UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ – UTFPR
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

LUCIANO RUFATTO

MINERAÇÃO DE DADOS APLICADA NA AGRICULTURA - ANÁLISE DE INSUMOS E SAFRA

TRABALHO DE DIPLOMAÇÃO

MEDIANEIRA
2015

LUCIANO RUFATTO

MINERAÇÃO DE DADOS APLICADA NA AGRICULTURA - ANÁLISE DE INSUMOS E 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 Claudio Leones Bazzi.

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

Mineração de dados aplicada na agricultura - Análise de insumos e safra

Por

Luciano Rufatto

Este Trabalho de Diplomação (TD) foi apresentado às 8:30 h do dia 03 de fevereiro de 2015 como requisito parcial para a obtenção do título de Tecnólogo no Curso Superior de Tecnologia em Manutenção Industrial, da Universidade Tecnológica Federal do Paraná, *Campus* Medianeira. Os acadêmicos foram argüidos pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado com louvor e mérito.

Prof. Dr. Claudio Leones Bazzi
UTFPR – *Campus* Medianeira
(Orientador)

Prof. Esp. Márcio Angelo Matté
UTFPR – *Campus* Medianeira
(Convidado)

Prof. Esp.Valter Erkbert
UTFPR – *Campus* Medianeira
(Convidado)

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

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

RESUMO

A modernização da agricultura fornece novas possibilidades para o campo, com o aumento constante do custo de produção gera-se a necessidade de uma melhor gerencia dos recursos utilizados. A modernização na lavoura faz com que o agricultor torne-se uma nova fonte de coleta de dados. Visando utilizar estes dados de maneira produtiva, este trabalho utiliza a mineração de dados para a descoberta de conhecimento na dosagem de insumos e sua influência na produtividade da soja. Sendo abordados os elementos que apresentaram maior influência nos índices de produtividade, abrindo portas para futuras análises com mineração de dados aplicada na agricultura.

ABSTRACT

The agriculture modernization provides new possibilities to the farm, with the rises of the production costs, emerge the needs for a better management of the resources. The farm modernization made the farmer a good source of data collection. Seeking for use this data in a productive way, this meta-paper use data mining looking for discovery knowledge in the fertilizers dosage and its influence on the soya bean production. Addressing the elements who represent bigger influence in the productive levels, opening doors for futures analysis with data mining applied in agriculture.

LISTAS

SUMÁRIO

1 INTRODUÇÃO

A necessidade do agricultor por melhorias na lavoura leva a modernização e inovação no campo.

> Através da modernização da agricultura, os produtores buscam melhores condições de enfrentar as dificuldades impostas pela natureza no que concerne à produção e melhorar alguns fatores necessários. Assim, através de uma artificial conservação e fertilização do solo, mecanização da lavoura, seleção de sementes, dentre outros recursos, busca-se a obtenção de maior produtividade. (TEIXEIRA 2005, p. 23)

Uma tecnologia a ser aplicada na agricultura é a mineração de dados, na qual por meio de algoritmos inteligentes em uma quantidade de dados, é possível encontrar padrões que passam despercebidos perante as análises comuns e/ou estatísticas, sendo possível a descoberta de novos conhecimentos na área.

A agricultura de precisão (AP) se mostra como uma potencial aliada na coleta de dados para a mineração, pois através do auxílio da área de sistemas de informação, analisa as áreas de cultivo não como um todo, mas divididas em partes, realizando um monitoramento mais preciso, buscando a aplicação mais eficiente dos insumos, para assim, obter uma produção mais homogênea. A agricultura de precisão também busca diminuir os impactos ambientais resultantes da aplicação excessiva de insumos em determinadas áreas.

Fazendo uso dos dados coletados pela AP, a mineração de dados pode ser aplicada na agricultura para diversas e indefinidas finalidades, neste projeto serão explorados os dados do solo e da produção de soja para a obtenção de conhecimento útil à agricultura.

1.1 OBJETIVO GERAL

Desenvolver um estudo e utilizar técnicas de mineração de dados buscando resultados satisfatórios na dosagem de insumos agrícolas.

1.2 OBJETIVOS ESPECÍFICOS

O trabalho possui os seguintes objetivos específicos:

- Obtenção de informações para a alimentação da base de dados;
- Efetuar a classificação dos dados;
- Determinar a melhor técnica de mineração de dados com base nos resultados obtidos;

- Desenvolver a modelagem da técnica selecionada;
- Avaliação dos resultados obtidos.

1.3 JUSTIFICATIVA

Devidos aos avanços na área de tecnologia da informação, as indústrias investem cada vez mais no gerenciamento de dados. No entanto, a quantidade de dados obtidos e armazenados é grande, mas a informação e o conhecimento extraídos são pequenos. De acordo com Han (2006), esta situação pode ser definida como "rico em dados, pobre em informação", dessa forma, as técnicas de mineração de dados são utilizadas para obter a maior quantidade de informações possíveis do montante de dados.

Assim procura-se aplicar a mineração de dados no processo de produção agrícola na busca de conhecimento, com o intuito de incentivar e expandir a utilização das técnicas de mineração de dados na agricultura.

1.4 ESTRUTURA DO TRABALHO

O trabalho foi dividido em duas partes:

- Revisão Bibliográfica: onde fala-se sobre a agricultura de precisão e as necessidades do agricultor em meio a tecnologia, e uma breve introdução sobre mineração de dados e suas aplicabilidades nos dias de hoje.
- Materiais e métodos: onde é apresentada a ferramenta de trabalho utilizada e os dados e opções da ferramenta utilizadas para este trabalho, seguido das análises resultantes da mineração.

2 REVISÃO BIBLIOGRÁFICA

2.1 AGRICULTURA DE PRECISÃO

As transformações que a agricultura moderna vem sofrendo nas últimas décadas acabaram por torná-la uma atividade altamente competitiva. O agronegócio exige dos produtores rurais um alto grau de especialização e profissionalismo.

Em conjunto com essa capacidade administrativa está a competência do agricultor de coletar dados relativos a sua área de produção, com a finalidade de inserir novas tecnologias a sua realidade. Visto que o agricultor está sujeito as condições de clima e impactos da natureza, a coleta de dados e gerenciamento de seu agronegócio que definem seu sucesso na produção agrícola.

A agricultura de precisão surgiu como um sistema de gerenciamento de informações, e potencializou seu crescimento com o avanço das tecnologias de referenciamento e posicionamento, como o GPS (do inglês, Global Positioning System) e de técnicas de sensoriamento remoto. Com o uso destas tecnologias emergiram alguns conceitos como os de aplicação de insumos em Taxas variáveis e dos Sistema de Informação Geográfica (SIG).

Segundo Nunes (2014), a forma clássica de realizar o plantio, é tratando uma área grande de forma homogênea, onde os insumos são distribuídos de forma atender uma necessidade média da lavoura, havendo assim, possíveis faltas ou sobras de insumos e determinada área. A agricultura de precisão (AP) pretende mudar esse cenário, permitindo assim a utilização correta dos insumos, de acordo com a necessidade de cada área.

A agricultura de precisão pode ser entendida como uma filosofia de gerenciamento agrícola, que utiliza como premissa informações exatas e precisas e é completada com decisões tomadas de forma exata. É uma forma de gerência de um campo produtivo metro a metro, onde cada parte da terra a ser cultivada tem diferentes características. O conceito principal é a aplicação dos insumos nos locais corretos e em momentos adequados, assim como a quantidade exata necessária para a produção, em áreas cada vez menores e mais homogêneas, de forma mais possível que os custos gerados permitam.

Com a agricultura de precisão à disposição do agricultor, é possível obter uma visão da variação espacial e temporal dos fatores climáticos e da composição do solo em cada área cultivável, em vez de tratá-la de forma geral e uniforme. Os problemas iniciais encarados para a utilização da AP são os custos elevados dos equipamentos e a dificuldade na interpretação da imensa quantidade de dados geradas. Porém, a popularização da sua utilização em

diversos países, vem causando evoluções em técnicas e equipamentos, permitindo assim cada vez mais a viabilidade da sua aplicação.

De acordo com Molin (2014), a AP tem várias formas de abordagem, porém, seu objetivo é único: utilizar estratégias para resolver os problemas de desuniformidades das lavouras e/ou até tirar proveito das mesmas. A AP pode ser praticada em diversos níveis de complexidade e com abordagens distintas. No Brasil, a AP predominante é o gerenciamento da adubação com fertilizantes e corretivos da lavoura com base na amostragem do solo, ou ainda georeferenciada baseada no GPS. Uma estratégia mais elaborada e ampla, leva em consideração os resultados da produção em ciclos anteriores para realizar a reposição correta dos nutrientes utilizados na colheita anterior. Dessa forma, é utilizada a geração de mapas de produtividade, o que exige mais trabalho e recursos sendo mais completa para a estratégia da gestão da lavoura.

2.2 MINERAÇÃO DE DADOS

"Mineração de dados é um termo coletivo para dezenas de técnicas para colher informações de dados e transformá-las em tendências e regras significativas para melhorar seu entendimento sobre os dados" (ABERNETHY, 2010).

De acordo com Abernethy (2010), a mineração de dados pode ser entendida basicamente como a transformação de grandes quantidades de dados em padrões divididos em dois tipos: direcionada e não direcionada. Na mineração de dados direcionada, busca-se prever um ponto de dados em particular, algo como tentar prever o preço de uma casa, baseada nos preços de casas de características semelhantes. Já na mineração não-direcionada, é necessária a criação de grupos de dados ou encontrar padrões em grupos de dados existentes.

A mineração de dados moderna teve início na década de 1990, quando os computadores atingiram um nível de processamento e armazenamento de dados atingiram níveis satisfatórios e seus custos permitiram as empresas utilizar seus recursos sem recorrer à utilização de recursos computacionais externos.

O termo mineração de dados é bastante abrangente, já que se refere a várias técnicas e procedimentos utilizados para examinar e processar elevadas quantidades de dados.

Em uma última instância, a mineração de dados é utilizada para criar um modelo para melhorar a forma de leitura dos dados existentes e permitir previsões mais próximas da realidade possível. Como existem várias técnicas para a mineração de dados, um passo importante para a construção de um bom modelo, é a escolha correta da técnica a ser utilizada.

Após esse passo, o modelo construído deve ser refinado, para permitir a exploração mais eficiente das suas utilidades.

2.2.1 Algoritmos de Mineração de Dados

Existem diversos algoritmos para a mineração de dados que são utilizados para resolver problemas específicos. São separados em categorias, sendo os algoritmos de associação, classificação, padrões sequenciais e agrupamento. Os algoritmos de associação se caracterizam por buscar todas as associações em que a presença de um determinado conjunto de itens em uma transação ocasiona outros itens. Os algoritmos de classificação ou geração de perfis, produzem perfis de diferentes grupos de dados. Os algoritmos de padrões sequenciais identificam tipos de padrões sequenciais com base em quantidades mínimas de informações fornecidas pelo usuário. Por fim, os algoritmos de agrupamento segmentam o grupo principal de dados em segmentos, ou subconjuntos. (Grupo de Sistemas Inteligentes - Mineração de Dados, 1998).

Os algoritmos de associação são utilizados em inúmeras situações, tais como planejamento de estoque em supermercados, envio de publicidade direcionada e planejamento de promoções de vendas. Como exemplo, uma regra de associação é derivada da mineração de dados em um banco de dados de transações, uma lista que contém um conjunto de produtos comprados em uma loja. A regra de associação poderia ser como a seguinte: "80% dos consumidores que compram carne também compram cerveja."

O número "80%" se refere ao fator de confidência, do inglês *confidence factor*, que representa a medida do poder preditivo da regra. O item do lado esquerdo da regra, o LHS, do inglês *left hand side*, é a carne, enquanto a cerveja está do lado direito da regra, sendo o RHS, *righ hand side*. Este algoritmo gera uma quantidade de informações considerável, necessitando assim que o usuário selecione o subconjunto de regras que possuam os maiores graus de confidência e a maior quantidade de listas que seguem esta regra. (Grupo de Sistemas Inteligentes - Mineração de Dados, 1998).

Algoritmos de classificação ou geração de perfis: Em um determinado conjunto de registros com atributos correspondentes, e um marcador para cada registro, uma função para classificação é utilizada para examinar o conjunto de registros marcados e gera uma descrição das características dos registros para cada classe. Como exemplo, pode-se citar uma análise de crédito, onde a empresa administradora de cartões de crédito terá registros de clientes com um

certo número de descritores. Logo, de acordo com o histórico de crédito de um consumidor, este terá o registro marcado como "bom", "médio" ou "ruim".

"Consumidores com histórico de crédito bom tem uma taxa de inadimplência de 5%."

Dessa forma, esta regra gerada pode ser utilizada para a classificação de novos conjuntos de dados.

Os algoritmos de padrões sequências buscam encontrar sequências de dados que venham a ocorrer ao longo de um determinado período de tempo. Por exemplo, um mercado pode descobrir que 80% dos clientes que comprar cerveja, compram carne.

"80% dos clientes compram cerveja seguida de carne."

Ou uma sequência semelhante pode ser:

"80% das vezes quando as vendas de cerveja aumentam, as vendas de carne também aumentam."

Dessa forma, podem ser identificados clientes alvo das promoções de carne, caso eles comprem bastante cerveja. Este algoritmo é bastante útil para empresas de catálogos e de investimentos financeiros, que são aptas para analisar sequências de eventos que afetam seus preços. Os algoritmos de agrupamento segmentam o conjunto de dados em subconjuntos ou grupos. Podem ser criados de forma estatística ou por métodos de indução não-supervisionados neurais ou simbólicos.(GSI, 1998).

Estes métodos se distinguem pelo tipo de valores de atributos que podem ser aceitos (numéricos, nominais, e objetos estruturados), a representação e organização de grupos (em hierarquias ou nível plano). Esta técnica de segmentação foi desenvolvida com o propósito de trabalhar com o processamento de pesquisas de consumidores. Como exemplo, pode-se citar um questionário com 25 questões de múltiplas alternativas, pode ser analisado por questões: Por exemplo, 45% responderam "b" questão de número 1 e assim por diante. O desafio desta técnica é analisar os 25 padrões realizados por cada consumidor. Através desta técnica os consumidores serão divididos de acordo com seu padrão de respostas, através da geração de grupos que contém a máxima similaridade e a máxima diferença entre eles (GSI, 1998).

2.2.2 Ferramentas de mineração de dados

Os algoritmos de mineração de dados são compostos basicamente de três componentes: A representação em forma de modelo, a avaliação desse modelo e o método utilizado para busca. De forma resumida, o modelo deve possuir limites flexíveis e suposições adequadas, para que os padrões possam ser encontrados, e a capacidade de predição deve ser validada, para depois os critérios utilizados na avaliação do modelo sejam optimizadas. (GSI, 1998).

As ferramentas utilizadas na mineração de dados, ou mecanismos de busca, são geralmente programas que utilizam de certa forma a inteligência artificial em banco de dados relacionais. Essas ferramentas buscam padrões pré-definidos e informam ao usuário a ocorrência de variações. Algumas ferramentas utilizadas são as redes neurais artificiais (RNA), árvores de decisão, indução de regras e visualização de dados.

As redes neurais são estruturas matemáticas com capacidade de aprendizagem. É baseada na representação do sistema nervoso humano. É muito eficiente para identificar padrões e detectar tendências muito complexas para serem percebidas pelo ser humano ou outras técnicas de mineração de dados menos eficientes. As RNA possuem um conjunto de elementos de processamentos, ou nós que fazem analogia aos neurônios humanos e tem a capacidade de aprender ao longo do tempo através do treinamento com base em padrões definidos, através da experiência. (GSI, 1998).

Já a indução trabalha com a inferência de informações em uma base de dados, sendo a dedução e a indução as duas técnicas principais para esse propósito.

- A dedução trabalha com uma consequência lógica de informação da informação do banco de dados, inferindo uma relação entre os dados analisados.
- A indução é a inferência de informações generalizada no banco de dados, vasculhando o banco de dados com padrões ou regularidades.

As árvores de decisão são representações simples do conhecimento e trabalham com a classificação de exemplos em um número finito de classes. Os nós possuem o rótulo dos atributos, os arcos são rotulados com os valores desse atributo e as folhas rotuladas com as diferentes classes. Os objetos são classificados por meio de um caminho percorrido na árvore, seguindo os arcos pelos valores correspondentes aos dos atributos do objeto. (GSI, 1998).

A indução de regras gera um conjunto de decisões não-hierárquicas, utilizado para prever valores em novos dados inseridos. A maioria das aplicações avaliam e refinam o conjunto de regras para selecionar as melhores. Essas regras são mais gerais e eficientes que as

árvores de decisão, pois utilizam as florestas de predição, que são um conjunto de árvores. (GSI, 1998).

Por fim, a análise de grupos e utilizado em ambientes de aprendizagem não supervisionada, agrupando os dados e descobrindo suas classes. O agrupamento basicamente particiona o banco de dados e faz com que cada partição possua dados similares, baseado em alguma métrica ou técnica de mineração.

2.3 ESTADO DA ARTE

Existem diversos casos de aplicação de mineração de dados na agricultura de precisão, tais como:

- Em Vriesmann et al. (2004), foi proposta a utilização de técnicas de inteligência artificial na mineração de dados de produtividade do solo;
- Em Souza et al. (2010), foi realizada a pesquisa sobre a análise de atributos do solo e da produtividade da cultura da cana de açúcar com o uso da geoestatística e árvore de decisão;
- O trabalho de Weber et al. () consistiu na aplicação de redes neurais artificiais para avaliação da influência da compactação do solo na produtividade de milho em dados agrícolas georeferenciados.
- A pesquisa realizada por Guimarães (2005), consistiu na aplicação da computação evolucionária da mineração de dados físico-químicos da água e do solo.

3 MATERIAIS E MÉTODOS

3.1 AQUISIÇÃO DAS INFORMAÇÕES

Para o desenvolvimento deste projeto serão analisados os dados georeferenciados de uma área de 16ha (hectares), sendo os dados de Soja: atributos físicos e químicos do solo,os dados de clima: chuva. E os dados de produção em Tonelada por hectare.

3.2 TRANSFORMAÇÃO DOS DADOS

3.2.1 Arquivo ARF

O arquivo de dados aceito pela Weka é a extensão ".arff". Os arquivos arff aceitam basicamente dois tipos de dados os do formato String (nominal) e Numeric. Normalmente o minerador utiliza valores nominais para trabalhar com classificação e associação e utiliza valores numéricos para tarefas de agrupamento, o que não é necessariamente uma regra.

O nome do arquivo é declarado na primeira linha com a anotação *@relation*, os atributos são declarados em seguida com a anotação *@attribute* seguido do nome do atributo e o tipo de dado. E por último a anotação *@data*, são adicionados os valores respectivos de cada atributo separados por vírgula, onde cada linha representa um registro de dados, conforme Figura 1.

```
@relation lavoura

@attribute umidade real
@attribute cobre real
@attribute zinco real
@attribute ferro real
@attribute manganes real
@attribute fosforo real
@attribute carbono real
@attribute ph real
@attribute HAl3 real
@attribute calcio real
@attribute magnesio real
@attribute potassio real
@attribute penetracao100 real
@attribute penetracao200 real
@attribute penetracao300 real
@attribute penetração400 real
@attribute mediaPenetracao real
@attribute massaGramas real
@attribute umidade2 real

@data
10.43,9.60,7.30,22.00,124.00,23.50,16.70,5.70,3.42,6.37,2.94,0.50,2261.43,5433.17,4038.08,5332.98,4266.41,461.07,11.65
10.73,7.80,8.20,18.00,114.00,16.30,18.18,5.80,3.18,7.15,3.57,0.25,5627.90,6396.80,5310.85,7124.32,6114.97,565.72,12.09
11.43,7.60,5.20,19.00,106.00,24.90,15.96,5.90,2.95,6.48,3.67,0.23,4657.08,4725.75,4269.09,6530.38,5045.57,446.27,15.28
10.83,7.20,10.20,18.00,135.00,26.20 19.29,5.90,3.18,6.88,3.41,0.71,2853.58,3996.65,4410.15,5755.65,4254.01,453.57,11.35
11.90,8.60,6.00,26.00,118.00,17.70,17.44,5.70,3.69,5.94,3.19,0.29,6171.68,5590.08,4977.65,5827.18,5641.64,398.99,12.22
11.83,8.60,6.00,26.00,118.00,17.70,17.44,5.70,3.69,5.94,3.19,0.29,6171.68,5212.93,5236.73,5827.18,5612.13,353.88,11.80
12.78,8.90,10.60,33.00,124.00,57.30,18.55,5.70,3.42,7.38,2.94,0.91,3115.13,3853.30,3451.90,4943.85,3841.04,396.79,10.86
14.25,8.40,5.90,21.00,142.00,12.00,19.29,6.20,2.95,7.06,3.63,0.24,2987.98,4539.40,4724.18,4820.13,4267.92,436.47,11.09
12.95,7.90,5.10,30.00,130.00,9.20,16.70,5.80,3.42,6.19,2.90,0.27,3189.93,3629.46,4734.03,5607.17,4290.15,400.16,10.42
15.70,7.30,11.70,24.00,144.00,18.10,17.44,6.10,2.95,6.79,3.62,0.30,3718.03,4436.45,6089.07,7377.53,5405.27,447.59,9.36
11.93,7.90,5.10,19.00,131.00,13.50,15.21,6.10,2.95,7.04,3.65,0.22,4445.55,5690.58,4764.10,5142.19,5010.60,416.17,10.99
11.23,7.30,6.30,18.00,157.00,36.60,19.29,6.10,2.95,5.96,3.02,0.27,3654.35,5432.63,4867.60,5925.33,4969.98,407.88,11.40
10.63,9.30,9.10,28.00,167.00,15.70,18.18,6.10,2.95,6.99,3.99,0.18,5355.73,4961.90,4406.13,5693.05,5104.20,378.13,11.39
```

Figura 1: Arquivo Arff

3.2.2 Preparação dos dados

Os dados obtidos neste trabalho são de formatos numéricos, o que limita a utilização de alguns algoritmos de mineração. Para melhor trabalhar os dados e a mineração, foi adicionado um atributo nominal ou atributo classe, que é um indicador de qualidade da produtividade, assim a weka poderá apresentar os dados de maneira mais amigável, além de permitir o uso de algoritmos de classificação, que apresentaram os melhores resultados para os dados em questão.

O atributo classe adicionado a este exemplo foi chamado de produtividade. Sendo classificada com base na quantidade produzida por hectare. Os dados foram classificados conforme a Figura 2.

< = 365	Baixa
365-417	Média
418-469	Boa
> 469	Excelente

Figura 2: Índices de produtividade - Ton/ha^{-1}.

3.3 PROCESSAMENTO DOS DADOS

Para o processamento dos dados obtidos foi utilizado a ferramenta WEKA. Esta ferramenta foi desenvolvida pela Universidade de Waikato, na Nova Zelândia, e foi implementado em sua forma moderna em 1997. É escrito na linguagem Java e contém uma interface gráfica para interagir com os arquivos de dados e resultados visuais (tabelas e gráficos). Também possui uma API (bibliotecas) para ser incorporada a outras aplicações.

Figura 3: Tela inicial da Weka

Para este trabalho, foram utilizadas as opções de ferramentas a partir do menu "Explorer".

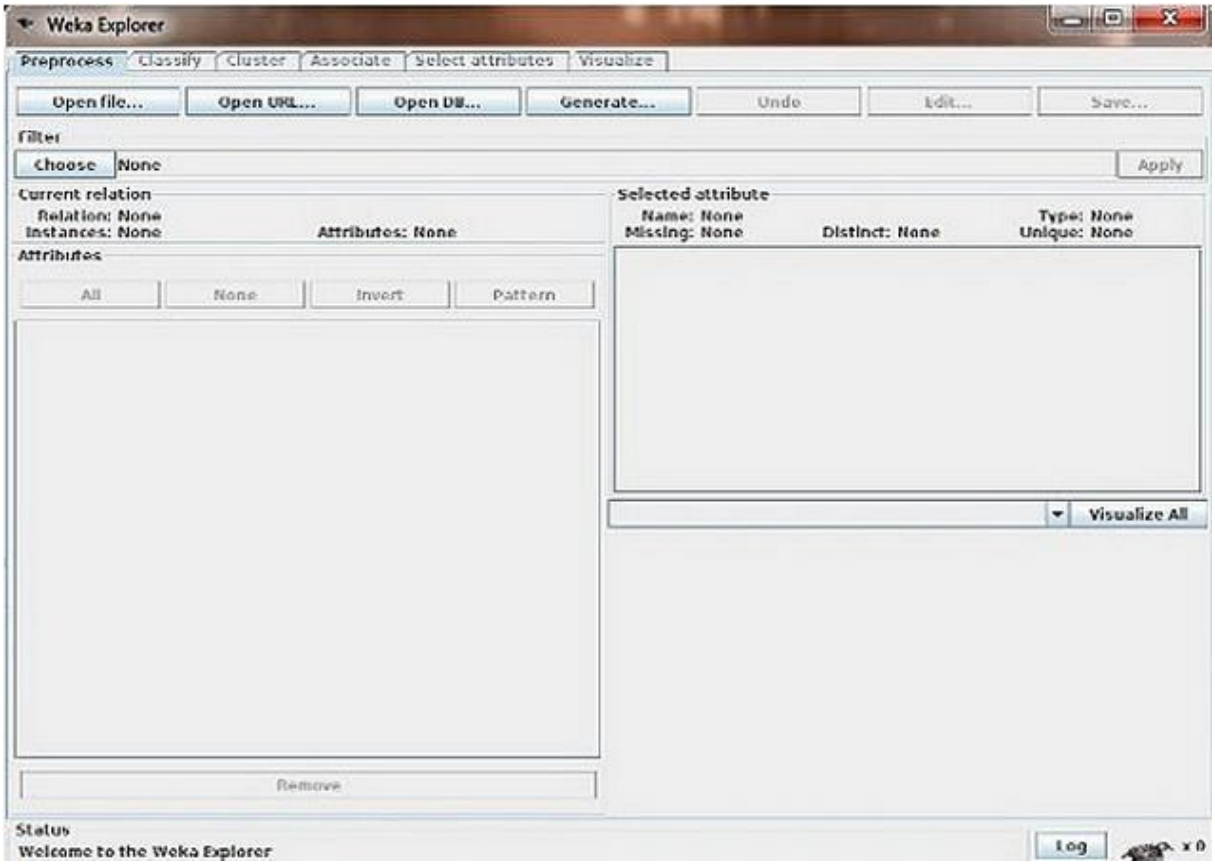

Figura 4: Tela do Weka Explorer

Neste contexto da Figura 4, serão carregados os dados do arquivo ARF preparado para a Weka.

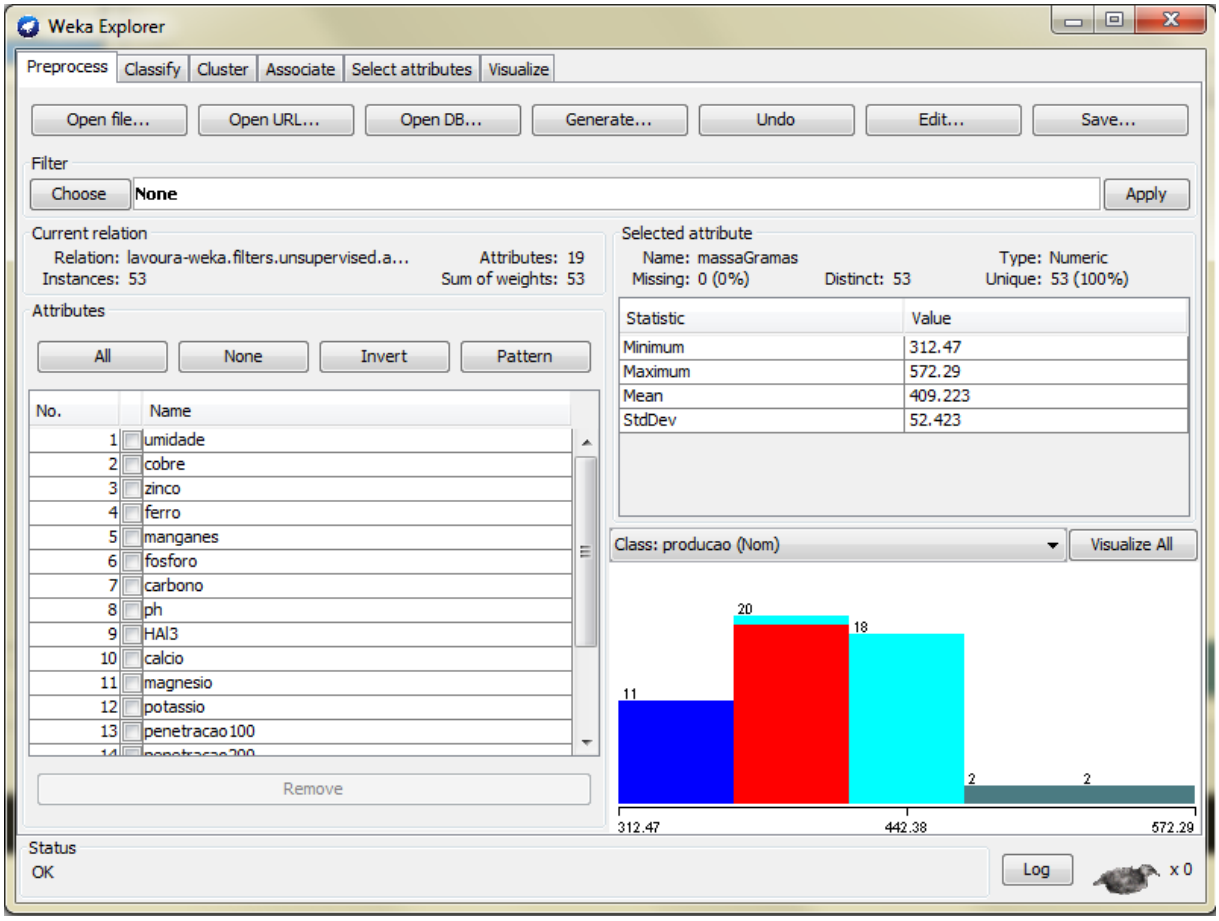

Figura 5: Dados carregados no contexto Weka explorer

Na Figura 5, pode-se observar a direita um gráfico relativo ao atributo nominal "produção" ou atributo classe, que qualifica a produção do ponto georeferenciado entre baixa, média, boa ou excelente. Sendo representadas respectivamente pelas cores: azul, vermelho, ciano-claro e ciano-escuro.

Arranhando a superfície do Weka Explorer, na aba *Preprocess* já é possível uma visualização da relação de cada atributo do solo cruzada com o atributo classe que qualifica a produtividade de cada ponto georeferenciado.

A partir deste contexto já é possível a visualização de alguns padrões nos dados.

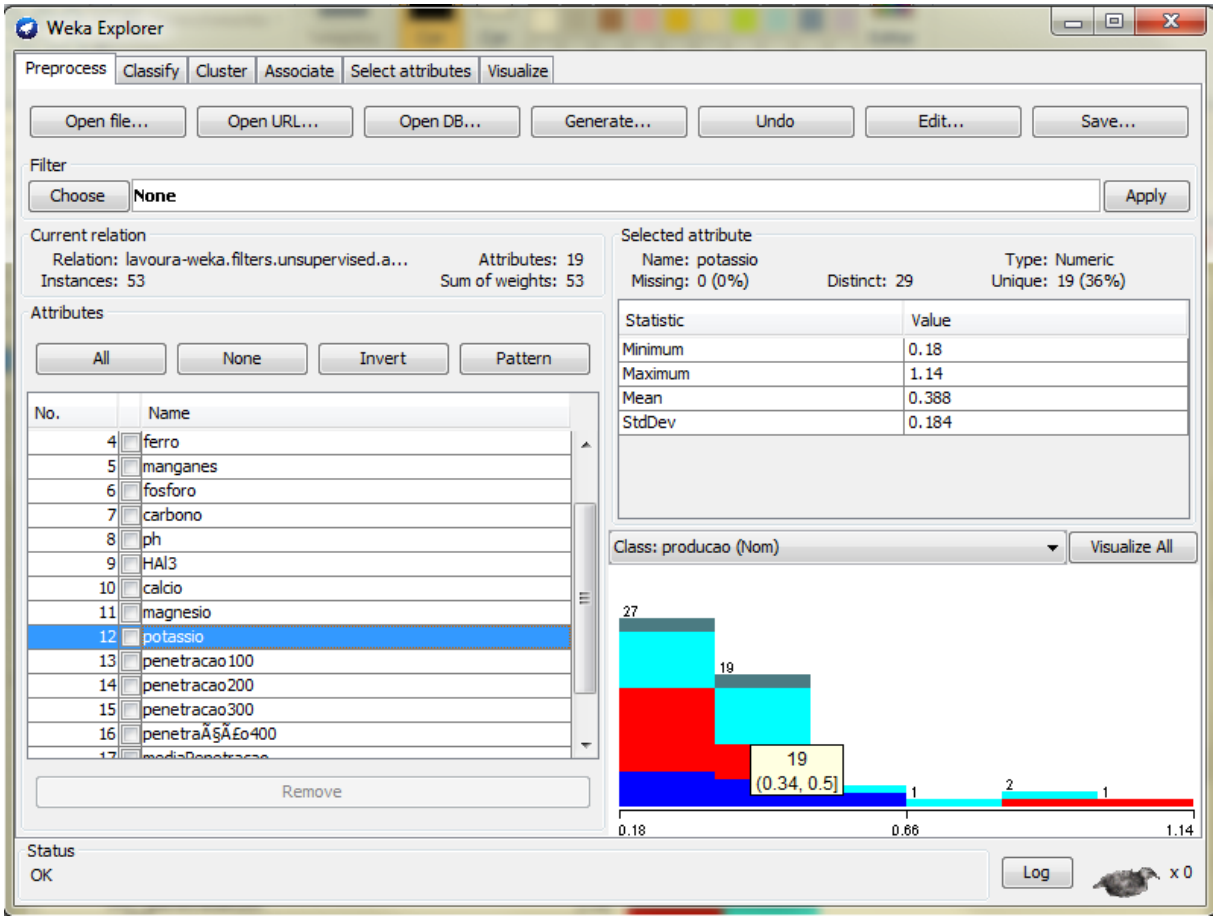

Figura 6:Gráfico de Potássio por produção

Na Figura 6, por exemplo pode-se verificar que os resultados associados a uma boa produção possuem uma quantidade de potássio(K) entre 0.18 e 0,5 cmolc dm^{-3}.

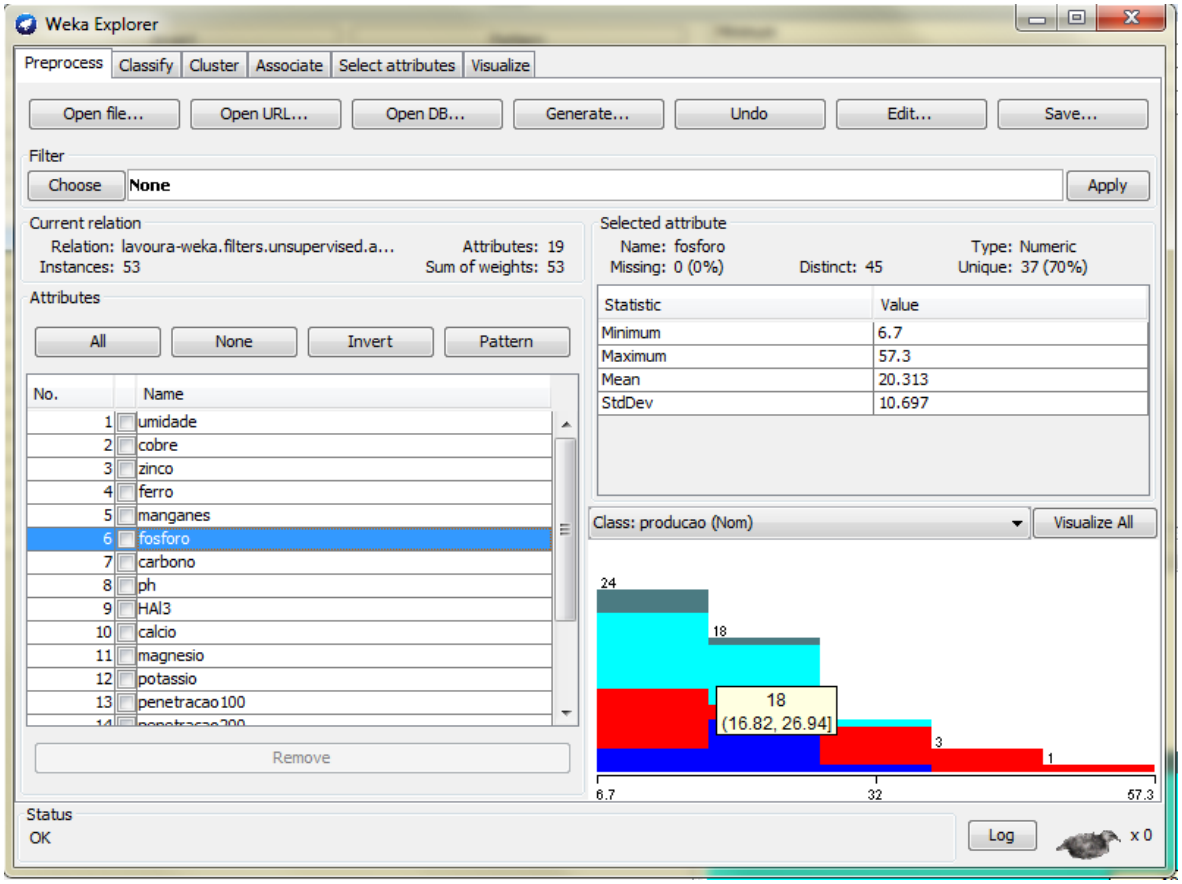

Figura 7:Gráfico de Fosforo por produção

Já na Figura 7, é possível verificar que os melhores índices de produção encontram uma quantidade entre 6.7 e 26.94 mg dm^{-3}de fósforo (P), e quase todos os pontos superiores a esta medida apresentaram apenas uma produção média, representada pela cor vermelha.

3.4 ALGORITMO DE MINERAÇÃO

Algoritmos de mineração utilizam funções matemáticas inteligentes para a realização de suas tarefas de mineração, neste trabalho será apresentado apenas aquele que mostrou resultados efetivos o objeto de estudo

3.4.1 Simple K-Means

O algoritmo K-Means foi selecionado por organizar os registros em grupos, denominados *clusters*, com membros que sejam familiares em algum aspecto, formando conjuntos de dados mais homogêneos entre si e mais heterogêneos possíveis entre os outros

conjuntos formados, facilitando a análise dos dados pelo reconhecimento de padrões e agrupamentos similares.

O K-Means permite a seleção de alguns atributos para serem ignorados antes que realize o agrupamento, pois alguns atributos podem influenciar negativamente o agrupamento dos dados, gerando resultados que escondem os padrões desconhecidos que busca-se encontrar.

Para a análise da Figura 8 foram ignorados cerca de quatro atributos referentes a Resistencia Mecânica do solo a Penetração (RMP) com a unidade medida MPa, a 0-0,1m, 0,1-0,2m, 0,2-0,3m e 0,3-0,4m. Deixando a RMP média que é a média entre os valores destes quatro atributos.

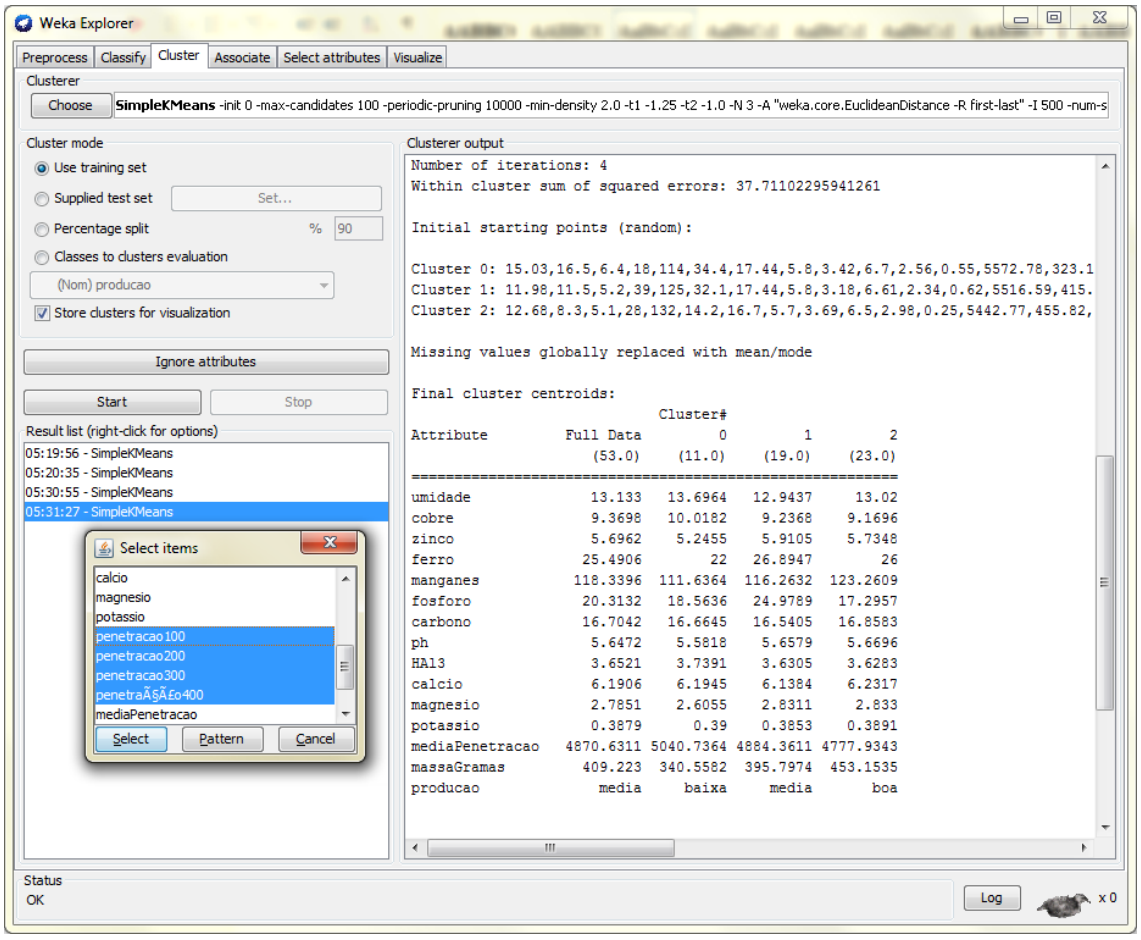

Figura 8:Algoritmo K-Means

O algoritmo foi configurado para dividir os dados em três agrupamentos, os valores mostrados nas tabelas, são referentes a média dos valores dos registros contidos em cada clustere no total dos dados. Para melhor compreensão do tipo de conhecimento que pode ser extraído basta analisar a Figura 9.

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(53.0)	(11.0)	(19.0)	(23.0)
umidade	13.133	13.6964	12.9437	13.02
cobre	9.3698	10.0182	9.2368	9.1696
zinco	5.6962	5.2455	5.9105	5.7348
ferro	25.4906	22	26.8947	26
manganes	118.3396	111.6364	116.2632	123.2609
fosforo	20.3132	18.5636	24.9789	17.2957
carbono	16.7042	16.6645	16.5405	16.8583
ph	5.6472	5.5818	5.6579	5.6696
HAl3	3.6521	3.7391	3.6305	3.6283
calcio	6.1906	6.1945	6.1384	6.2317
magnesio	2.7851	2.6055	2.8311	2.833
potassio	0.3879	0.39	0.3853	0.3891
mediaPenetracao	4870.6311	5040.7364	4884.3611	4777.9343
massaGramas	409.223	340.5582	395.7974	453.1535
producao	media	baixa	media	boa

Figura 9:Análise de 3 clusters com algoritmo K-means

Nos dados em destaque na Figura 9, pode-se observar que no cluster 0, o qual representa a menor média de produção, o elemento Ferro(Fe) encontra-se em uma quantidade significativamente inferior 22mg $dm^{-3,}$ em relação à média de todos os dados 25.49mg dm^{-3} e dos demais clusters 1 26.89mg dm^{-3} e 2 26mg dm^{-3}.

O elemento Manganês(Mn)também aparece em concentração menor onde há baixa produtividade no cluster 0 com 111.6cmolc dm^{-3} e em níveis respectivamente maiores no cluster 1 com 116.2cmolc dm^{-3} que representa produtividade média e no cluster 2 com 123.2cmolc dm^{-3} que representa produtividade boa.

Ainda na Figura 9, os dados com baixa produtividade no cluster 0 apresentam um índice maior de RMP media com 4870.6MPa, enquanto a média produtividade no cluster 1 apresenta 4884.3MPa e o cluster 3 com maior média de produtividade apresenta a menor média, 4777.1MPa.

Efetuando uma segunda análise, mas desta vez configurando o algoritmo a utilizar apenas 2 clusters para agrupamento dos dados. Reforça-se a evidencia da influência de elementos como o Ferro, o Manganês e a Resistencia Mecânica a Penetração do solo, e apresenta o elemento Zinco e Magnésio como possíveis influente na produtividade, conforme a análise da figura 10.

Final cluster centroids:

Attribute	Full Data (53.0)	Cluster# 0 (20.0)	1 (33.0)
umidade	13.133	13.7095	12.7836
cobre	9.3698	9.86	9.0727
zinco	5.6962	5.015	6.1091
ferro	25.4906	23.5	26.697
manganes	118.3396	104.65	126.6364
fosforo	20.3132	20.295	20.3242
carbono	16.7042	16.104	17.0679
ph	5.6472	5.515	5.7273
HAl3	3.6521	3.8645	3.5233
calcio	6.1906	5.9505	6.3361
magnesio	2.7851	2.5675	2.917
potassio	0.3879	0.367	0.4006
mediaPenetracao	4870.6311	4993.6835	4796.0539
massaGramas	409.223	363.1375	437.1536
producao	media	baixa	boa

Figura 10:Análise de 2 clusters com K-means

Na Figura 10 é possível evidenciar que o cluster 0, relativo a baixa produtividade, possui quantidades significativamente inferiores de Ferro e Manganês, e apresenta RMP novamente superior, em relação a estes mesmos atributos no cluster 1 que representa maior produtividade. Não só isso, apresentou desta vez um novo elemento com diferença significativa entre os dois cluster, o Zinco, que apresentou uma média cerca de 20% superior, com 6.1mg dm^{-3} onde houve melhor produtividade.

Em uma terceira análise com o K-means, agora utilizando 4 clusters para agrupar os dados de maneira mais homogênea entre si, e adicionando o atributo de RMP a 0,1m. Observa-se os seguintes resultados na Figura 11.

Final cluster centroids:

		Cluster#			
Attribute	Full Data	0	1	2	3
	(53.0)	(22.0)	(11.0)	(5.0)	(15.0)
umidade	13.133	13.0991	13.6964	12.148	13.098
cobre	9.3698	9.1	10.0182	8.8	9.48
zinco	5.6962	5.8409	5.2455	8.78	4.7867
ferro	25.4906	26.1818	22	27.2	26.4667
manganes	118.3396	125.0909	111.6364	142.6	105.2667
fosforo	20.3132	17.1091	18.5636	37.52	20.56
carbono	16.7042	16.8318	16.6645	18.552	15.93
ph	5.6472	5.6818	5.5818	5.9	5.56
HAl3	3.6521	3.6127	3.7391	3.184	3.802
calcio	6.1906	6.2355	6.1945	6.786	5.9233
magnesio	2.7851	2.8732	2.6055	3.108	2.68
potassio	0.3879	0.3905	0.39	0.566	0.3233
penetracao100	3965.8458	3632.1214	4077.0245	4244.238	4280.98
mediaPenetracao	4870.6311	4808.2345	5040.7364	4725.5	4885.7793
massaGramas	409.223	447.7382	340.5582	394.748	407.9133
producao	media	boa	baixa	media	media

Figura 11: Análise de 4 clusters com K-means

Com um agrupamento de 4 clusters, foi possível uma melhor identificação dos atributos que merecem mais atenção para análise. Conforme a Figura 11, são os mesmos atributos: Ferro, Manganês e os atributos de RMP se apresentam com as diferenças mais significativas entre e boa e a baixa produtividade, enquanto que os atributos apresentados anteriormente com diferenças significativas como o Zinco e o Magnésio se mostram irrelevantes para posteriores análises.

Visto que os dados são agrupados nos clusters conforme a proximidade dos valores de cada atributo, para reforçar os resultados faz-se necessário realizar novas análises apenas com os atributos que se mostraram mais significantes para os resultados: Ferro, Manganês, RMP.

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(53.0)	(23.0)	(19.0)	(11.0)
ferro	25.4906	26	26.8947	22
manganes	118.3396	123.2609	116.2632	111.6364
mediaPenetracao	4870.6311	4777.9343	4884.3611	5040.7364
massaGramas	409.223	453.1535	395.7974	340.5582
producao	media	boa	media	baixa

Figura 12: Análise de 3 clusters – Atributos significativos

Os resultados se reforçam. Novamente onda há baixa produtividade, no cluster 2, encontra-se valores inferiores de Ferro e Manganês, e valor superior de RMP, conforme Figura 12.

Para melhor validação da importância destes atributos foi realizada uma análise isolada dos mesmos.

3.4.1.1 Manganês (Mn)

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(53.0)	(11.0)	(23.0)	(19.0)
manganes	118.3396	111.6364	123.2609	116.2632
massaGramas	409.223	340.5582	453.1535	395.7974
producao	media	baixa	boa	media

Figura 13: Análise de 3 clusters – Manganês

Mesmo com o elemento isolado, assim como nas análises anteriores o Mn aparece com uma média superior no cluster relativo a melhor produtividade, com 123.26cmolc dm^{-3}.

Em todas análises de agrupamento realizadas, conforme maior o índice de Mn, melhores foram os resultados de produtividade. Segundo Melarato et. al:

> O manganês desempenha funções importantes na vida da planta. Entre essas, estão a ativação de enzimas e a participação na reação de fotólise da água e na evolução do O2 no sistema fotossintético, na formação de clorofila e na formação, multiplicação e funcionamento dos cloroplastos. (MELARATO, PANOBIANCO, VITTI, VIEIRA, 2002).

Conforme o gráfico apresentado pela Weka na Figura 13, pode-se verificar que os pontos em que houveram produtividade baixa ou média, representados respectivamente pelas cores azul e vermelha, se mostraram mais evidentes onde há quantidades de Mn fora do intervalo de 125-146cmolc dm^{-3}.

De acordo com os dados analisados, uma quantidade entre 125 e 146cmolc dm^{-3} de Mn, apresentou maior ocorrência de índices de boa produtividade. Enquanto que abaixo de 125cmolc dm^{-3} houveram muitas ocorrências de baixa ou média produtividade.

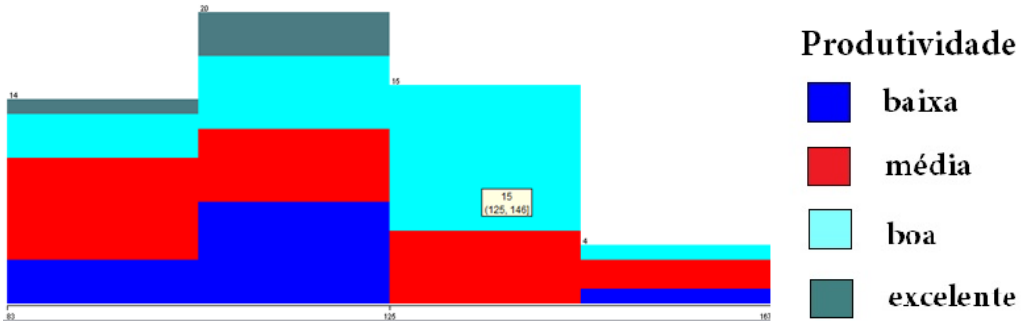

Figura 14: Gráfico de Manganês por produtividade.

No gráfico da Figura 14 fica evidente os índices de baixa produtividade o Manganês é inferior a 125cmolc dm^{-3}.

3.4.1.2 Ferro (Fe)

Segundo Sfredo (2008), como 75% do Fe das folhas estão nos cloroplastos, quando há deficiência de Fe, diminui o teor de clorofila e o número de cloroplastos nas folhas de soja.De acordo com Lacerda et. Al.(2007) que produziu uma série de apostilas sobre Fisiologia Vegetal, especificamente na Unidade V – Fotossíntese, os cloroplastos contêm um pigmento verde chamado de clorofila o qual absorve a luz necessária para realização da fotossíntese.

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(53.0)	(11.0)	(23.0)	(19.0)
ferro	25.4906	22	26	26.8947
massaGramas	409.223	340.5582	453.1535	395.7974
producao	media	baixa	boa	media

Figura 15: Análise com 3 clusters – Elemento Ferro

3.4.1.3 Resistência Mecânica a penetração do solo (RMP)

De acordo com Foloni(2006), “um solo compactado tem o arranjo estrutural, a porosidade total, a difusão de gases, a infiltração e o armazenamento de água comprometidos, que, por consequência, afetam o crescimento radicular das plantas”.

Conforme a análise da Figura 16, o cluster 0 com melhor produtividade apresentou uma média menor de compactação do solo, enquanto que os clusters relativos a média e baixa produtividade apresentaram respectivamente índices maiores de compactação do solo.

```
Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1          2
                      (53.0)     (23.0)     (19.0)     (11.0)
=============================================================
mediaPenetracao    4870.6311 4777.9343 4884.3611 5040.7364
massaGramas         409.223   453.1535   395.7974   340.5582
producao              media       boa      media      baixa
```

Figura 16: Análise de 3 cluster – RMP

A análise do gráfico na Figura 17, evidencia que os índices relativos a baixa e média produtividade, representados pelas respectivas cores azul e vermelho, apresentam maior ocorrência onde o RMP foi superior a 4295.826KPA.

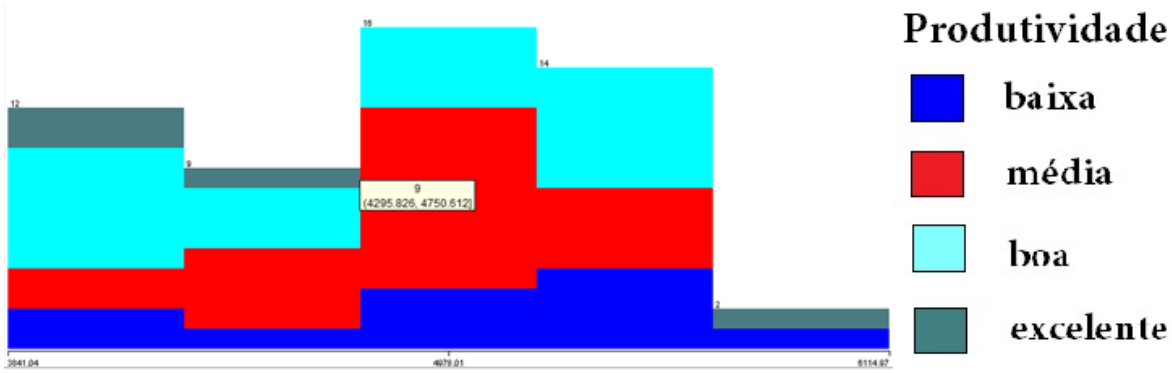

Figura 17: Gráfico de RMP por qualidade de produção

O gráfico da Figura 17 evidencia a diferença de produtividade onda há maior resistência de penetração ao solo.

4 CONSIDERAÇÕES FINAIS

4.1 CONCLUSÃO

Os dados necessitaram de um atributo de classificação para o uso de alguns algoritmos de mineração utilizados pela Weka.

Após várias análises com algoritmos como J48, RandomTree, RandomForest, NavBayes, ZeroR e LinearRegression, o algoritmo K-Means foi selecionado como a melhor técnica de mineração, sendo o único que apresentou resultados significativos.

As análises dos dados com o K-Means apontaram que atributos como o Manganês, Ferro e a Resistencia Mecânica a Penetração do solo se apresentaram de maneira influente na produtividade da soja.

A falta de dados não permitiu uma análise mais ampla dos resultados, visto que quando se trabalha com agrupamentos, uma quantidade maior de dados permite uma variedade maior de agrupamentos e uma análise melhor das anomalias encontradas no meio.

4.2 TRABALHOS FUTUROS/CONTINUAÇÃO DO TRABALHO

O projeto mostrou-se eficiente em termos de análise dos resultados da safra, para futuros trabalhos uma coleta mais ampla de dados faz-se necessária.

A mineração de dados ainda tem muito potencial a ser aplicado na agricultura, espera-se que este trabalho venha a abrir portas para a utilização desta tecnologia no campo, trazendo a descoberta de novos conhecimentos e soluções de problemas relativos a produtividade que são ainda desconhecidos na área.

REFERÊNCIAS BIBLIOGRÁFICAS

TEIXEIRA, Jodenir Calixto. Modernização da agricultura no Brasil: Impactos econômicos sociais e ambientais. **Revista Eletrônica da Associação dos Geógrafos Brasileiros** [online], Três Lagoas v. 2, n. 2, ano 2, setembro de 2005. Disponível em <http://www.cptl.ufms.br/geo/revista-geo/Revista/Revista_ano2_numero2/jodenir.pdf> Acesso em 12 de Janeiro de 2015.

GEOREFERENCIAMENTOAgricultura de precisão. In:AGROLINK "JOSÉ LUIZ DA SILVA NUNES": disponível em <http://www.agrolink.com.br/georreferenciamento/ AgriculturaPrecisao.aspx>Acesso em 11 de janeiro de 20015.

Mineração de dados com Weka, parte 2: Classificação e armazenamento em cluster. In IBM "MICHAEL ABERNETHY": disponível em <http://www.ibm.com/developerworks /br/opensource/library/os-weka2/> Acesso em 13 de janeiro de 2015.

ALGORITMOS de mineração - tecnologias relacionadas. In: Grupo de Sistemas Inteligentes – Mineração de dados. Desenvolvido por DIN Departamento de Informática da Universidade Estadual de Maringa. Apresenta textos sobre Mineração de dados. Disponível em <http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html> Acesso em 10 de janeiro de 2015

MINERAÇÃO DE DADOS Extração de arvores de decisão com a ferramenta weka de data mining weka. In Devmedia "EDUARO CORREIA GONÇALVES": disponível em <http://www.devmedia.com.br/extracao-de-arvores-de-decisao-com-a-ferramenta-de-data-mining-weka/3388> Acesso em 15 de janeiro de 2015

VRIESMANN, Leia Maria, et al. Análises de resultados obtidos por técnicas de inteligência artificial na mineração de dados de produtividade do solo. **Revista Brasileira de Agrocomputação** [online], Ponta Grossa v.2, n.1, p.11-18 junho de 2004. Disponível em <http://www.leb.esalq.usp.br/molin/analiseinteligencia.pdf> Acesso em 15 de janeiro de 2015.

SOUZA, Zigomar Menezes de, et al. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciência Rural**, Santa Maria, v. 40, n.4, p.840-847, abril de 2010 [online]. Disponível em <http://www.scielo.br/pdf/cr/v40n4/a527cr1043.pdf> Acesso em 15 janeiro de 2015.

WEBER, A. R. H.; JUNIOR, J. I. O.; PROENÇA, C. A.; GUIMARÃES, A. M.; ROCHA J. C. F.; POZO, A. T. R. Aplicação de Redes Neurais Artificiais para avaliação da influência da compactação do solo na produtividade de milho em dados agrícolas georeferenciados. In. CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 8., 2011, Bento Gonçalves. **Anais eletrônicos**... Florianópolis: UFSC, 2011. Disponível em <http://www.gse.ufsc.br/sbiagro/wp-content/anais/anais/apresentacaoPoster/poster_sessaoI/89866_1.pdf> Acesso em 10 de janeiro de 2015.

GUIMARÃES, A. M. **Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo**. Botucatu: UNESP, 2005. Disponível em <http://base.repositorio.unesp.br/bitstream/handle/11449/101851/guimaraes_am_dr_botfca.pdf?sequence=1&isAllowed=y> Acesso em 11 de janeiro de 2015.

MELARATO, Marcelo, et. al. Manganês e potencial fisiológico de sementes de soja. **Ciência Rural**, Santa Maria, v.32, n.6, p.1069-1071, ano de 2002[online]. Disponível em<http://www.scielo.br/pdf/cr/v32n6/12757.pdf> Acesso em 23 de janeiro de 2015.

FISIOLOGIA VEGETAL – UNIDADE V - FOTOSSÍNTESE. Universidade Federal do Ceará. Apostila em pdf. Disponível em <http://www.fisiologiavegetal.ufc.br/APOSTILA/FOTOSSINTESE.pdf > Acesso em 27 de janeiro de 2015.

FOLONI, J. S. S, et. al. Crescimento Aéreo e Radicular da soja e de plantasde cobertura em camadas compactadas de solo. Revista Brasileira de Ciência do Solo. v.30, p.49-57, ano de 2006[online]. Disponível em <http://www.scielo.br/pdf/rbcs/v30n1/a06v30n1.pdf >. Acesso em 28 de janeiro de 2015.

MOLIN, José Paulo. Boletin Técnico – Agricultura de precisão. Brasília, p.5-27, 2011. Disponível em <http://www.agricultura.gov.br/arq_editor/Boletim%20T%C3%A9cnico%20AP.pdf>. Acesso em 10 de janeiro de 2015.