMinOUT numeric @attribute tempMinNOV numeric @attribute tempMinDEZ numeric
@attribute tempMinJAN numeric @attribute tempMinFEV numeric @attribute tempMinMARC numeric
@attribute insolacaoOUT numeric @attribute insolacaoNOV numeric @attribute insolacaoDEZ numeric
@attribute insolacaoJAN numeric @attribute insolacaoFEV numeric @attribute insolacaoMARC numeric
@attribute umidadeOUT numeric @attribute umidadeNOV numeric @attribute umidadeDEZ numeric
@attribute umidadeJAN numeric @attribute umidadeFEV numeric @attribute umidadeMARC numeric
@attribute AreaPlantio numeric
@attribute ProdutividadeSOJA numeric

@data
1,119,214,187,129,123,54,28,29,29,30,30,30,15,17,18,18,17,17,242,256,239,305,277,256,70,68,78,76,83,84,581022,1629003
1,311,233,150,319,0,64,27,28,29,28,31,31,15,17,17,17,17,18,200,206,237,158,251,271,74,87,89,94,86,84,600226,1473365
```

Figura 11 - Arquivo ARFF baseado em dados climáticos com médias mensais

Fonte: Autoria Própria

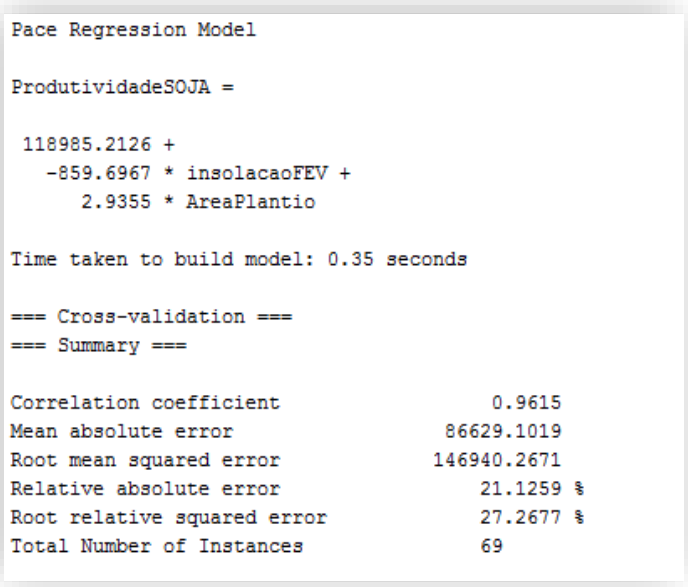
Após configurar e salvar o arquivo ARFF, foi aplicado o arquivo na ferramenta Weka, a fim de analisar o coeficiente de correlação para cada algoritmo proposto. Conforme é analisado na Tabela 1, o algoritmo de regressão Pace teve os melhores

resultados, com o maior número de coeficiente de correlação e o menor número de erros, o que torna o modelo gerado confiável.

Tabela 1- Avaliação dos algoritmos com dados de médias mensais

Modelos de Regressão 1			
	LinearRegression	***PaceRegression***	***LeastMedSqRegression***
Coeficiente de Correlação	0.8328	**0.9615**	0.6603
Erro médio Raiz Quadrada	164324.8049	**86629.1019**	167074.5415
Erro Médio Absoluto	367119.6349	**146940.2671**	478243.154
Erro Relativo Absoluto	40.0733 %	**21.1259 %**	40.7439 %
Erro Relativo Raiz Quadrada	68.1264 %	**27.2677 %**	88.7475 %

A Figura 12 apresenta o modelo de regressão do *Pace Regression*, conhecendo a base de dados, o algoritmo formulou um modelo de previsão aplicada em apenas dois atributos de substituição (*insolacaoFEV e AreaPlantio*), aparentemente um número pequeno de atributos para fazer a substituição por valores para determinar uma previsão de safra.



```
Pace Regression Model

ProdutividadeSOJA =

 118985.2126 +
   -859.6967 * insolacaoFEV +
      2.9355 * AreaPlantio

Time taken to build model: 0.35 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9615
Mean absolute error                  86629.1019
Root mean squared error             146940.2671
Relative absolute error                 21.1259 %
Root relative squared error             27.2677 %
Total Number of Instances               69
```

Figura 12 - Modelo de regressão Pace aplicado em dados climáticos com médias mensais

Fonte: Autoria Própria

5.2.2 Modelo 2 – Baseado em dados climáticos agrupados por média bimestral

No segundo modelo, o arquivo ARFF foi desenvolvido por meio dos dados referentes a 3 bimestres de cada safra, sendo eles 1° Bimestre (média dos valores de Outubro e Novembro), 2°Bimestre (média dos valores de Dezembro e Janeiro) e 3°Bimestre (média de Fevereiro e Março) da cultura do soja, sendo os mesmos avaliados individualmente. A Figura 13 representa a organização do arquivo ARFF:

```
@relation  DadosProducaoSOJA

@attribute regiao numeric
@attribute chuva1BIM numeric
@attribute chuva2BIM numeric
@attribute chuva3BIM numeric
@attribute TempMax1Bim numeric
@attribute TempMax2Bim numeric
@attribute TempMax3Bim numeric
@attribute TempMinima1Bim numeric
@attribute TempMinima2Bim numeric
@attribute TempMinima3Bim numeric
@attribute Insolacao1Bim numeric
@attribute Insolacao2Bim numeric
@attribute Insolacao3Bim numeric
@attribute Umidade1Bim numeric
@attribute Umidade2Bim numeric
@attribute Umidade3Bim numeric
@attribute AreaPlantio numeric
@attribute ProdutividadeSOJA numeric


@data
1,166,158,89,29,29,30,16,18,17,249,272,266,69,77,83,581022,1629003
1,272,235,32,27,29,31,16,17,18,203,198,261,81,91,85,600226,1473365
1,221,97,167,28,30,30,17,18,17,207,248,207,85,79,81,573240,1472966
1,124,211,190,29,29,30,18,20,19,225,185,219,78,83,82,569144,1759363
1,168,142,126,30,30,30,16,18,17,213,231,228,83,88,88,568024,1742947
1,117,153,118,29,30,30,15,17,17,219,252,219,78,80,81,560945,1307864
```

Figura 13 - Representação do arquivo .ARFF organizado com valores bimestrais

Fonte: Autoria Própria

Após configurar e salvar o arquivo ARFF, foi aplicado o arquivo no Weka, a fim de analisar o coeficiente de correlação para cada algoritmo proposto, os dados gerados podem ser visualizados na Tabela 2.

Tabela 2 - Avaliação dos algoritmos com dados de médias bimestrais

Modelos de Regressão 2			
	LinearRegression	***PaceRegression***	***LeastMedSqRegression***
Coeficiente de Correlação	0.9268	**0.982**	0.9571
Erro médio Raiz Quadrada	110840.0823	**74100.2142**	91327.5477
Erro Médio Absoluto	214639.2746	**100472.5461**	154206.2672
Erro Relativo Absoluto	27.1674 %	**18.1623 %**	22.3848 %
Erro Relativo Raiz Quadrada	40.0577 %	**18.751 %**	28.7792 %

De acordo com os dados da Tabela 2, o algoritmo de *Pace Regression* teve novamente melhores resultados, comparando com outros dois algoritmos. O algoritmo teve o maior número de coeficiente de correlação e por sua vez o menor número de erros.

A Figura 14 representa o modelo de regressão do *Pace Regression*, sendo a fórmula para se gerar um modelo de previsão contém 3 atributos de substituição (*chuva3BIM, Umidade2Bim e AreaPlantio*). Comparando esse mesmo modelo com o modelo anterior, esse teve melhores resultados, não somente pela taxa de coeficiente de correlação, e uma menor quantidade de erros, mas também o tempo de execução do algoritmo para ser gerado a função, 0.08 segundos contra 0.35 segundos do primeiro modelo:

```
Pace Regression Model

ProdutividadeSOJA =

-841388.5183 +
    746.8938 * chuva3BIM +
   9346.6215 * Umidade2Bim +
      2.7816 * AreaPlantio

Time taken to build model: 0.08 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.982
Mean absolute error                  74100.2142
Root mean squared error             100472.5461
Relative absolute error                 18.1623 %
Root relative squared error             18.751  %
Total Number of Instances               69
```

Figura 14 - Modelo de regressão Pace aplicados em médias bimestrais

Fonte: Autoria Própria

5.2.3 Modelo 3 – Baseado em dados climáticos agrupados por média trimestral

No terceiro modelo, o arquivo ARFF foi desenvolvido por meio dos dados referentes a 2 trimestres de cada safra, sendo eles 1° Trimestre (média dos valores de Outubro, Novembro e Dezembro), 2°Trimestre (média dos valores de Janeiro, Fevereiro e Março) da cultura do soja. A Figura 15 representa a organização do arquivo ARFF:

```
@relation  DadosProducaoSoja

@attribute regiao numeric
@attribute chuva1TRI numeric
@attribute chuva2TRI numeric
@attribute tempMax1TRI numeric
@attribute tempMax2TRI numeric
@attribute tempMin1TRI numeric
@attribute tempMin2TRI numeric
@attribute insolacao1TRI numeric
@attribute insolacao2TRI numeric
@attribute umidade1TRI numeric
@attribute umidade2TRI numeric
@attribute AreaPlantio numeric
@attribute ProdutividadeSOJA numeric


@data
1,173,102,29,30,17,17,246,279,72,81,581022,1629003
1,231,128,28,30,16,18,214,227,83,88,600226,1473365
1,164,159,29,30,17,17,228,213,83,80,573240,1472966
1,146,204,30,30,18,19,220,199,78,83,569144,1759363
1,156,136,30,29,17,17,226,221,84,88,568024,1742947
1,111,147,30,28,16,17,248,212,77,83,560945,1307864
1,206,208,29,30,19,19,205,191,85,91,570010,1886760
1,222,178,28,29,16,19,217,173,80,81,576447,1958305
```

Figura 15 - Representação do arquivo .ARFF organizado com valores trimestrais

Fonte: Autoria Própria

Configurando o arquivo ARFF com os dados trimestrais, foi aplicado o arquivo no Weka, a fim de analisar o coeficiente de correlação para cada algoritmo proposto, os dados gerados podem ser visualizados na Tabela 3.

Tabela 3 – Avaliação dos algoritmos com dados de médias trimestrais

Modelos de Regressão 3			
	LinearRegression	***PaceRegression***	***LeastMedSqRegression***
Coeficiente de Correlação	0.9569	0.8167	**0.9622**
Erro médio Raiz Quadrada	100775.1334	152228.6191	**91592.2442**
Erro Médio Absoluto	155349.9094	380229.5084	**146116.6954**
Erro Relativo Absoluto	24.5867 %	37.1401 %	**22.3463 %**
Erro Relativo Raiz Quadrada	28.8405 %	70.5892 %	**27.1264 %**

Ao observar os dados da Tabela 3, o algoritmo *Least Med Sq* acabou superando os outros dois algoritmos, ele obteve os melhores resultados. Ao contrário do que foi observado nos dois primeiros modelos, o algoritmo de *Pace Regression* teve os índices mais baixos, fazendo a entender que o mesmo não trabalha bem com uma quantidade de dados reduzida.

A Figura 16 representa o modelo de regressão do *Least Med Sq*, sendo a fórmula para se gerar um modelo de previsão com vários atributos de substituição.

```
ProdutividadeSOJA =

    366.729  * chuva1TRI +
    218.7723 * chuva2TRI +
  -3647.9651 * tempMax1TRI +
  11547.1434 * tempMax2TRI +
  12694.9356 * tempMin1TRI +
 -29971.5215 * tempMin2TRI +
    921.1964 * insolacao1TRI +
  -1219.8659 * insolacao2TRI +
   8991.2563 * umidade1TRI +
  -4670.3979 * umidade2TRI +
      3.1629 * AreaPlantio +
-281068.2549

Time taken to build model: 0.38 seconds
```

Figura 16 - Modelo de regressão Pace aplicados em médias trimestrais

Fonte: Autoria Própria

5.3 VALIDAÇÃO DOS MODELOS DE PREVISÃO DE SAFRA

Após gerar os modelos foi realizado as validações dos mesmos por meio de amostras, que após geradas as funções foi comparado o resultado real com o resultado indutivo de cada modelo.

As amostras foram realizadas sobre 2 etapas, sendo estabelecidos padrões para inserção de instâncias cujo os valores são reais, a fim de testar a eficácia dos modelos. Fazem parte do padrão de inserção a instância mediana e a instância que contém o desvio padrão de cada modelo. A primeira amostra foi realizada com base no desvio padrão de cada atributo de cada conjunto de dados. A segunda amostra foi efetivada pela instância mediana de cada conjunto de dados (instância que ocupa a posição central do modelo).

Os resultados das amostras serão apresentadas por tabelas, elas conterão os valores das variáveis de substituição de cada modelo, e colunas em comum, ValorEstimado, ValorReal e Diferença. A coluna ValorEstimado representa os valores resultantes da fórmula do modelo sobre cada instância, a coluna ValorReal representa os valores reais, quanto mais próximo a estimativa chegou perto dos valores dessa coluna mais confiável será o modelo, ambas colunas correspondem os valores em toneladas. E por fim a coluna Diferença representa a porcentagem da diferença entre os valores estimados pelos valores reais. Quanto menor for os valores da coluna Diferença, mais alta será a confiança sobre o modelo de previsão.

Formaram os conjuntos de dados de cada arquivo ARFF, respectivamente a cada modelo utilizados no subcapitulo anterior.

5.3.1 Amostra 1

A primeira validação refere-se ao modelo de regressão 1, gerado pelo algoritmo *Pace Regression*. Para esta validação, empregou-se um conjunto de dados meteorológicos cujos valores foram agrupados pelas médias mensais.

A Tabela 4 representa os dados que foram aplicados nas variáveis de substituição insolacaoFEV (Média de Insolação do Mês de Fevereiro) e AreaPlantio (Área que corresponde ao espaço do plantio), valores utilizados na Formula 1:

$$ProdutividadeSoja = 118985.2126 + (-859.6967 * \text{insolacaoFEV}) + (2.9355 * \text{AreaPlantio}). \quad \textbf{(1)}$$

Tabela 4 – Dados Amostra 1

	InsolaçãoFev.	ÁreaPlantio	ValorReal	ValorEstimado	Diferença (%)
Desvio Padrão	51	183518	529402	613858	13,76
Mediana	165	73180	193440	191764	0,87

Segundo a Figura 17, ao analisar a coluna Diferença, é possível ver a que a divergência entre os valores reais e a estimativa não foi tão alta, sendo 13,76 % para a instância que continha os valores do desvio padrão, e 0,87% para mediana, nos dois casos a predição se aproximou muito do valor real. O modelo teve um bom desempenho com os dados indutivos e apresentou uma boa aceitação sobre os valores das instâncias, considerando que o coeficiente de correlação não chegou a 100%.

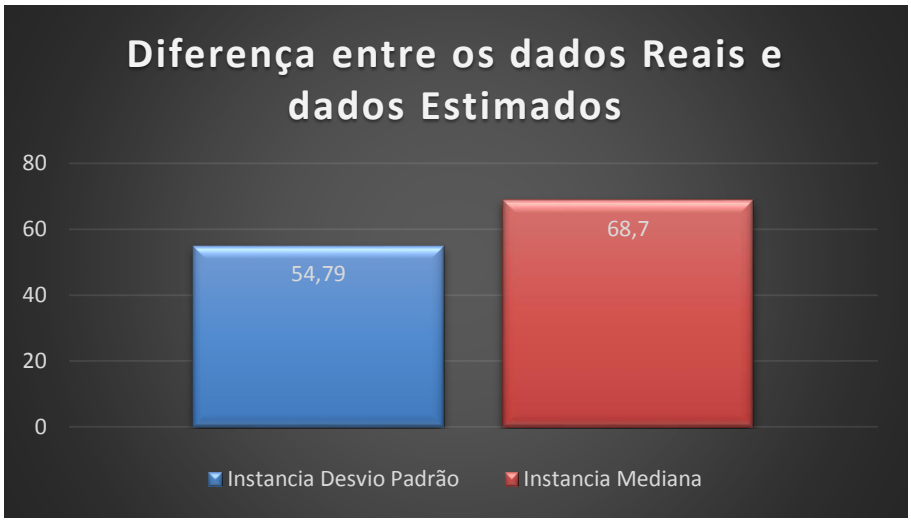


Figura 17 - Diferença entre os dados reais e dados estimados do conjunto de média mensal

Fonte: Autoria Própria

5.3.2 Amostra 2

A segunda etapa da validação foi correspondente ao modelo de regressão 2, também gerado pelo algoritmo *Pace Regression.* O modelo foi testado sobre as instâncias de média e mediana, do conjunto de dados com registros de médias bimestrais. A Tabela 5 apresenta os dados que foram aplicados nas variáveis de substituição Chuva3Bim (Média de chuva do 3 Bimestre), Umidade2Bim (Media de umidade do Segundo Bimestre) e AreaPlantio (Área que corresponde ao espaço do plantio), valores utilizados na Formula 2:

$$\text{ProdutividadeSOJA} = (-841388.5183) + (746.8938 * \text{Chuva3Bim}) + (9346.6215 * \text{Umidade2Bim}) + (2.7816 * \text{AreaPlantio}) \quad \textbf{(2)}$$

Tabela 5 - Dados Amostra 2

	Chuva 3° Bim.	Umidade 2° Bim.	Área Plantio	Valor Real	Valor Estimado	Diferença (%)
Desvio Padrão	61	5	183518	527809	238621	54,79
Mediana	108	91	76500	219225	699614	68,7

Levando em consideração que o modelo de regressão 2, aplicando o algoritmo *Pace Regression* teve o maior índice de coeficiente de correlação dentre todos os modelos, correlação de 0,98, ou seja uma correlação muito forte. O modelo não soube corresponder a expectativa e fracassou ao realizar estimativa sobre os dados bimestrais, o que torna o mesmo não muito confiável para ser aplicado quando o objetivo é prever uma safra.

Após os cálculos do modelo, os dados estimados ficaram longe dos valores reais, para instância do desvio padrão, a diferença correspondeu a 54,79%, e para instância mediana a diferença foi ainda maior, 68,7% de divergência sobre os valores reais. A Figura 18 representa a diferença dentre as instâncias do desvio padrão e mediana do conjunto de dados da média bimestral.

Figura 18 - Diferença entre os dados reais e dados estimados do conjunto de média bimestral

Fonte: Autoria Própria

5.3.3 Amostra 3

A última etapa da validação foi correspondente ao modelo de regressão 3, gerado pelo algoritmo *Least Med Sq.* O modelo foi testado sobre as instâncias de média e mediana, do conjunto de dados com registros de médias trimestrais. A Tabela 6 apresenta os dados que foram aplicados nas variáveis de substituição C1 (Média de Chuva do primeiro Trimestre), C2 (Média de Chuva do segundo Trimestre), X1 (Média de Temperatura Máxima do primeiro Trimestre), X2 (Média de Temperatura Máxima do segundo Trimestre), M1 (Média de Temperatura Mínima do primeiro Trimestre) , M2 (Média de Temperatura Mínima do segundo Trimestre) ,I1 (Média de Insolação do primeiro Trimestre), I2 (Média de Insolação do segundo Trimestre), U1 (Média de Umidade do primeiro Trimestre), U2 (Média de Umidade do segundo Trimestre), e AreaPlantio (Área que corresponde ao espaço do plantio), valores utilizados na Formula 1 :

$$\begin{aligned} \text{ProdutividadeSOJA} = {} & (366.729 * \text{C1}) + (218.7723 * \text{C2}) + \\ & (-3647.9651 * \text{X1}) + (11547.1434 * \text{X2}) + (12694.9356 * \text{M1}) + \\ & (-29971.5215 * \text{M2}) + (921.1964 * \text{I1}) + (-1219.8659 * \text{I2}) + \end{aligned} \quad \textbf{(3)}$$

$(8991.2563 * U1) + (-4670.3979 * U2) + (3.1629 * AreaPlantio) + (-281068.2549)$.

Tabela 6 – Dados Amostra 3

	C1	C2	X1	X2	M 1	M 2	I1	I2	U1	U2	AP	VR	VE	D(%)
Desvio Padrão	39	50	8	21	16	18	185	186	77	80	98895	292625	257057	12,15
Mediana	108	154	26	28	16	17	154	175	90	91	76500	219225	268862	18,5

Em conformidade com os dados apresentados na coluna diferença (D%), o modelo foi baseado no algoritmo *Least Med Sq,* e apresentou para a instância do desvio padrão uma diferença de 12,15% entre o valor real e valor estimado, já a instância mediana a diferença foi de 18,5%, conforme é analisado na Figura 19.

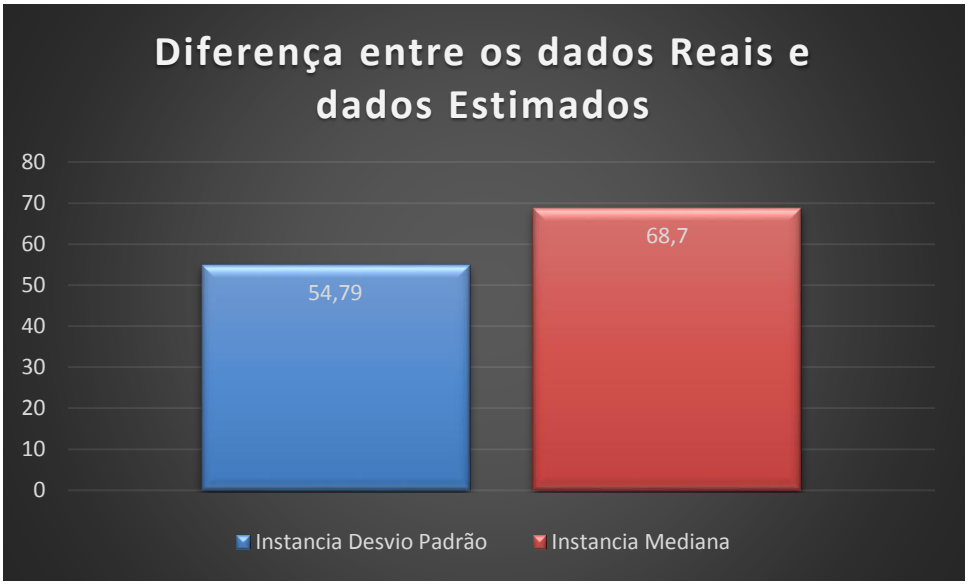


Figura 19 - Diferença entre os dados reais e dados estimados do conjunto de média trimestral

Fonte: Autoria Própria

Se analisado o coeficiente de correlação apresentado nesse modelo, 0,95 de correlação, o mesmo não teve uma boa regularidade sobre os valores das instâncias e se mostrou instável na hora de gerar o resultado. Adotando como referência os dados climáticos de várias regiões distintas o modelo de regressão 3 não teria um bom desempenho, podendo apresentar valores aproximados e valores muito além deles, tornando o modelo não muito confiante e pouco aceitável.

6 CONSIDERAÇÕES FINAIS

6.1 CONCLUSÃO

O principal objetivo desse trabalho foi apresentar alguns métodos de mineração de dados para estimativa de safra usando os métodos de regressão da ferramenta Weka *Linear Regression*, *Pace Regression* e *Least Med Sq Regression*, algoritmos que geram um modelo de saída.

Com o objetivo de verificar a precisão dos diversos métodos de regressão apresentados, foram realizados 3 experimentos (amostras) para calcular a precisão desses métodos na predição de safra sobre os conjuntos de dados agrupados por médias mensais, bimestrais e trimestrais. O algoritmo *Pace Regression* obteve os melhores resultados para os conjuntos de dados mensais e bimestrais e o algoritmo *Least Med Sq Regression* obteve o melhor resultado para o conjunto de dados trimestrais.

Por meio dos experimentos realizados, constatou-se que ambos algoritmos tiveram um alto índice de correlação dentre os atributos. Entretanto apenas o algoritmo *Least Med Sq Regression* que trabalhou com dados de médias mensais gerou um modelo confiável, insto é, sem muita diferença entre a indução e os valores reais.

Por fim, é importante enfatizar que os experimentos não levaram em consideração outros fatores além da precisão, como por exemplo, a discrepância dos dados e percas produtivas das safras, o que pode ter influenciado no desenvolvimento das amostras.

6.2 TRABALHOS FUTUROS/CONTINUAÇÃO DO TRABALHO

O desenvolvimento da agricultura aliada com a tecnologia está cada dia mais importante. Um modelo de previsão de safra poderia auxiliar de várias maneiras, muitas famílias que tem como renda a agricultura, a ter uma perspectiva com a

variância do clima sobre a produtividade do soja. Baseado nesses estudos pode se propor duas propostas para a continuação desse trabalho.

O presente trabalho efetuou 3 modelos de previsão de safra, e somente um modelo apresentou um desempenho satisfatório, tendo em ideia esse raciocínio, a primeira proposta para um trabalho futuro corresponderá a testes de transformação dos dados climáticos e produtivos, tendo em consideração que os dados estão em escalas de medidas distintas. Por exemplo, os dados de precipitação pluviométrica foram coletados com a medida em milímetros, já os dados de insolação estão representados por hora mensal. Outro caso é quando as medidas climáticas apresentam uma diferença muito alta de um mês para o outro, por exemplo, se em um mês choveu 46 milímetros no mês seguinte choveu 250 milímetros.

Uma aplicação de modificação dos dados é a transformação logarítmica. A transformação logarítmica aumenta a distância entre os valores pequenos e reduz as distancias entre os valores grandes, o que tornaria os dados das medidas climáticas mais simétricos a serem trabalhados por estarem numa mesma escala de medida. Por sugestão, poderia ser realizados testes com os dados transformados e dados não transformados, afim de descobrir modelos de previsão de safra com os melhores resultados.

Após obter um modelo confiante que pudesse prever uma safra com exatidão, uma segunda proposta para trabalho futuro seria um desenvolvimento de uma interface voltada para usuário, com modelos de previsão implementados e campos de inserção de dados climáticos tendo como saída os valores de uma previsão de safra. Tendo como base uma aplicação em que o próprio produtor rural poderia interagir aplicando os dados coletados da sua região.

7 REFERÊNCIAS

ABERNETHY, Michael. Mineração de dados com Weka – Introdução e Regressão. Disponível em <<http://www.ibm.com/developerworks/br/opensource/library/os-weka1>>. Acesso em 11 JAN 2015.

BOSCHI, Raquel. OLIVEIRA, Stanley. ASSAD, Eduardo. *Data Mining techniques for decennial analysis of rainfall* in Rio Grande do Sul. Disponível em << http://www.scielo.br/scielo.php?script=sci_arttext&pid=S010069162011000600016>> Acesso em 31 AGO 2014;

DERAL. Soja – Analise da Conjuntura Agropecuária. Disponível em <<http://www.agricultura.pr.gov.br/arquivos/File/deral/Prognosticos/soja__2013_14>> Acesso em 9 SET 2014.

DOSUALDO, Daniel. REZENDE, Solange. Análise da Precisão de Métodos de Regressão. Disponível em << http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_197 >> Acesso em 2 SET 2014.

EMBRAPA. Tecnologias de Produção de Soja. Disponível em <<http://www.cnpso.embrapa.br/producaosoja/index.htm>>. Acesso em 2 SET 2014.

FREITAS, Eduardo. As definições de agricultura. Disponível em <<http://www.brasilescola.com/geografia/agricultura-5.htm>> Acesso em 31 AGO 2014;

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data mining: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. 261 p.

GONÇALVES, Anderson. Os principais produtos da agricultura paranaense. Disponível<<http://www.gazetadopovo.com.br/vidaecidadania/especiais/retratosparana/curiosidades/os-principais-produtos-da-agricultura-paranaense>> Acesso 2 SET 2014.

GONÇALVES, Constantino. O papel da inovação tecnológica no desenvolvimento do agribusiness. Disponível em<< http://www.unopar.br/portugues/revista_cientificaj/artigosderevisao/opapeldaino/opapeldaino.html>> Acesso em 31 AGO 2014;

IBGE. Levantamento Sistemático da Produção Agrícola. Disponível em <<http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/defaulttab.shtm>> Acesso em Acesso em 31 AGO 2014;

LUCHINI, Cristina. BIANCHI, Benê. O grande salto da Produção Agrícola do Paraná. Disponível em <<http://revistacrea.crea-pr.org.br/o-grande-salto-da-producao-agricola-no-parana>> Acesso em 31 AGO 2014;

MATOS, Tauller Augusto. Uma Visão Geral das Principais Tarefas de Mineração de Dados.Disponível<<http://faa.edu.br/portal/sistemas/revistas/saber_digital/2012/6/12-47-1-PB>>. Acesso em 11 SET 2014.

MICROSOFT. Conceitos de mineração de dados. Disponível em <<http://msdn.microsoft.com/pt-br/library/ms174949.aspx>> Acesso dia 03 MAI 2014;

MINISTERIO DA AGRICULTURA. Brasil lidera produtividade agrícola na América Latina. Disponível em <<http://www.brasil.gov.br/economia-e-emprego/2009/11/brasil-lidera-produtividade-agricola-na-america-latina>> Acesso em 31 AGO 2014;

MINISTERIO DA AGRICULTURA. Dados de Exportação de Alimentos. Disponível em <<http://www.agricultura.gov.br/vegetal/exportacao/alimentos>> Acesso em 2 SET 2014.