UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA DA USP

GABRIEL DUARTE GRANDULPHO

Big Data Aplicado a Indicadores Econômicos do Agronegócio

São Paulo
Dezembro de 2016

Big Data Aplicado a Indicadores Econômicos do Agronegócio

Gabriel Duarte Grandulpho

Apresentado por:

_____

Gabriel Duarte Grandulpho

Aprovado por:

_____

Prof. Dr. Pedro Luiz Pizzigatti Corrêa
Departamento de Engenharia de Computação e Sistemas Digitais
Escola Politécnica da Universidade de São Paulo

São Paulo
Dezembro de 2016

Gabriel Duarte Grandulpho

Big Data Aplicado a Indicadores do Setor de Agronegócio

Monografia final do Curso de Big Data: Inteligência na Gestão de Dados da Escola Politécnica da USP como parte dos requisitos para obtenção do título de Especialista em Big Data.

Área de Concentração:
Tecnologia da Informação

Orientador:
Prof. Dr Pedro Luiz Pizzigatti Corrêa

São Paulo
Dezembro de 2016

Dedico esse trabalho aos meus pais, Izaura e Carlos, por todo o apoio em tornar esse sonho possível.

.

Agradecimentos

Gostaria de agradecer a minha família por todo o apoio durante toda essa caminhada, que serviram de base para que eu pudesse ter a oportunidade de chegar até aqui. Um agradecimento a todo o corpo docente do curso, por compartilhar conosco seu conhecimento e experiência que será de grande valia em minha carreira profissional e acadêmica.

Um agradecimento ao Prof. Pedro Luiz Pizzigatti Corrêa, por toda a ajuda e aprendizado compartilhado durante esses 3 meses de desenvolvimento da monografia.

Também gostaria de agradecer ao Prof. Fernando Elias Correa, que com seu conhecimento e por ter compartilhado informações de estudos realizados previamente ao CEPEA, fizeram com que esse estudo de caso fosse possível.

Motivação Profissional do Curso

O curso de *Big Data* - Inteligência na Gestão de Dados foi de grande importância no meu desenvolvimento profissional pois proporcionou a abertura de novas oportunidades em minha carreira profissional. As aulas ministradas foram de grande importância para o aprendizado de novas técnicas e tecnologias inovadoras que estão sendo difundidas no mercado para apoiar na administração de dados de forma eficiente e eficaz, apoiando os mais diversos setores na gestão de informações de forma inteligente.

Através do conhecimento adquirido durante o curso, pude participar de projetos relacionados a *Big Data*, onde tive contato com ferramentas de processamento massivo de dados, como Spark e Hadoop, além de ferramentas de análises preditivas e mineração de dados.

De forma geral, o curso foi importante para meu desenvolvimento profissional e acadêmico.

RESUMO

A tecnologia em geral tem revolucionado inúmeros segmentos das cadeias produtivas do agronegócio. A utilização de novos métodos e tecnologias, tornou-se essencial para garantir maior competitividade e lucro, com o menor investimento possível. Tecnologias embarcadas, equipamentos de monitoramento e outros sensores permitem a captura de dados desde o plantio até a colheita. Após o surgimento do conceito de *Big Data*, tem-se discutido sobre suas possibilidades de aplicação e contribuições que possam ser aplicadas para apoiar o agronegócio. Através do processamento e análise massiva de dados é possível trazer *insights*, e dessa forma otimizar a produção de determinado produto ou serviço. O intuito deste trabalho é realizar um estudo de caso sobre o Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Escola Superior de Agricultura “Luiz de Queiroz” da Universidade de São Paulo, para o qual são estudados os métodos e técnicas aplicadas pelo Centro para a coleta e a manipulação de dados relacionados ao Agronegócio, além de discutir possíveis aplicações de *Big Data*, análise e mineração de dados estudados durante o curso de Especialização em Big Data.

Palavras-chave: Banco de Dados. Big Data. Indicadores do Agronegócio.

ABSTRACT

Technology in general has revolutionized numerous segments of agribusiness production chains. The use of new methods and technologies, has become essential to ensure greater competitiveness and profit, with the least investment possible. Embedded technologies, monitoring equipment and other sensors allow the capture of data from planting to harvesting. After the emergence of the Big Data concept, it has been discussed about its possibilities of application and contributions that can be applied to support agribusiness. Through the processing and massive analysis of data it is possible to bring insights, and in this way to optimize the production of a given product or service. The purpose of this paper is to carry out a case study about the Center for Advanced Studies in Applied Economics (CEPEA) of the "Luiz de Queiroz" School of Agriculture of the University of São Paulo, for which the methods and techniques applied by the Center For the collection and manipulation of data related to Agribusiness, in addition to discussing possible applications of Big Data, analysis and data mining studied during the course of Specialization in Big Data.

Keywords: Database. Big Data. Agribusiness Indicators

LISTA DE ABREVIATURAS E SIGLAS

CEPEA	Centro de Estudos Avançados em Economia Aplicada
GPS	Global position System
UNESP	Universidade Estadual Paulista
KDD	Knowledge-Discovery in Databases.
PIB	Produto Interno Bruto
ETL	Extract Transformation Load
KNN	K-Nearest Neighbor
PIB	Produto Interno Bruto
OLAP	Online Analytical Processing
PAA	Piecewise Aggregate Approximation
MDIC	Ministério do Desenvolvimento, Indústria e Comércio Exterior
CRM	Customer Relationship Management
IBGE	Instituto Brasileiro de Geografia e Estatística
RNA	Redes Neurais Artificiais
ERP	Enterprise Resource Planning
AVT	Aplicações em Taxas Variáveis
BM&F	Bolsa de Mercadorias e Futuros
IPEA	Instituto de Pesquisa Econômica Aplicada
SAX	Symbolic Aggregate Approximation
DTW	Dynamic Time Warping

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Com o crescimento exponencial da tecnologia no mundo, e a criação de novos dispositivos eletrônicos, fizeram com que a produção de dados tornar-se ainda maior nos últimos anos. Diariamente usuários adquirem novos equipamentos, e estes geram novos dados brutos. A massa de dados que está sendo produzida atualmente é gigantesca. E ela tende a crescer. Isso faz com que esses dados possam ser trabalhados e estudados para que novos conhecimentos sejam adquiridos. Entretanto, somente uma pequena quantidade desses dados são analisados e tornam-se informação estruturada para tomada de decisão dos mais diversos setores.(SAS, 2016)

Com o avanço tecnológico e a concorrência cada vez mais acirrada para conquistar um espaço maior de mercado em todos os segmentos, a análise de dados tornou-se essencial para que isso fosse possível. No setor de agronegócio não foi diferente. Na década de 90 iniciou-se a Agricultura de Precisão. O uso de dispositivos de posicionamento global (GPS), além de sensores em equipamentos de agricultura proporcionam maior precisão, devido a geração de dados em todo o processo, desde o plantio até a colheita.(DATASTORM, 2016).

Um estudo realizado por uma companhia que realiza análise de dados com foco em agricultura (Climate Corporation), descreve que utilizando-se de sensores, foi possível detectar anomalias em lavouras, além de monitorar a quantidade de nitrogênio em algumas plantas (essencial para seu desenvolvimento), fazendo com que fosse possível tomar medidas preventivas que diminuísse a perda e trouxesse um produto com maior qualidade.(DATASTORM, 2016).

A UNESP de Jaboticabal, em parceria com a Esalq/USP/Piracicaba, realizaram um estudo no solo paulista a partir de amostra de dados coletados. O intuito é mapear os micronutrientes presentes no solo, dividindo-se este em áreas em relação a características semelhantes.(REUTERS, 2016).

Grande parte desses estudos tornaram-se possíveis graças a técnicas de Big Data (que serão apresentadas nos próximos capítulos), para armazenamento, processamento e análise massiva de grande quantidades de dados. O uso dessas técnicas e ferramentas se tornaram popular e tem potencial para gerar ganhos de até 24 bilhões nos próximos 5 anos, segundo a consultoria McKinsey (empresa americana líder em consultoria empresarial). (REUTERS, 2016).

Segundo Ceper, sócio associado da McKinsey, “Nos últimos quinze anos, a média de aumento de rendimento agrícola no mundo é da ordem de 1% ao ano. Se conseguir 5% ao longo de cinco anos [com o uso de big data], já seria um excelente resultado, mas pode ser até mais”.

Com os pontos levantados, é possível identificar o potencial que a análise de dados pode trazer e se tornou essencial para apoio na tomada de decisão nos mais diversos setores, inclusive no setor agrícola. Técnicas e ferramentas de análise massiva de dados foram desenvolvidas para gerar conhecimento com intuito de aumentar a produtividade, com menor gasto e impacto ao meio ambiente.

O Centro de Estudos Avançados em Economia Aplicada (CEPEA) vem levantando dados do agronegócio desde 1995 sobre indicadores do agronegócio brasileiro. Alguns desses indicadores são, conforme Corrêa (2014), preços de grãos (como milho, soja e derivados), índices de inflação, preço de venda e compra, demanda de exportação, entre outras. Entretanto, essas informações coletadas não se encontram em uma forma estruturada e facilmente disponível para ser estudada e trabalhada para geração de informações relevantes para o negócio. Grande parte das informações está dividida em locais computacionais diferentes, com granularidades diversas, unidades de medidas distintas, informações divergentes ou em duplicidade, além de algumas delas não possuírem completude de conteúdo.

Dessa forma, o desafio é realizar um estudo de caso do CEPEA e propor possíveis soluções de Big Data que apoiem na solução dos problemas enfrentados hoje em seu cotidiano.

1.1. OBJETIVO

O objetivo geral deste trabalho é estudar problemas de análise de dados aplicados ao setor agronegócio brasileiro, tomando como base de estudos o Centro de Estudos (CEPEA) da Escola Superior de Agricultura "Luiz de Queiroz" da Universidade de São Paulo (ESALQ/USP), localizado na cidade de Piracicaba, interior do s São Paulo.

Os objetivos específicos desta monografia é identificar técnicas e ferramentas de *Big Data*, Mineração de Dados, Análises Preditivas, que possam ser aplicadas ao setor. Para isso será utilizado como base de estudos o CEPEA, onde serão exploradas técnicas e métodos que possam ser aplicadas, baseando-se em literatura científica sobre o assunto.

Além disso, a ideia é discutir possíveis aplicações de análise de dados que possam ser desenvolvidas futuramente para o setor.

1.2. MOTIVAÇÃO E JUSTIFICATIVA

Segundo Corrêa (2009), o setor de agronegócio (um dos principais setores da economia brasileira), possui diversas peculiaridades dentro da sua cadeia produtiva.

> Em abril de 2016, o Brasil exportou 10,08 milhões de toneladas de soja grão, segundo o Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC). Foi o maior volume mensal já embarcado.
> A média foi de 504,29 mil toneladas por dia, 32,5% a mais que o exportado diariamente em março último. Já na comparação com abril do ano passado, as exportações cresceram 54,0%.
> De janeiro a abril de 2016, o país embarcou 20,89 milhões de toneladas de soja grão. São 59,5% ou 7,79 milhões a mais este ano, em relação a 2015.
> A previsão da Companhia Nacional de Abastecimento (Conab) para a temporada 2015/2016 é de 55,35 milhões de toneladas exportadas, frente as 54,32 milhões em 2014/2015. (INTERACTIVE, 2016)

Através da análise dos números apresentados, é possível identificar o potencial do setor agrário no país, que tende a crescer. Dessa forma, a motivação para desenvolvimento da monografia é entender como esse setor (que é tão rico de informações e que é economicamente ativo e fundamental para nosso país) funciona, e quais ferramentas ainda podem ser aplicadas para apoiar na geração de

informações mais sólidas para tomada de decisões mais precisas e possíveis estudos na área.

Sendo assim, foi realizado um estudo sobre o CEPEA, que coleta dados desse gênero. Atualmente, diversos dados são coletados no Centro de Estudos, mas estes não possuem uma estrutura de *Big Data* que seja compatível para realização de estudos mais aprofundados.

Portanto, o que justifica o projeto é entender quais são as técnicas que são utilizadas para obtenção de dados do setor agrícola, como essas informações são armazenadas e estruturadas em seu banco de dados e propor possíveis soluções e ferramentas de *Big Data* que possam contribuir com o Centro para mineração dos dados disponíveis, além de técnicas de análises que possam ser aplicadas para gerar novos *insights.*

1.3. METODOLOGIA

A realização deste trabalho será conduzida através de estudos de caso do CEPEA. Inicialmente serão estudadas quais são as formas de coleta, armazenamento e apresentação dos dados atuais que são utilizadas no Centro. Para isso serão estudadas teses de Mestrado e Doutorado de alunos, relacionadas à representação de dados através de modelos de *Data Warehouse*, além de aplicações de mineração de dados para análise de preços do setor de agronegócio.

Foram realizadas reuniões semanais com o orientador para passagem de conhecimento referentes ao CEPEA e sobre conhecimentos gerais do setor, além de acompanhamento do Cronograma e escrita da monografia em si.

Inicialmente planejou-se realizar visitas ao CEPEA para conhecer o processo de coleta, análise e publicação dos indicadores do agronegócio *in loco*, para melhor entendimento e compreensão do ciclo de vida dos dados. Porém não foi possível, pois não houve liberação para se ausentar das atividades profissionais.

Em paralelo, foram levantadas quais ferramentas podem ser aplicadas ao cenário atual do Centro de Estudos, tais como Data Warehouse para armazenamento de dados, decomposição de Tucker para mineração de dados.

Em seguida, foram abordadas outras técnicas *Open Source* e alternativas relacionadas a *Big Data*, que poderiam ser aplicadas para análise dos indicadores do agronegócio estudados pelo CEPEA/USP.

1.4. CONTRIBUIÇÃO DO TRABALHO

O intuito do tema escolhido é adquirir conhecimento e desenvolvimento pessoal na área de agricultura. Por ser um setor que cresce exponencialmente no Brasil, fez com que me motivasse para conhecer mais a fundo sobre o tema.

A possibilidade de atrelar o setor de agronegócio com ferramentas atuais de análise de dados e processamento massivo de informações tornou-se de grande interesse levando em consideração o quão rico o setor é em relação a geração de dados.

Essa monografia também busca atingir alunos e pessoas que buscam adquirir conhecimento sobre os princípios de *Big Data* e como eles estão sendo aplicados nos dias atuais, e como essas ferramentas e técnicas podem ajudar o setor de agronegócio a possuir informações mais ricas para apoiá-los com conhecimento mais sólidos do negócio.

2. CONCEITOS DE *BIG DATA* E ANÁLISE PARA AGRONEGÓCIO

No Brasil, a agricultura e a pecuária vêm encontrando dificuldades em acompanhar o avanço tecnológico desde a produção até a comercialização dos produtos, o que fez com que a tecnologia viesse para agregar valor e continuar tornando o setor competitivo perante ao mercado. Para isso, tecnologias vêm sendo aplicadas no setor para torná-lo ainda mais produtivo. Serão apontadas as principais técnicas e tecnologias bastante difundidas atualmente no mercado tecnológico, realizando um breve descritivo e fazendo uma breve introdução sobre o CEPEA, onde será realizado o estudo de caso aplicando as técnicas descritas.

2.1. *BIG DATA*

É o termo que vem sendo empregado no âmbito tecnológico (e que permeia todas as áreas de negócio de alguma forma) para descrever a quantidade massiva de dados que vem sendo produzida diariamente pelos mais diversos equipamentos e dispositivos tecnológicos. Há grandes estudos em relação a esse conceito. Desenvolver técnicas de coleta, armazenamento, análise e apresentação desses dados com o intuito de tirar proveito dessas informações se tornou essencial (SAS, 2016).

Existem vários conceitos em relação a *Big Data.* Segundo a Amazon (2016), *Big Data* consiste no gerenciamento de Dados que possui algumas características em comum, chamada de “três Vs” de *Big Data*:

Volume: dados que variam desde Terabytes até Petabytes, (mas que podem crescer ainda mais), que necessitam ser coletados, armazenados e transformados em conhecimento para o negócio.

Variedade: dados que são coletados de variadas origens (Bancos de Dados Estruturados e Não Estruturados, Planilhas, Logs, Redes Sociais, ferramentas de CRM, Fontes Externas de Pesquisas como IBGE, entre outras) e também de variados formatos (Texto, Imagem, Sons, vídeos, entre outros).

Velocidade: os dados precisam ser processados em curtos períodos de tempo, para atender as necessidades do negócio em tempo real, ou o mais próximo disso.

Algumas outras empresas do ramo como a Stefanini, consideram ainda mais duas características para Big Data, que são:

Veracidade dos Dados: nem todas as informações que são coletadas são de fato úteis e gerarão informações ricas para o negócio. Dessa forma é necessário filtrar e analisar o que ajudará o negócio ou setor a entender melhor o comportamento de uma variável específica.

Valor: às informações coletadas precisam gerar valor para que seu propósito tenha sido atendido. Ou seja, o trabalho realizado de coleta e estudo, precisam gerar informações e responder perguntas que antes não eram possíveis, ou mesmo, complementar ou consolidar informações já existentes do negócio. (DATASTORM, 2016)

2.2. DATA WAREHOUSE

Técnicas de *Data Warehouse* passaram a fazer parte do agronegócio e auxiliar consideravelmente as análises do sistema e do processo, além de criar padrões e suportar grandes volumes de dados desde a plantação até a comercialização do produto.(CORRÊA, 2010)

Data Warehouse (DW) ou depósito de dados trata-se de uma arquitetura capaz de armazenar informações colhidas interna ou externamente através de históricos organizados, corrigidos e restaurados, a fim de colaborar nas tomadas de decisões, sem que afete as operações do sistema.(RICARDO, 2015)

Segundo Inmon (1997 *apud* Ricardo, 2015), que foi um dos primeiros no assunto, *Data Warehouse* é uma coleção de dados orientados por assunto, integrado, variável com o tempo e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisões. Na figura 1 temos a arquitetura de um *Data Warehouse*

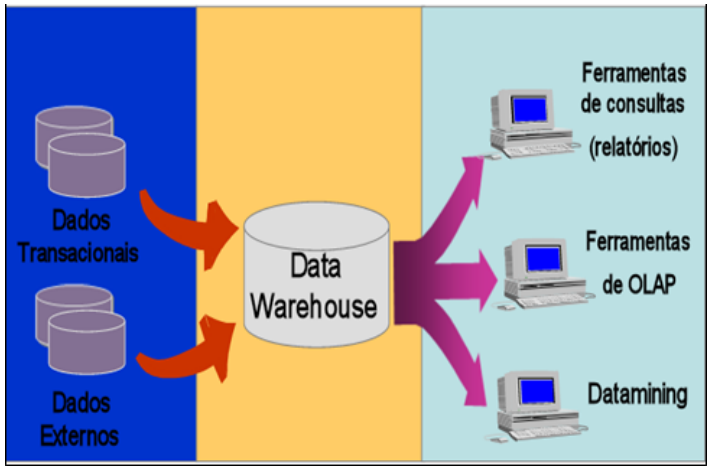


Figura 1: Arquitetura de um Datawarehouse. Fonte: Caiçara Júnior.

Um *Data Warehouse* também pode ser dividido em sub-conjuntos de dados conhecidos como *Data Marts*. O intuito é apoiar e subdividir as informações por setores ou departamentos específicos. Na figura 2 temos uma imagem representativa da estrutura de um *Data Mart*. (DATA MART, 2015)

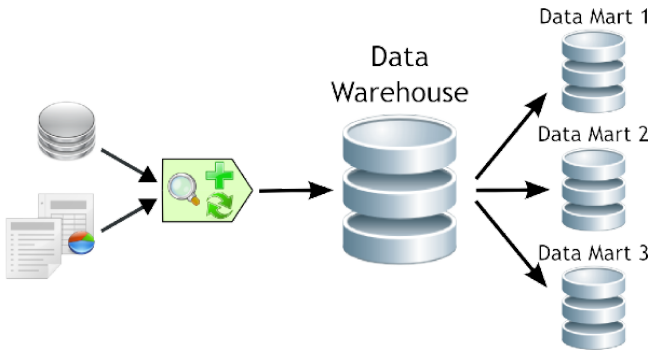


Figura 2: Disposição de um Data Mart. Fonte: Data Mart

As características *Data Warehouse* consiste em:

- Orientado por Temas: armazena informações específicas.
- Integrado: dados transformados em estados uniformes, isto é, em um mesmo assunto.
- Não-Volátil: Dados estáveis por tempos longos, geralmente de leitura.
- Variante no Tempo: Compara dados históricos em diversos períodos.

2.2.1. Modelos Dimensionais

Consiste na organização dos dados onde informações interagem na forma de um cubo, tornando as pesquisas simples de serem identificadas, melhorando assim, o desempenho das consultas (ELIAS, 2014). Podem ser:

- Modelo Estrela (*Star Schema*): Fica no centro a tabela de fatos, tendo ao redor as dimensionais como se fossem estrelas, facilitando desta forma as pesquisas e consultas.

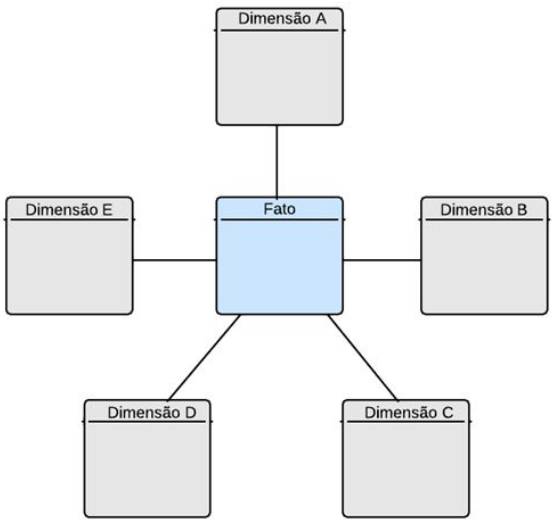


Figura 3 : Modelo Estrela. Fonte: Elias, 2014

- Modelo Constelação: há múltiplas tabelas fato com mesma dimensão. Exemplo na figura 4:

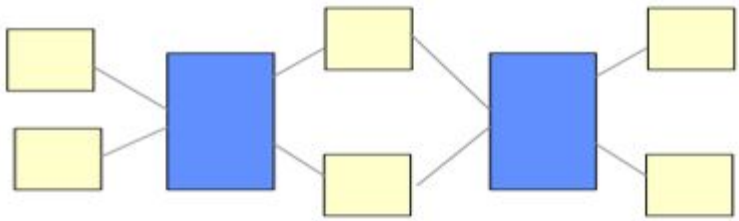


Figura 4: Modelo Constelação. Adaptado de Asterio K. Tanaka.

2.3. MINERAÇÃO DE DADOS

Mineração de dados consiste em analisar massas de dados com o intuito de buscar conhecimento.(CAMILO, 2009)

Através dos padrões identificados, é possível entender comportamentos e predizer alguma situação, apoiando na tomada de decisão.

A mineração de dados é uma tecnologia em ascensão e está permeando diversos setores como Medicina (fornecendo diagnósticos mais precisos), Segurança (identificando possíveis atividades terroristas), Tomada de Decisão (analisando quais dados são relevantes, em um determinado caso), RH (filtrando padrões em currículos para identificar competências) entre outros.

O processo de extração de conhecimento de uma massa de dados é conhecida pelo termo KDD (do inglês knowledge-discovery in databases) e possui algumas etapas conforme exibido na figura 5:

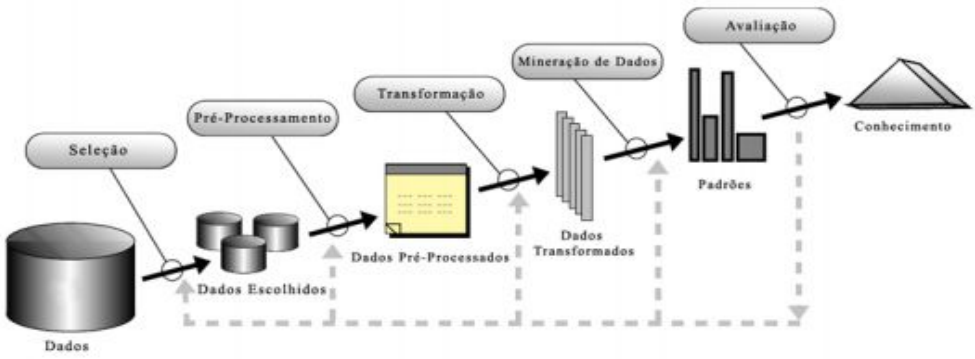


Figura 5: Figura representando o processo de KDD. Fonte: Camilo, 2009

2.3.1. Decomposição de Tucker

Segundo Bader e Kolda (2009 *apud* Corrêa, 2014), trata-se de uma técnica para realização de análise de dados multidimensionais, onde é realizado uma sumarização dos dados, gerando uma estrutura de forma reduzida para ser analisada.

O intuito dessa técnica é decompor as multidimensões de uma determinada amostra de dados, onde haverá um tensor principal que será multiplicado por matrizes de cada uma das dimensões da amostra, representando somente as informações mais relevantes da amostra inicial.

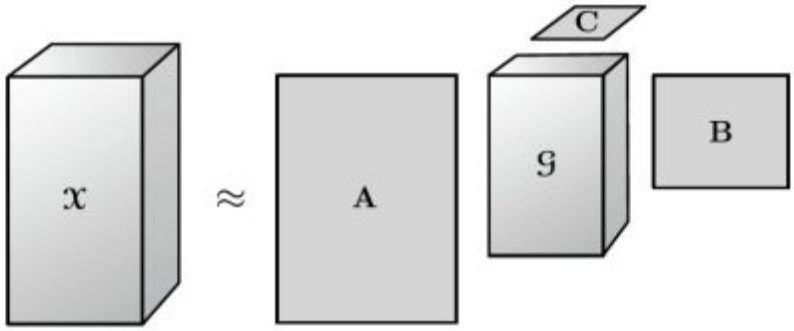


Figura 6: Decomposição de uma amostra de dados. Fonte: Kolda e Bader (2009)

2.4. ANÁLISE PREDITIVAS

São técnicas que analisam bases históricas de dados utilizando ferramentas e técnicas estatísticas com o intuito de prever alguma ação. (FACELI, K.*at al.* 2011). Essas análises se tornaram cada vez mais essenciais, devido à competição acirrada do mercado.

Segundo a SAS (2016), as análises preditivas estão sendo usadas atualmente para executar 5 tarefas diferentes:

- Identificar tendências
- Entender o comportamento de algo ou alguém
- Melhorar o desempenho
- Promover a tomada de decisão
- Prever comportamento

2.4.1. Aprendizado de Máquina

De acordo com Faceli *et al.* (2011), o aprendizado de máquina pode ser classificado em 4 modelos principais. Sendo eles:

a) Baseado em Distância: parte do princípio de que o dados precisam estar na mesma escala, ou seja, precisam estar normalizados. São diretamente influenciados pela proximidade dos dados que é essencial para que se realize as predições. Exemplo: classificadores k-vizinhos.

a.1) Classificadores K-vizinhos

A técnica aplicada é estimada em distância e a sua classificação é realizada em dois tempos (fase de treinamento e fase de testes). Os k–vizinhos surgiram com a proposta de eliminar ruídos causados na classificação dos tradicionais algoritmos, esta mudança é realizada no início da fase onde não se utiliza mais um vizinho único. Os dados são armazenados e expostos em uma única tabela com estimativas classificadas.(FACELI *et al.* 2011)

b) Métodos Probabilísticos: permite constatar teoremas para aplicações, combinações, álgebra, números e computação. São algoritmos que recolhe provas para prever dados, uma prévia que pode ser confiável ou não. Este método estuda fenômenos aleatórios, analisando se as informações obtidas podem ou não ser verdadeiros. Exemplo: Aprendizado Bayesiano

b.1) Aprendizado Bayesiano

Algoritmo usado para cálculos de probabilidade e fórmulas estatísticas para determinar a sua classificação. Trabalham com probabilidades de previsões e diagnósticos baseado no Teorema de Bayes. De posse dos dados obtidos determinam-se hipóteses, formando um conjunto de informações prováveis porém imprecisas.(FACELI *et al.* 2011)

c) Método Baseado em Procura: ocorre quando há a necessidade de encontrar possíveis soluções de problemas quanto à classificação, através de estratégias competentes, velozes e padronizada. Exemplo: Árvore de Decisão

c.1) Árvore de Decisão

Constitui-se em uma tabela com estrutura em forma de árvore. Tanto pode ser usada para tomar decisões como para obter conhecimentos futuros. São parecidas com a regra if-then, e normalmente são implantadas quando existem problemas nas classificações. (FACELI *et al.* 2011)
São realizados alguns testes e após a formação de um conjunto de perguntas classifica-se o caso particularmente.

d) Métodos de Otimização: utilizado quando há problemas para solucionar de alto valor ou mesmo de importância reduzida, e que se faça necessário criar condições favoráveis e obter a melhor decisão possível, ainda que existam certas restrições. Exemplo: Redes Neurais

d.1) Redes Neurais
O algoritmo Redes Neurais (RNA) consiste em uma técnica usada para detectar e reparar problemas através de regras. Possui unidades de processamento bem simples.(FACELI *et al.* 2011)
Na entrada são apresentados sinais e calcula-se um determinado limite, o erro ocorre quando na saída este limite for excedido.

e) Modelos Múltiplos: consiste na combinação de um ou mais métodos citados anteriormente para que seja possível atender a necessidade do negócio com maior eficácia.

2.4.2 - Séries Temporais

Chama-se Séries Temporais, um conjunto de dados coletados sequencialmente, que possui uma ordem cronológica entre eles. A principal utilidade de uma Série Temporal, é realizar estudos e análises em cima desse conjunto sequencial de dados, na tentativa de encontrar padrões em determinados períodos com o intuito de gerar *insights* e trazer conhecimento sobre o assunto ou área que está sendo analisado.(CORRÊA, 2014)

Segundo Morettin e Toloi (2004 *apud* Corrêa, 2014), existem alguns pontos que se destacam que tornam as Séries Temporais essenciais:

- Entender qual é o fator inicial que inicia uma série temporal. Entender como uma série temporal se inicia, facilita no entendimento dos dados obtidos e ajuda na busca de padrões nos dados que estão sendo analisados.
- Através da análise histórica da Série Temporal é possível predizer valores futuros. Períodos semelhantes podem apresentar padrões semelhantes, fazendo com que seja possível predizer como se comportará determinada variável em um período seguinte, por exemplo
- Identificar períodos relevantes. Esse ponto é de se destacar pois analisando uma série temporal é possível identificar onde há picos positivos ou negativos na variável que está sendo analisada, dando a possibilidade de extrair informações através dessa análise para uma tomada de decisão.

Em relação ao contexto do Agronegócio, entender o histórico de oscilações de preços de produtos, histórico de índices de inflação, preço de compra e venda dos mais diversos produtos, torna-se essencial para predizer qual será o comportamento futuro dessas variáveis, apoiando os pesquisadores e pessoas interessadas ao negócio em identificar qual a probabilidade de determinado evento ocorrer.

2.5. AGRONEGÓCIO

O agronegócio tem passado por diversas mudanças, tornando-se muito mais complexo. Antigamente suas atividades se limitavam a propriedade rural, com atividades não tão estruturadas e amplas. Atualmente, as atividades percorrem um extenso ciclo de vida, desde a produção de insumos, passando por atividades de armazenamento, processamento e sua atividade final de distribuição dos produtos agrícolas, além de permear outras áreas de negócio, como setores financeiros, marketing, logística, entre outros. (Gestão no Campo, 2016)

De forma simplificada, agronegócio consiste no conjunto de atividades envolvendo a cadeia produtiva agrícola e pecuária.

A figura 7 exemplifica o Ciclo de Vida Agrícola, desde a produção até o produto final para o consumidor:

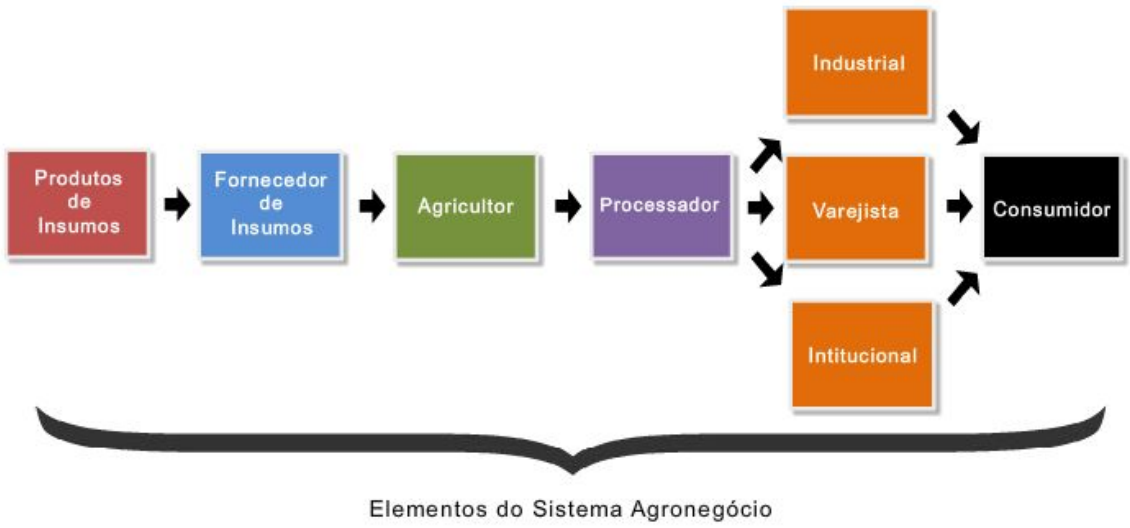


Figura 7. Elementos do Sistema Agronegócio. Fonte: Gestão no Campo, 2016

Segundo Lopes (2015), o agronegócio vem crescendo de maneira surpreendente nos últimos anos, não somente no Brasil mas em todo o mundo. Para acompanhar tal procedimento houve a necessidade de implementar uma tecnologia que conduzisse os processos agrícolas com a mesma agilidade.

Para obter uma agricultura de precisão (AP) foram investidos vários recursos, porém o mais importante trata-se das pesquisas em torno do assunto, que instituições, órgãos públicos e outros, apresentaram nas mais diversas informações. Dentre estes dados foram estudado os custos de maneira geral, os impactos ambientais e os meios mais viáveis tanto para empresários como para agricultores.

O volume de dados adquirido é intenso, tanto na sua diversidade quanto na qualidade e muitas vezes dificultam no processo de tomada de decisões.

A indústria de manufatura através de um Sistema de Informação Gerencial (Management Information System) oferece soluções dentre as quais ferramentas como o ERP (Enterprise Resource Planning), contribui em ações relacionadas à gestão, além do suporte de TI e comunicação.

Dentre as ferramentas tecnológicas utilizadas para levar informações de dados com mais presteza e agilidade na lavoura são:

- Sensores e dispositivos para armazenamento e processamentos;
- Sistemas de atuador para comunicação sem fio (Zigbee and WiFi), e protocolo ISOBUS (ISO 2013).
- Aplicações em taxas variáveis (AVT)

Ainda assim, existem muitos problemas a serem enfrentados nos processos agrícolas, como exemplo, o clima, o solo, impactos ambientais, entre outros que em uma linha de produção não aconteceria.

O Brasil abraçou a causa no que se refere a tecnologia para qualificar e quantificar o volume de dados do setor do agronegócio e tornou-se o pioneiro na exportação e produção de vários itens, tais como, café, laranja e açúcar e vem se consagrando em segundo na exportação de outros.(LOPES 2015)

Para a economia brasileira este acontecimento é extremamente significativo, pois além de representar um grande desenvolvimento, apresenta 26% do PIB, uma

agradável importância na balança comercial e ainda traz estratégias satisfatórias nos setores econômicos, ambientais e sociais.

2.5.1. CEPEA

Atualmente, um dos principais fornecedores de dados relacionados ao agronegócio é o Centro de Estudos Avançados em Economia Aplicada da USP (CEPEA).

> "O Centro de Estudos Avançados em Economia Aplicada (CEPEA) é parte do Departamento de Economia, Administração e Sociologia (DEAS) da Esalq/USP. Foi criado por docentes deste Departamento com a finalidade de atender mais eficientemente às demandas por estudos, pesquisas e informação nas áreas da economia, administração e ciências sociais[...].
> Através de pesquisas diárias sobre as principais cadeias de matérias-primas agropecuárias e seus derivados, o Cepea elabora indicadores de preços de produtos, insumos e de serviços (como frete) que buscam refletir com precisão o movimento do mercado físico." (Cepea 2016).

Conforme apresentado acima, o CEPEA coleta diariamente grande quantidade de dados relacionados ao meio agropecuário, com o intuito de entender de forma mais eficiente a movimentação do mercado para fornecer informações mais sólidas e precisas para produtores, cooperativas, e varejistas em geral.

Os principais dados do agronegócio que são coletados, segundo CEPEA (2016), são:

- Índice de Preços de Exportação
- Índice de Cálculos de Volume e Preços
- Índice de Câmbio Efetivo
- Índice de Atratividade (Relação de Câmbio X Preço Externo)
- Índice de Volume (Volume de Exportação do Agronegócio (medido pelo IVE-Agro-Cepea)
- PIB

As atividades do CEPEA abrangem estudos, pesquisas e difusão de informações através de variados meios e são estruturados segundo cadeias produtivas. Utiliza tanto informações nacionais como internacionais trazidas de universidades,

instituições, órgãos públicos e privados. Os dados coletados colaboram para todas as categorias envolvidas no assunto, sejam pesquisadores, agricultores, empresas, etc. Disponibiliza ao mercado informações atualizadas, e ainda realiza um estudo histórico com os dados acumulados há anos.

Segundo o CEPEA (2016), suas pesquisas contam com vários órgãos no mercado, tanto nacionais como internacionais. Alguns delas são:

- CNPq,
- Fapesp,
- BM & Bovespa,
- Agência do Estado
- Reuters
- Centro de Pesquisas da USP
- CNA - Confederação da Agricultura e Pecuária do Brasil,
- Ministério do Meio Ambiente,
- Embrapa - Empresa Brasileira de Pesquisa Agropecuária
- Monsanto

2.6. TRABALHOS RELACIONADOS

Segundo BRONSON *et al*. (2016) a agricultura está passando por uma transformação digital e a aplicação da abordagem de *Big Data* mostrou ser uma grande arma para acompanhar esta revolução.

Tanto que, ferramentas de análise e coleta de dados estão sendo implantadas na agricultura no intuito de colaborar com novos conhecimentos, um maior desenvolvimento e análise de informações tanto para pequenos agricultores como para grandes corporações.

Os equipamentos de precisão são notoriamente utilizados na tomada de decisão sobre exportações, e os processos de *Big Data* tornaram-se essenciais para o desenvolvimento dos produtos.

Empresas passaram a buscar soluções desse tipo, como atualmente acontece com a Agriculture Canada's conforme explicado por Bronson *et al.* (2016) que através de processos de *Big Data*, faz uso de app digital que fornece ao agricultor práticas que detectam de forma eficiente futuras produções, além de colaborar no desempenho individual interagindo-os com as informações de que necessita.

A empresa de Tratores John Deere acompanha a evolução e é considerada uma das maiores indústrias de ferramentas agrícolas. Ela utiliza sensores em seus maquinários para adquirir dados extensos sobre solo e plantio e o *Big Data* auxilia-os na hora de filtrar as informações. BRONSON *et al.* (2016)

A empresa Monsanto Corporation no passado teve problemas com as restrições na tecnologia, por isso criou a ferramenta digital IFS (Integrated Farming Systems) ou Sistemas Integrados de Produção para trazer maior produtividade e rentabilidade, já que proporciona todos os dados e informações sobre solo e clima, e ainda detecta plantas daninhas. BRONSON *et al.* (2016)

A Sysmos Corporation criou a ferramenta Heartbeat, que através de dados da mídia auxilia a conhecer as prioridades e interesses dos consumidores, além de um aplicativo Agroclimate Reporter (AR) que divulga dados do clima. BRONSON *et al.* (2016)

Diante da repercussão do Big Data, muitas dúvidas surgiram. A compreensão dos dados obtidos, as vantagens reais alcançadas e à quem estes resultados têm beneficiado, são alguns dos pontos que estão sendo estudados para compreender melhor seu benefício.

De acordo com Carolan (2016) a agricultura cresceu demasiadamente nos últimos anos, assim a quantidade de dados e informação aumentaram ao ponto de se fazer necessário um processo que pudesse suportar tais dados e que facilitasse o setor agrícola.

Surge então o *Big Data*, que além de suprir tal necessidade ainda trouxe uma agricultura de precisão, com redução do consumo de água, aumento nas safras, menores impactos ambientais e a qualidade de todo processo.

Big data e a Agricultura de Precisão trazem informações de como plantar, colher, comercializar e o tratamento de plantações diferentes por metro quadrado de terra, além dos devidos cuidados com pragas.

Muitas empresas resolveram então, investir neste ramo, que é portanto o que há de mais moderno no agronegócio.

A empresa Deere investiu em sensores e equipamentos de conexão a fim de integrar todos os interessados com as informações adquiridas.

Já a empresa Monsanto adquiriu a Climate Corporation que trabalha com uma plataforma de software capaz de trazer dados e soluções sobre o solo e clima de cada plantação e proporciona mapear toda a zona de cultivo.

Outras empresas como a Foodscapes criaram grupos e técnicas para usar a agricultura de precisão em sistemas alimentar.

Portanto o *Big Data* e a Agricultura de Precisão tornaram-se o sistema mais rentável nos últimos anos e esta inovação trouxe à agricultura uma tecnologia capaz de agregar informações desde o plantio até a exportação, de maneira segura.

À cada ano esta tecnologia vem se destacando e fazendo com que agricultores, administradores e empresas procurem projetos novos capazes de oferecer o que há de mais modernos, confiável, de qualidade e que o lucro seja certo.

Segundo Antunes (2015), o Brasil teve um excelente desenvolvimento no setor agrícola, tornando-se o maior exportador de açúcar e a colocação de segundo lugar como produtor de etanol, fato que gerou 35% na balança comercial e o setor chegou a 6% de empregos. A responsabilidade de 90% dessa produção vem do Centro-Sul do Brasil.

O ciclo de desenvolvimento da cana-de-açúcar pode ser entre 12 ou 18 meses, isto vai depender de alguns fatores, como o clima, da região, da época ou do solo, sendo que, após a primeira colheita geralmente seu ciclo passa a ser de 12 meses, e após cortada pode ser replantada outras vezes.

No Estado de São Paulo, predominam os ciclos de produção da cana de ano e meio, plantada de janeiro a maio, e da cana de ano, plantada de setembro a dezembro, com a colheita estendendo-se de abril a dezembro (ANJOS & FIGUEIREDO, 2010 *apud* ANTUNES, 2015).

Atualmente existem satélites de monitoramento ambiental que possibilitam através de sensores orbitais estudar e analisar todo o desenvolvimento da processo da planta, área de cultivo, análises do solo, perfis temporais, comportamento espectral ao longo do tempos, entre outros dados.

À exemplo, o sensor MODIS, que através de detectores são capazes de cobrir e mapear toda a área e todo o processo agrícola com imagens com resolução espacial, qualidade radiométrica, baixo custo de aquisição.

O sensor MODIS (Moderate Resolution Imaging Spectroradiometer), a bordo das plataformas orbitais do programa internacional EOS (Earth Observing System), liderado pela NASA (National Aeronautics and Space Administration), tem gerado dados processados para estudos globais da vegetação. O satélite TERRA foi lançado em dezembro de 1999 e tem passagem pelo Equador às 10h30 (horário local), em órbita descendente (SOARES *et al*., 2007 *apud* ANTUNES, 2015).

Este trabalho através de perfil temporal de dados busca avaliar todo o processo da cana-de-açúcar do Estado de São Paulo, fazendo uso de informações do sensor MODIS referente a safras de 2004/2005 a 2011/2012.

A empresa Raízen disponibiliza dados de algumas áreas de cultivo em séries temporais do Estado de São Paulo. Estas áreas são terras arrendadas e algumas unidades produtoras de cana-de-açúcar, onde foi verificado se com a presença de

ruídos do sensor MODIS pode-se alterar o reconhecimento do padrão espectral da produção das safras.

A Transformada de Wavelet foi utilizada para análise de ruído em relação ao desenvolvimento da vegetação através de uma curva de remoção de alta frequência destacando as de baixa, a fim de moderar os perfis temporais.(MARTÍNEZ & GILABERT, 2009 *apud* ANTUNES, 2015)

O trabalho mostrou a importância em adquirir padrões para o monitoramento temporal da Cana de Açúcar em relação aos eventos periódicos, assim como a percepção do ambiente e as variações interanuais.

Com a utilização da Transformada de Wavelet Daubechies 8 aplicada à série temporal do EVI2 do MODIS determinamos que:

- trata-se de uma técnica robusta,
- é eliminador de ruídos,
- traz conhecimento geral da série temporal da cana-de-açúcar em relação ao desenvolvimento,
- os perfis temporais suavizados do EVI2 identificou dados de mudança de solo e cobertura da terra, de acordo com eventos anuais e todas as fases do plantio à colheita da cana-de-açúcar.

Após a contextualização demonstrada acima, onde é possível entender a grande aplicação de técnicas e tecnologias aplicadas ao setor de agronegócio, a abordagem principal dessa monografia será focar no estudo de caso do CEPEA. Abaixo será detalhado o mecanismo atual de coleta e apresentação dos dados de preços do agronegócio praticado pelo CEPEA.

3. ESTRUTURA E ANÁLISE DE INDICADORES PELO CEPEA

3.1. VISÃO GERAL

O Centro de Estudos Avançados em Economia Aplicada da Escola Superior de Agricultura “Luiz de Queiroz” da Universidade de São Paulo foi criado oficialmente em 1982 por docentes e pesquisadores com o objetivo principal de apoiar o setor de agronegócio com o levantamento de dados relevantes de forma estruturada e organizada para apoiar o setor e instituições (tanto de áreas públicas como privadas) carentes de informações.(CEPEA, 2016)

Professores, pesquisadores e alunos, trabalham em conjunto diariamente com o intuito de levantar informações econômicas relacionadas direta e indiretamente ao setor de agronegócio, estruturar a informação levantada e comunicar esses dados obtidos nos mais diversos meios, entre eles Boletins, Revistas, Informativos e Indicadores, Jornalismo de TV e Rádio, além da própria Internet.

O CEPEA também identifica pontos de melhoria para otimizar a produção dos fornecedores e também apontar novos espaços que possam ser preenchidos por cooperativas, agricultores e produtores em geral, buscando aprimorar a produção.

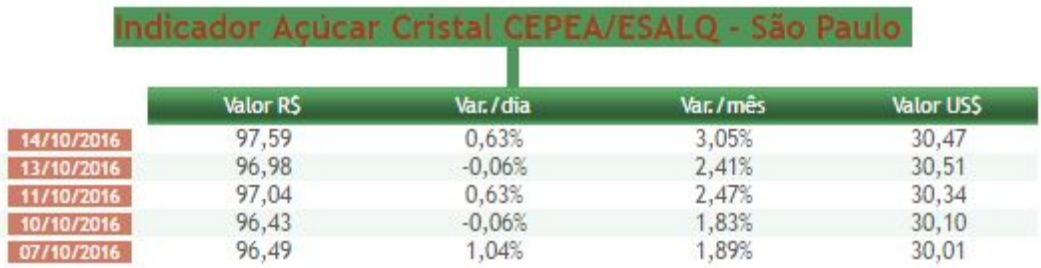
Alguns dos produtos são realizados estudos no CEPEA. Esses produtos podem ser observados na Tabela 1:

Agricultura	Pecuária	Outros
Açúcar	Bezerro	Etanol
Algodão	Boi	PIB Agronegócio
Arroz	Frango	Economia Ambiental
Café	Suíno	Economia Florestal
Citros		Economia Internacional
Ovos		Administração Rural
Leite		
Mandioca		
Milho		
Soja		
Trigo		

Tabela 1: Áreas de Pesquisa. Adaptado: CEPEA

Todos os dias, são coletados e disponibilizados indicadores. Entre eles podemos citar informações relacionadas ao Produto Interno Bruto (PIB), preço dos produtos que estão sendo praticados no setor, frete dos produtos, além de comparações com o mercado internacional, para fornecer informações ainda mais detalhadas.

Na figura 8, temos alguns exemplos de indicadores que são fornecidos no site do CEPEA, com informações do agronegócio. Pode-se observar na Figura 8, temos dados como a média dos valores praticados tanto em Real como em Dólar em cada um dos dias que foram obtidos os dados, além da variação por dia e mês do produto.



Indicador Açúcar Cristal CEPEA/ESALQ - São Paulo

	Valor R$	Var./dia	Var./mês	Valor US$
14/10/2016	97,59	0,63%	3,05%	30,47
13/10/2016	96,98	-0,06%	2,41%	30,51
11/10/2016	97,04	0,63%	2,47%	30,34
10/10/2016	96,43	-0,06%	1,83%	30,10
07/10/2016	96,49	1,04%	1,89%	30,01

Figura 8: Indicador de Açúcar Cristal. Fonte: CEPEA

Os estudos também permeiam o bem-estar e qualidade de vida no meio agropecuário e na sociedade como um todo, abordando assuntos como saúde e educação no meio rural.

3.2. COLETA E ESTRUTURAÇÃO DE DADOS

Atualmente, para que seja possível realizar os estudos relacionados a área e disponibilizar os indicadores para os órgãos envolvidos, o CEPEA realiza um levantamento de dados realizado em campo, obtendo informações diretamente com pecuaristas, agricultores, cooperativas, além de associações chamadas de mistas (onde há compradores e vendedores de determinado produto ou serviço).

São levantados quais são os valores praticados pelos fornecedores de cada um dos produtos que são áreas de estudo do CEPEA. Esse levantamento é realizado com produtores dos principais Estados do país, para fornecer um valor mais próximo do real.

O fornecimento dessas informações para o agronegócio pode variar por períodos. Dependendo do mercado, essas informações podem ser geradas diariamente como é o caso de produtos de agricultura como Açúcar, Arroz, Soja, entre outros; ou podem ser fornecidas semanalmente como é o caso do valor praticado do Etanol.

Os dados individuais não são disponibilizados. São calculadas médias dos valores praticados pelo produto e assim essa informação é divulgada. Caso o produto a ser analisado possua somente uma fonte de informação, ou seja, caso a informação que

foi obtida seja de um fornecedor somente, essa informação não é disponibilizada ao público, com o intuito de preservar os valores praticados.

Atualmente, o CEPEA fornece informações mais relacionadas a preços praticados no mercado, cabendo a outros órgãos como BM&F e IPEA, fornecer informações mais abrangentes cruzando informações com outras bases de dados. Esses dados que são coletados periodicamente e armazenados, geram séries temporais, utilizadas para estudos e análises de dados.

Considerando a massa de dados coletada diariamente, torna-se necessária a gestão de dados eficiente, para que seja possível a extração de informações relevantes.

Para fazer as análises, além dos dados primários gerados pelo CEPEA, são necessários mapear todas as fontes de dados externas, para que seja possível estruturar a informação de uma forma coerente e que apoie na tomada de decisão. Também se faz necessário identificar na base de dados que é coletada, quais são os dados relevantes para análise histórica, e sua granularidade de informação para que seja estruturado de uma forma que atenda as necessidades dos consumidores da informação.

Após mapeado quais dados serão relevantes para o contexto, é realizada a modelagem dos dados, contendo as mesmas unidades de medida e granularidade semelhantes para que seja possível realizar uma análise de mesmo nível. Conforme citado inicialmente, no capítulo de introdução, o modelo aplicado é o Estrela, no qual as informações detalhadas de medidas, passíveis de análise, ficam estruturadas em uma tabela principal chamada de dimensão, e as tabelas de apoio, que são utilizadas para intersecção de todas as informações para geração dos dados históricos, são chamadas de dimensão.

Após realizada a modelagem das tabelas, esses dados precisam ser armazenados para que possam ser disponibilizados para os usuários interessados. Para isso faz-se uso de um *Data Warehouse*, através de processos de *Extract, Transformation and Load* (ETL), é possível centralizar os dados das mais diversas fontes para que seja possível integrar as informações, e alimentar as tabelas que foram

estruturadas. A visualização desses dados é possível através de ferramentas de *Online Analytical Processing (*OLAP), que permitem analisar os dados de uma forma mais específica ou macro, cabendo ao pesquisador definir período a ser analisado, a granularidade da informação e qual segmento deseja realizar o estudo.

Dessa forma é possível obter as informações desejadas com muito mais agilidade, pois o cruzamento das informações de bases distintas é realizado de forma automatizada, alimentado automaticamente pelo processo de ETL, e de uma forma prática, pois as informações relevantes estarão estruturadas de uma forma padronizada e centralizada.

3.3. ANÁLISE HISTÓRICA ATRAVÉS DE *DATA MINING*

Após a estruturação dos dados, faz-se necessário entender como as variáveis dos produtos que são estudados pelo CEPEA relacionam-se e como elas se modificam na trajetória do tempo. Em um segundo momento, também é importante realizar comparações entre períodos semelhantes nas variáveis a serem analisadas, em busca de padrões entre elas, para que seja possível entender qual fato que ocorre no período que faz com que o comportamento das variáveis se assemelham. Essa análise poderá trazer informações positivas ou negativas, dependendo da situação.

Após identificação, técnicas de redução de dimensões como a Decomposição de Tucker são adequadas, pois garantem com que somente a informação que será de fato útil seja analisada. O intuito dessa técnica é garantir a redução das informações sem que tenhamos perda de informação relevante para extração de *insights.*

Para isso, utiliza-se uma técnica chamada *Piecewise Aggregate Approximation* (PAA). Essa técnica consiste na divisão dos dados principais em conjuntos. Nesses conjuntos de dados será aplicada uma média para que se tenha a mesma ideia do conjunto original, mas em quantidades menores para facilitar o processamento das informações.(Li *et al*, 2003 *apud* CORREA, 2014).

Após realizada a sumarização dos dados, é necessária realizar a discretização dos dados. Uma das técnicas mais utilizadas para realização desse processo é o

método SAX (Corrêa, 2014), onde de forma sintetizada, consiste em atribuir símbolos ou letras do alfabeto, em um conjunto aplicando limites entre eles, gerando níveis de cortes. Esses níveis de cortes serão utilizados para busca de padrões entre as séries, onde serão encontrados níveis de cortes com as mesmas características em comum.

Após realizado o pré-processamento dos dados, é necessário a aplicação de técnicas que busquem padrões nas séries temporais dos dados coletados do CEPEA. As que serão abordadas, serão Técnicas de Distância Euclidiana, MINDIST e Dynamic Time Warping (DTW).(CORRÊA, 2014)

- Técnica de Distância Euclidiana, consiste na comparação de 2 conjuntos de informações de mesmo tamanho para identificar similaridades entre elas. Caso a comparação entre as posições dos conjuntos seja igual, exibirá o valor 0. Quanto maior o número inteiro resultante, mais diferença há entre os conjuntos de dados.
- Técnica de Distância MINDIST, consiste em identificar a distância entre de cada caractere presente em cada uma das séries
- Dynamic Time Warping (DTW), técnica muito utilizada para identificar possíveis deslocamentos que podem ocorrer nos eixos das séries temporais estudadas. Diferentemente da técnica de distância euclidiana que consiste na comparação de um par similar, a DTW compara pares que podem ter se deslocado no decorrer do tempo em uma série temporal. Caso isso tenha ocorrido, esse deslocamento só será possível de ser identificado utilizando DTW.

Baseado no CEPEA como estudo de caso, as dimensões que fazem mais sentido em aplicar técnicas de mineração de dados são: Produtos, Tempo e Mercado. Entende-se como Produtos, qualquer item analisado pelo CEPEA que é realizado algum tipo de indicador, seja ele do ramo da Agricultura ou de Agropecuária. Entre eles, temos Leite, Trigo, Milho, Boi, Suíno, entre outros. Considera-se Dimensão tempo, o período em que deverá ser analisado para obter indicadores. Essa comparação pode ser realizada entre meses, trimestres, semestres, anos

semelhantes. Mercado consiste no local praticado dos produtos. Atualmente o CEPEA obtém dados de mercados internacionais como a Bolsa de Valores de Chicago por exemplo.

De forma sumarizada, as técnicas de Data Warehouse são aplicadas para estruturar os dados diariamente, onde são extraídas, transformadas e carregadas através de um processo de ETL. As informações estruturadas em um cubo permitem a visualização de dados de uma forma multidimensional, sendo possível relacionar diversas variáveis.

As técnicas de mineração de dados buscam entender o comportamento das séries temporais de produtos do CEPEA, onde através da Decomposição de Tucker e SAX, os dados originais passam por um processo de redução de dimensionalidade e discretização. Em seguida, técnicas de obtenção de padrões são utilizadas (como a distância euclidiana e MINDIST), que buscam comparar períodos e encontrar informações semelhantes.

3.4. OUTRAS ANÁLISES APLICADAS AOS INDICADORES

Baseando-se nos estudos das técnicas citadas anteriormente, a seguir são apresentados estudos sobre os indicadores do agronegócio. As análises mostraram-se efetivas, pois após a aplicação de técnicas de mineração de dados, foi possível identificar alguns relacionamentos entre as dimensões estudadas.

Para análise realizada por Corrêa (2014), foram criados 6 cubos, onde cada um deles representa um ano de análise (2007-2012). Em seguida, para entender o comportamento das dimensões Produto, Mercado e Tempo, foi realizado um estudo para identificar quais variáveis se relacionam melhor. Sendo assim, cada uma das variáveis de cada uma das dimensões foram relacionadas e foram divididas em fatores, onde fator será o menor número de variáveis possíveis em uma dimensão, que possam ter maior índice de explicação dos resultados. (CORRÊA, 2014).

Seguindo essa linha, o fator que teve maior índice de qualidade foram a relação 2, 3, 3 ou seja, 2 fatores para produtos, 3 para mercados e 3 para tempo, que conseguem

ter um grau de explicabilidade (ou seja, variáveis que quando relacionadas consegue descrever o comportamento mais próximo do real) de 90%. (CORRÊA, 2014). O gráfico de análise realizado por Corrêa (2014), pode ser visto na figura 9:

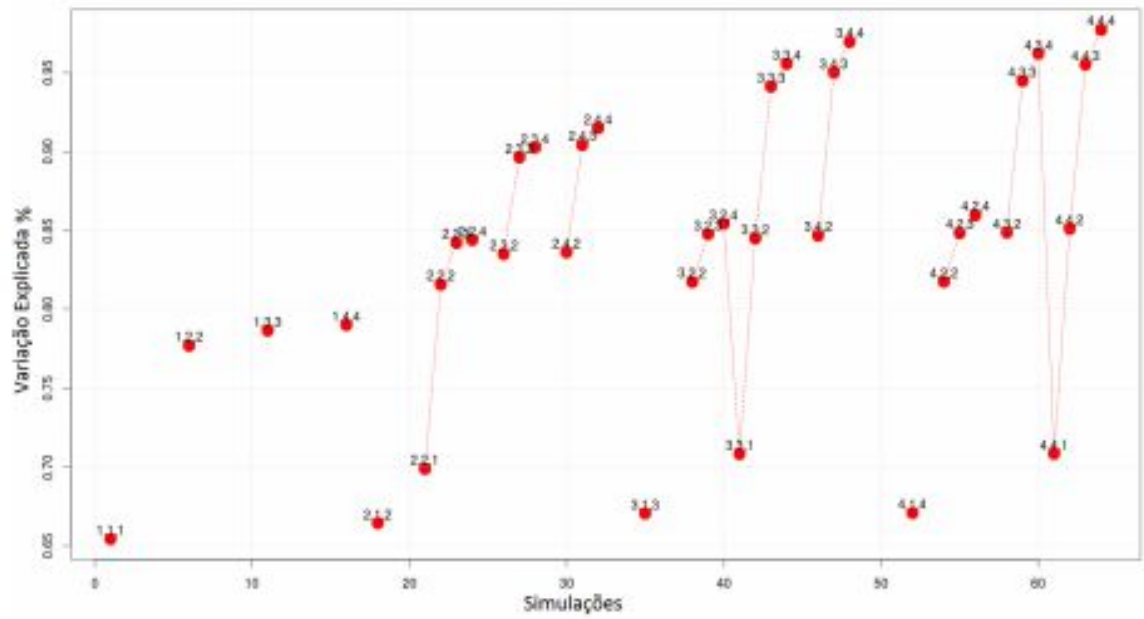


Figura 9: Análise de Combinações para um dos cubos. Fonte: Corrêa (2014)

Para cada um dos anos, foi realizado esse relacionamento, no qual foi possível identificar os fatores que melhor se relacionam em cada um dos anos. O resultado final pode ser observado abaixo:

Conjunto de dados Cubos / Ano	Fatores (P,Q,R)	Variabilidade Explicada
2007	2,3,3	90%
2008	3,3,3	88%
2009	3,3,3	85%
2010	3,3,3	88%
2011	3,3,3	86%
2012	2,3,3	89%

Figura 10: Análise de fatores em cada um dos cubos. Corrêa (2014)

Em seguida, é possível identificar o comportamento dos produtos analisados através dos gráficos das trajetórias temporais. Baseado no estudo realizado por Corrêa (2014), foram feitas análises dos comportamentos dos produtos, baseado nos fatores que obtiveram melhor explicabilidade conforme figuras 9 e 10. Sendo assim, é possível identificar alguns relacionamentos similares entre os produtos e alguns comportamentos. Entre eles, soja e farelo, que apresentaram variações semelhantes e que milho apresentou uma inversão em sua oscilação graças a competitividade que ocorre entre as culturas. Essa análise pode ser obtida através da figura 11:

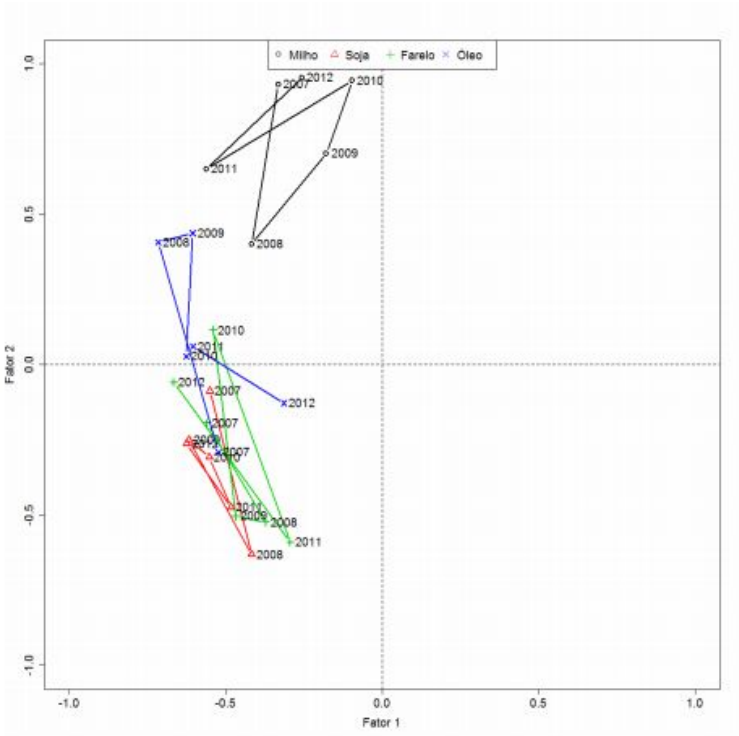


Figura 11: Trajetória dos produtos analisados do CEPEA. Corrêa (2014)

Conforme analisado por Corrêa (2014), é possível identificar que soja e farelo destacados respectivamente com as cores vermelho e verde tiveram um comportamento semelhante e que milho teve uma oscilação temporal. Também é possível identificar que tanto soja como derivados tiveram um comportamento tendendo negativamente.

Também foi possível identificar na dimensão “Mercado” para o qual foram realizados estudos entre a Bolsa de Valores de Chicago e o Mercado Interno Brasileiro

(CEPEA), os quais não apresentaram trajetórias inter-relacionadas, apesar de que a trajetória de preços de Chicago e CEPEA, apresentam semelhanças.(CORRÊA, 2014)

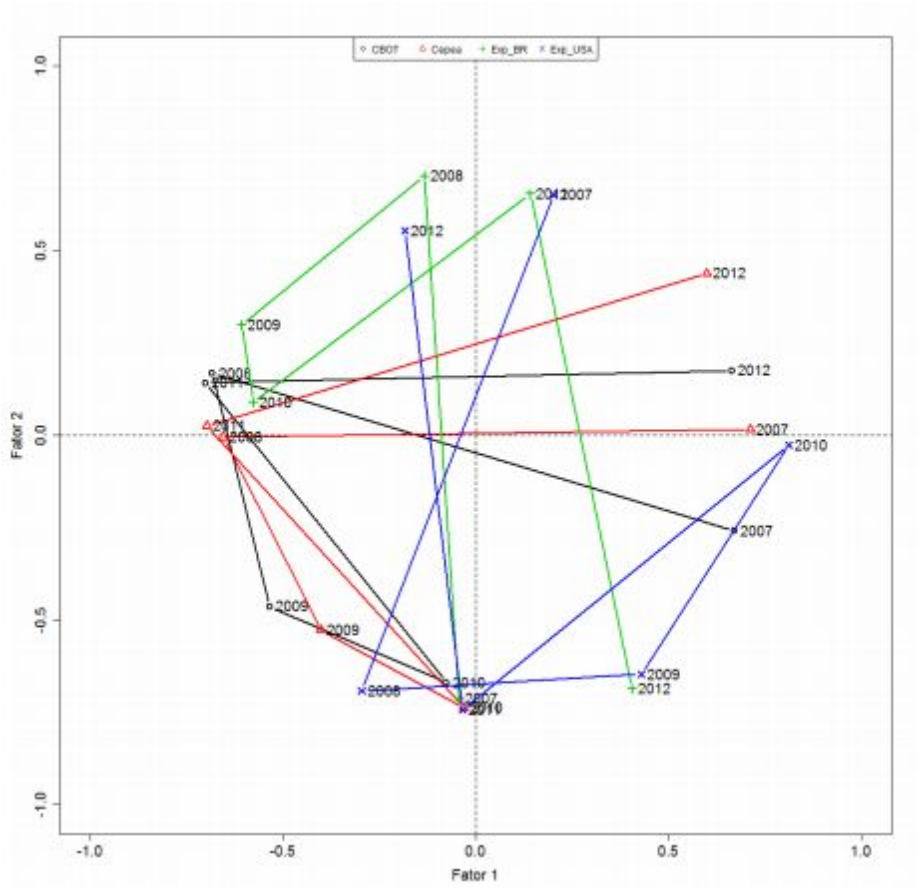


Figura 12: Análise de Trajetória de Mercados. Corrêa (2014)

No gráfico é apontado o comportamento do mercado brasileiro (CEPEA) e a Bolsa de Valores de Chicago. Foi observado por Corrêa (2014), que a variação dos preços tanto do CBOT como do CEPEA, apresentam comportamentos semelhantes como podemos observar nas linhas destacadas como preto e vermelho. É possível identificar que há crescimento e declínio entre esses dois mercados em anos semelhantes, e que a variação ocorre inversamente ao se tratar de exportações quando analisado as linhas verde e azul.

Outras análises obtidas por Corrêa (2014) é a variação do preço praticado de produtos como Milho e Soja, se comparado com outros mercados, como a Bolsa de

Valores de Chicago, onde é possível acompanhar a trajetória semelhante dos produtos em ambos os mercados, conforme figura 13:

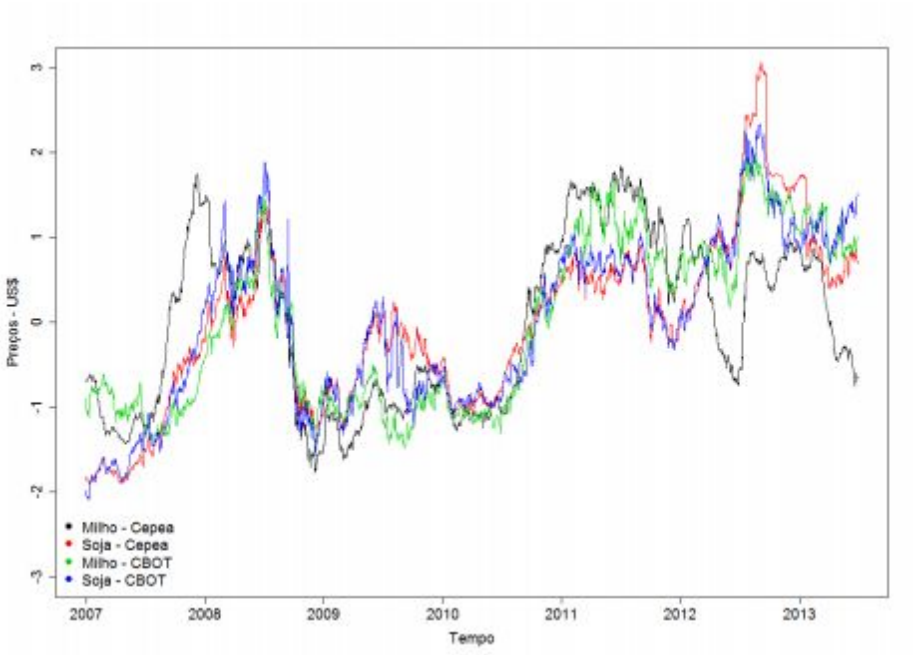


Figura 13: Comparação CEPEA X Bolsa de Valores de Chicago. Fonte: Corrêa (2014)

No gráfico é possível observar que os produtos milho e soja tanto no CEPEA como Bolsa de Valores de Chicago (CBOT) tiveram seus preços oscilando de forma semelhante .

Esses estudos se mostram eficazes e úteis, pois apoiam estudiosos a entender de forma mais detalhada, como os produtos praticados pelo CEPEA se comportam, quais produtos possuem um comportamento semelhante entre si, além de entender os mercados e como eles se relacionam. (CORRÊA, 2014)

Dessa forma, é possível entender baseando-se nos números obtidos por Corrêa (2014), comportamentos como as trajetórias de produtos como milho e soja que se apresentaram distintas e inversas nas análise temporal, relacionamento de preços praticados entre a bolsa de Valores de Chicago e o Mercado Interno Brasileiro, que

tiveram uma variabilidade nos anos de 2010 e 2011 em meses específicos como Janeiro e Fevereiro.

Baseando-se nas análises já realizadas, o capítulo seguinte apresentará novas técnicas que possam ser aplicadas, para otimizar a análise já existente, utilizando-se de técnicas de análise e relacionando-as com soluções de *Big Data*.

4. TÉCNICAS E FERRAMENTAS COMPUTACIONAIS DE BIG DATA

Conforme apresentado no capítulo anterior, as técnicas de estruturação, exibição e mineração de dados mostraram-se adequadas quando aplicadas aos indicadores de agronegócios estudados diariamente pelo CEPEA. A estruturação dos dados que são recebidos diariamente, se tornam mais coerentes e fáceis quando estruturados em um *Data warehouse*, pois sua modelagem permite que Dados Históricos fiquem disponíveis de forma mais clara, facilitando sua análise. Processos de ETL para tratamento e padronização dos dados se tornam essenciais, para garantir com que os dados estejam na mesma granularidade, unidades de medida e formatos semelhantes, para que não haja distorção na análise a ser realizada.

Conforme discutido por Corrêa (2014), análises de dados relacionados ao Agronegócio tornam-se complexas pois podem envolver muitas dimensões, além da quantidade massiva de dados e variáveis que estas possuem, comportamentos peculiares de sazonalidades e tipos de comercialização. Tudo isso somado a análise de vários anos de histórico, torna-se um verdadeiro desafio.

O desafio de analisar grandes massas de dados com o intuito de adquirir processos mais estruturados, além de trazer insights valiosos para as áreas é uma meta não só do setor do agronegócio, como de todos os outros.

Obter indicadores que apontem o comportamento de uma variável ou produto, ou entender o andamento de seu concorrente, tornou-se essencial em um cenário cada vez mais competitivo. Mas para isso, petabytes de dados são gerados diariamente e esses precisam ser estruturados em ambientes cada vez mais robustos, que suportem a quantidade de dados que são gerados, além de sistemas que permitem o processamento desses, levando em consideração também, a variabilidade de informações com estruturas diferentes, o números significante de variáveis que fazem com que, quando analisadas em conjuntos, tragam informações relevantes para o negócio.

Devido a essa necessidade, foi criado no início dos anos 2000 um termo chamado *Big Data,* que permeiam todas as áreas de Dados, desde a coleta massiva dos

dados, até a disponibilização da informação para análise e apoio na tomada de decisão da área de negócio.(SAS, 2016)

Neste capítulo, são apresentadas técnicas e ferramentas computacionais de *Big Data* e *Analytics* que podem ser aplicadas ao CEPEA para gerar visualizações e análises, que poderão complementar as já realizadas pelo Centro.

São abordadas técnicas, tecnologias e ferramentas de *Big Data*, traçando um comparativo das técnicas que foram apresentadas nos capítulos anteriores, baseando-se no conhecimento adquirido durante o curso de especialização e também de conhecimentos profissionais e externos adquiridos durante a carreira profissional.

4.1. ANÁLISE PREDITIVA DE DADOS TEMPORAIS

Analisar dados temporais, tornou-se um desafio, devido a quantidade massiva de dados. O algoritmo K-Vizinho ou do inglês (*K-Nearest Neighbor*) que nesta monografia será tratado pela sigla KNN é um método de previsão de dados temporais. O algoritmo em questão é bastante utilizado por ser simples e a sua vasta utilização na análise de séries temporais não-lineares e no comportamento sazonal da série temporal. (FERRERO, 2009)

O KNN, conforme citado no item 2.4.1, consiste em rotular novos dados baseando-se em exemplos similares, que já possuem uma classificação. Ou seja, conforme McNames (1998 *apud* Ferrero, 2009) o intuito é encontrar K-Sequências semelhantes dentro da Série Temporal baseando-se em uma sequência de referência e também nos valores futuros dessas séries.

Ou seja, o intuito será encontrar os K exemplos que já estão identificados e rotulados na base histórica e compará-lo com os novos dados não classificados, com o intuito de prever os valores seguintes da série temporal. Os K-exemplos precisam ser identificados de forma correta, pois os recursos computacionais utilizados para a classificação são altos pois requer a comparação, na pior das hipóteses, com todos os dados da base a ser analisada.(FERRERO, 2009). Caso o

K seja definido como 1 por exemplo, o novo exemplo será classificado com a mesma classe do primeiro vizinho mais próximo.

Na Figura 14 é possível observar como exemplo que, no atributo 1 (k=1), o novo dado (Ei) será classificado como positivo, enquanto que no atributo 2, (K=4) será classificado como negativo.

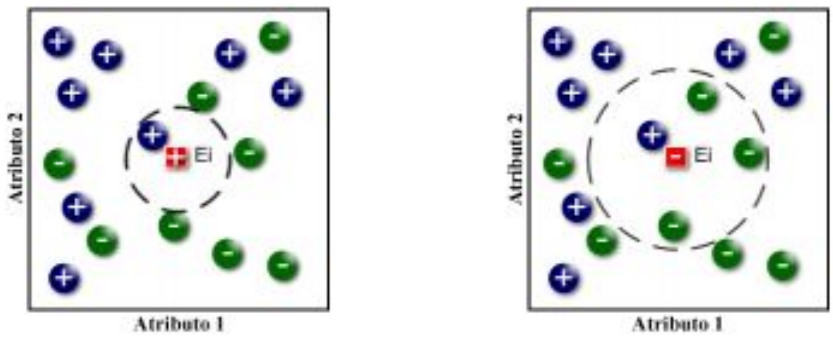


Figura 14: Exemplo de aplicação KNN. Fonte: Ferrero, 2009

Dessa forma, seria interessante utilizar somente os exemplos mais representativos das dimensões definidas para o CEPEA (Tempo, Produto e Mercado), sumarizando as informações em dados relevantes e que fazem sentido ao negócio, e utilizando um K-Exemplo apropriado, para que seja possível chegar a uma informação que faça sentido sem que haja muito gasto computacional.

Outro ponto importante na utilização do KNN é a definição da Medida de Similaridade. A utilização dessa se faz importante pois permite com que a distância entre 2 pontos não seja somente quantificada, mas também considerar seu comportamento e forma. Só assim será possível analisar se a sequência temporal é de fato semelhante ou não.(FERRERO, 2009). Atualmente a Distância Euclidiana é bastante utilizada para esse propósito, assim como outras técnicas, conforme descrito no item 3.3. Após definido esses pontos, é interessante realizar a normalização das séries que serão analisadas, pois essas podem apresentar

granularidades diferentes, ou mesmo são semelhantes mas possuem tamanhos diferentes conforme pode ser observado na Figura 15:

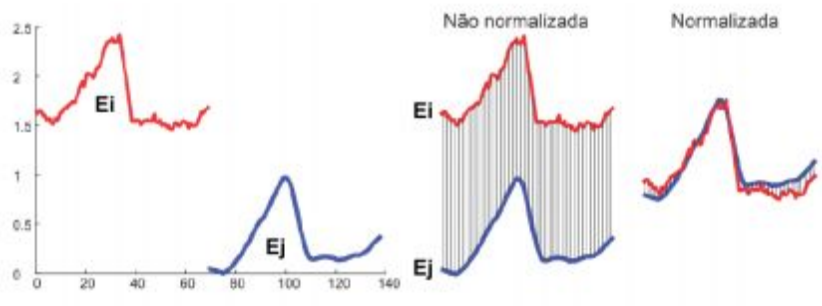


Figura 15. Aplicação de técnica de Normalização. Fonte: Ferrero, 2009

4.2. APLICAÇÃO DO MÉTODO NO CONTEXTO DO CEPEA

No capítulo seguinte serão abordadas técnicas de análises preditivas que são geralmente utilizadas com séries temporais.

4.2.1. KNN

Levando em consideração o estudo de caso do CEPEA, que apresenta séries temporais relacionados a preços do agronegócio e dimensões de Tempo, Mercado e Produtos, a aplicação do método de KNN pode seguir um padrão de 3 fases principais, conforme já analisado por Ferrero (2009), aplicando essa técnica em outros cenários:

- **Pré-Processamento da base histórica**: consiste em adequar a base realizando normalizações e padronizações para que essa se adeque ao formato correto para realizar métodos de classificação, ou mesmo de previsão.
- **Configuração e previsão:** Consiste em determinar quais e quantos dados serão estudados para realizar a previsão/classificação
- **Avaliação dos Resultados e Pré-Processamento**: etapa onde será avaliada a eficiência do modelo de classificação.

Uma aplicação seria prever novos valores de indicadores de preços de produtos praticados pelo CEPEA. No estudo, baseado nos dados históricos que o CEPEA possui de aproximadamente 20 anos de coleta, é possível realizar predições que melhor explicam a movimentação do mercado. Baseando-se na série temporal, é possível classificar os preços em 4 grupos principais: “QUEDA”, “ALTA”, “MÉDIA” e “SAZONAL”, sendo que podemos classificar em:

- **QUEDA**, os valores em que, baseado no histórico temporal dos preços dos produtos, possivelmente terão uma cotação MENOR no próximo período, baseando-se na comparação em um período anterior semelhante.

- **ALTA**, valores que se comparados a períodos anteriores, possivelmente terão uma cotação MAIOR do que o valor cotado no período semelhante anterior

- **MÉDIA,** valor que possivelmente NÃO OSCILARÃO seu valor, baseando-se na comparação com período semelhante anterior.

- **SAZONAL**, valores considerados “**outliers**”, ou seja, valores que deverão ter seu valor acrescido ou diminuído devido a algum agente externo temporário ou eventual. Esse valor pode ter ocorrido, devido a feriados no período, desastres naturais, ou algum agente imprevisto que influenciasse no indicador, e possivelmente ocorrerá novamente.

Baseando-se no relacionamento das principais variáveis de cada uma das dimensões, o histórico de indicadores seriam classificados, onde seria possível avaliar onde houve variações, além de ser possível estimar novos dados encaixando-os nas classificações acima, caso houvesse variação conforme critérios da Tabela 2:

TIPO DE VARIAÇÃO	VARIAÇÃO (%)
QUEDA	-10 à -20
ALTA	+10 à 20
MÉDIA	-10 à 10
SAZONAL	Abaixo de -20 ou Acima de 20

Tabela 2: Valores estimados de classificação dos indicadores. Fonte: Autor

Baseando-se nesses critérios será possível classificar os dados históricos de indicadores do CEPEA, além de realizar possíveis análises preditivas dos indicadores. Lembrando que esses valores podem não coincidir integralmente com a análise realizada, levando em consideração que o mercado de agronegócio sofre variações devido a diversos fatores externos, o que dificulta a predição de valores.

4.2.2. Arima

Para realização do processo de análise é necessário entender os princípios do problema que será analisado, e à partir desse ponto entender o comportamento da série que será estudada e assim entender de fato quais são os fatores que influenciam a série em si.

Para isso, será abordado sobre o *Autoregressive Integrated Moving Average* (ARIMA) ou do português (modelo auto-regressivo integrado de média móvel) consiste em um modelo muito utilizado para modelagem e previsão de séries temporais. Através dele é possível identificar e ajustar o melhor modelo que se adeque a sua necessidade entre um conjunto de modelos concorrentes (FELIPE, 2012). Ou seja, o objetivo deste é descrever de forma mais adequada os dados com o menor número de variáveis ou parâmetros, trazendo mesmo assim, previsões precisas.

O modelo é bastante utilizado com séries temporais não estacionárias, ou seja, que apresentam sazonalidades e tendências, como é o caso dos dados do CEPEA, que possuem valores oscilando no decorrer do tempo.

Segundo Morettin & Toloi (2004), entre os principais pontos para análise de séries temporais se destacam alguns, como a identificação do motivo gerador da série e a descrição do comportamento, gerando a previsão dos valores futuros, com suas tendências e sazonalidades.

Para chegar a conclusão de qual modelo se adequa melhor aos dados, é necessária a realização de 5 etapas chave, sendo elas:

1. Identificação. identificar variáveis e problemas que querem ser analisados

2. Seleção do modelo: identificar modelo que se adeque a análise

3. Estimação. estimar e testar modelos e sua aplicabilidade

4. Checagem. verificar se o modelo é efetivo

5. Previsão. prever se os valores foram assertivos e se estão próximos do real.

Essas etapas podem ser melhor identificadas na figura 16, conforme Makridakis *et al.* (1998, *apud* Medeiros*,* 2006), onde o modelo ARIMA é dividido em algumas etapas principais:

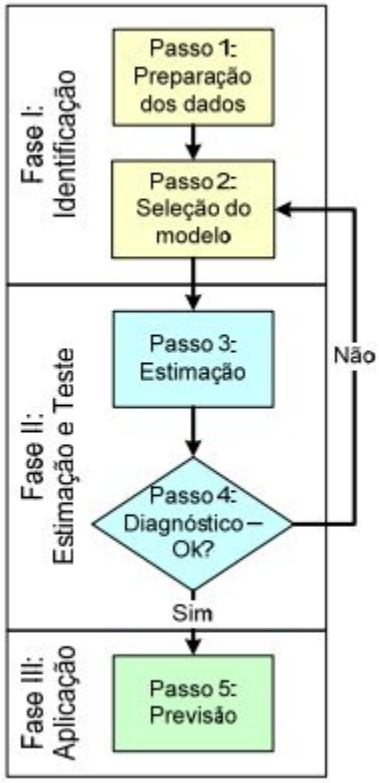


Figura 16. Etapas do Modelo ARIMA. Fonte: Medeiros (2006)

No contexto do CEPEA, e baseando-se no modelo e passo destacados, na fase I seria necessária a criação de gráficos para identificar padrões existentes nas variáveis a serem analisadas, buscando autocorrelação nos dados a serem analisados.

Para seleção de melhores modelos é recomendado a utilização de softwares especialistas, para identificar o modelo mais adequado para o cenário.

Na fase de estimação e testes, é necessário gerar as estatísticas como erro padrão dos dados, testes de significância e variância dos resíduos, e assim medir se o modelo que está sendo utilizado se é de fato eficiente e atende as necessidades da análise, e também se é possível melhorá-lo. (MEDEIROS, 2006)

Nesta etapa também será analisada se não há outros tipos de padrões que não foram contemplados no modelo. Caso haja será necessário realizar novamente a seleção de modelos.

Em seguida, caso os testes se adequem a necessidade, é realizada a etapa de previsão propriamente dita. Nessa etapa normalmente é necessária a utilização de softwares computacionais para processamento do modelo, pois dependendo da modelagem matemática que está sendo aplicada, pode ser necessária um processamento robusto das informações. (MEDEIROS, 2006)

4.3. CLASSIFICAÇÃO DOS DADOS - CEPEA

Para entender de forma mais detalhada o comportamento das séries temporais do CEPEA, é possível utilizar de técnicas de *clustering* de dados, através do comportamento das séries temporais é possível identificar as semelhanças de suas trajetórias e classificá-las em grupos.

Para realizar essa classificação, atualmente utiliza-se uma técnica chamada K-means. Segundo Koerich (2008), trata-se de uma técnica simples, interativa e poderosa para clusterizar um conjunto de dados em grupos, onde o valor K (número de Grupos) precisa ser pré-determinado. Dentre as principais vantagens deste algoritmo, o que a torna tão utilizada atualmente é a sua simplicidade computacional.

Na prática os **x** dados coletados são particionados em **n** conjuntos de dados, no qual cada novo valor pertence ao grupo que esteja mais próximo das médias dos valores próximos. Consiste na definição de centróides (um para cada *cluster*). Esses centróides precisam estar localizados equidistantemente um dos outros. Após realizado esse procedimento, é necessário que cada novo dado que será inserido, seja incluído dentro do cluster do centróide mais próximo. Após realizado esse procedimento e nenhum ponto esteja fora de um conjunto pré-definido, é possível realizar a definição de novos centróides, até que o modelo se adapte às necessidades. Na figura 17 é possível observar o gráfico gerado à partir da aplicação do método de classificação. É possível observar que foram gerados a partir da análise dos dados, 3 grupos.

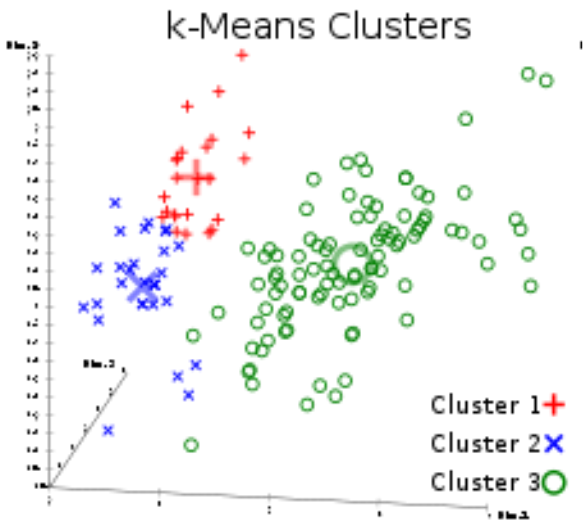


Figura 17. Exemplo de gráfico plotado utilizando K-Means. Fonte: Conteúdo Aberto

No Contexto do estudo de caso do CEPEA, a aplicação do K-Means seria através da classificação das séries temporais semelhantes de produtos distintos. Ou seja, caso produtos diferentes como Soja e Milho por exemplo apresentem seus dados no mesmo *cluster*, podemos dizer que ambos os produtos possuem características semelhantes.

5. CONCLUSÕES E TRABALHOS FUTUROS

Após o desenvolvimento da monografia, foi possível observar através das pesquisas e análises realizadas, a quantidade massiva de dados que vem sendo gerada nos mais diversos setores, e com o agronegócio não é diferente, conforme observado.

Diariamente, essa quantidade de dados e de variáveis como podemos citar a variação do dólar, clima, solo, possíveis acidentes naturais, entre outras, são essenciais e se relacionam para que se forme os preços dos produtos que são utilizados pelo setor de agronegócio. Com a variação desses fatores, os valores podem sofrer oscilações durante o tempo, e esses valores conforme foi descrito nos capítulos anteriores precisam ser analisados para apoiar pessoas e/ou indústrias que necessitam dessa informação para possíveis tomadas de ação, como agricultores, produtores, cooperativas, entre outros.

Para que seja possível realizar estudos e trazer *insights* nas informações que são geradas, essas precisam ser estruturadas de uma forma que seja possível analisá-las com a mesma granularidade e períodos semelhantes, para que tenha-se resultados mais assertivos quanto a análise. Dessa forma, estruturar e armazenar a informação corretamente em um *Data Warehouse* que torna a análise simples, fácil e de acesso rápido torna-se essencial.

Após a estruturação da informação, foi possível retirar informações relevantes de dados brutos utilizando-se técnicas de mineração de dados, onde utilizando-se de técnicas de PCA e Decomposição de Tucker, é possível reduzir a massa de dados em uma amostra menor (mas que seja coerente e que reflita a massa original de dados) para que haja recursos computacionais suficientes para processamento desses dados.

Em seguida, utiliza-se técnicas que busque encontrar padrões nos dados obtidos. Uma das mais utilizadas é a Distância Euclidiana, na qual através de seu uso é possível identificar similaridades entre dados de períodos temporais semelhantes.

Com as técnicas aplicadas, foi possível obter informações relevantes para o setor como variações similares entre produtos no decorrer do tempo, assim como mercados estrangeiros e nacionais que possuem variações semelhantes nos preços praticados em suas regiões.

Através da caracterização do CEPEA e de seus dados coletados, foi possível identificar aplicações de técnicas como algoritmos de classificação, sendo possível identificar *clusters,* quando produtos distintos apresentam o mesmo comportamento de variação no decorrer do tempo.

Também através da análise realizada, foi possível identificar a possibilidade de aplicar métodos de análises preditivas (que é considerado como um diferencial desse trabalho com os demais já realizados para o CEPEA, devido a sua complexidade), como o método ARIMA que é muito utilizado para modelagem e previsão de séries temporais, além do método KNN (k-Nearest Neighbors), onde baseando-se na comparação de dados históricos é possível prever um novo valor, através de quão próximos seus K-Vizinhos estão da nova amostra de dados.

Para possíveis trabalhos futuros, consideramos a aplicação de novos métodos como utilização de bancos de dados NoSQL e Spark e Hadoop para processamento massivo de Dados.

6. REFERÊNCIAS

AMAZON. Big Data. **O que é Big Data**? Disponível em: <https://aws.amazon.com/pt/big- data/what-is-big-data>. Acesso em: 11 Set 2016

ANTUNES, J. F. G. ; LAMPARELLI, RUBENS A. C. ; RODRIGUES, LUIZ H. A. **Avaliação da dinâmica do cultivo da cana-de-açúcar no estado de São Paulo por meio de perfis temporais de dados MODIS**. Engenharia Agrícola (Online), v. 35, p. 1127-1136, 2015.

BRONSON, K.; KNEZEVIC, I.. **Big Data in food and agriculture**. Sage Journals. Canadá, p. 1-5. 01 jun 2016.

CAIÇARA J.C. **Tópicos avançados em sistema de informações gerenciais**: Aula 6. 2014. Disponível em: <http://slideplayer.com.br/slide/294751/>. Acesso em: 05 Out 2016

CAMILO, C. O; SILVA, J.C **Mineração de dados: conceitos, tarefas, métodos e ferramentas**. Instituto de Informática Universidade Federal de Goiás: Goiás, Agosto de 2009.

CAROLAN, M. **Publicising food**: **big data, precision agriculture, and co-experimental techniques of addition**. European Society For Rural Sociology. Colorado, p. 1-20. jan. 2016.

CEPEA. **Centro de estudos avançados em economia aplicada**. Disponível em: <http://www.cepea.esalq.usp.br>. Acesso em: 10 Out 2016

GESTÃO NO CAMPO**. Conceito de agronegócio**.Disponível em: <http://www.gestaonocampo.com.br/conceito-de-agronegocio/>. Acesso em: 18 Set 2016.

CORRÊA, F. E. **Modelo integrado de mineração de dados para análise de séries temporais de preços de indicadores agroeconômicos**. 2014. 116 f. Tese (Doutorado) - Curso de Engenharia da Computação, Centro de Informática de São Carlos, Universidade de São Paulo, São Paulo, 2015.

CORRÊA, F. E.. **Representação de comercialização agropecuária através de modelo de data warehouse**. 2010. 70 f. Dissertação (Mestrado) - Curso de Sistemas Digitais, Centro de Informática de São Carlos, Universidade de São Paulo, São Paulo, 2010.

DATA MART**. DATA MART**. Disponível em: <http://futuroinformacao.blogspot.com.br/2015/06/data-mart.html>. Acesso em: 05 Out 2016.

DATASTORM. **Big Data. 5V de Big Data**: **Entenda a estrutura de grandes dados.** Disponível em: <http://datastorm.com.br/5v-big-data-estrutura/>. Acesso em: 11 Set 2016

ELIAS, D. **Dimensões e fatos no contexto do business intelligence**. 2014. Atualizado em 29/06/2015. Disponível em: <https://corporate.canaltech.com.br/noticia/business-intelligence/dimensoes-e-fatos-no-contexto-do-business-intelligence-bi-18710/>. Acesso em: 31 Mar 2016

FACELI, K. ; LORENA, A.; GAMA, J. ; CARVALHO, A. P. L. **Inteligência Artificial – Uma abordagem de aprendizado de máquina**, LTC, 1a Edição, 2011.

FELIPE, I.J. DOS S. **Aplicação de modelos arima em séries de preços de soja no norte do paraná.** 2012. 17 f. Dissertação (Mestrado) - Curso de Administração, Ufrn, Botucatu, 2012.

FERRERO, C.A. **Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia.** 2009. 129 f. Dissertação (Mestrado) - Curso de Ciência da Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2009.

HIGHBEAM RESEARCH. **Researchers from institute of agriculture describe findings in geoscience (sugarcane mapping in tillering period by quad polarization terrasar-x data)**. China, 2015. Disponível em: https://www.highbeam.com/doc/1G1-415379921.html/. Acesso em: 07 Ago 2016

INSPER. **O que é um estudo de caso?** Disponível em: <http://www.insper.edu.br/casos/estudo-caso/>. Acesso em: 17 Set 2016.

INTERACTIVE, O. W. **Brasil exporta 10 milhões de toneladas de soja**. Disponível em:<http://www.uagro.com.br/editorias/agricultura/soja/2016/05/09/brasil-exporta-10-milhoes-de-toneladas-de-soja.html>. Acesso em: 17 Set. 2016.

KOERICH, A.L. **Aprendizagem de máquina:** Paraná:, 2008. 35 slides.Color. Dissertação (Mestrado) - Curso de Informática Aplicada, Pontifícia Universidade Católica do Paraná.

KOLDA, T.G.: BADER, B.W. Tensor decompositions and applications. **SIAM review**, v. 51, n.3, p. 455-500, 2009.

LOPES, W.C. **Arquitetura de integração entre máquinas, implementos e sistemas de gestão agrícola: uma abordagem orientada a serviço e a tecnologias embarcadas isobus**. 2015. 157 f. Tese (Doutorado) - Curso de Engenharia Mecânica, Universidade de São Paulo, São Carlos, 2015.

MEDEIROS, A.L. **Regressão múltipla e o modelo arima na previsão do preço da arroba do boi gordo**. 2006. 124 f. Dissertação (Mestrado) - Curso de Engenharia de Produção, Universidade Federal de Itajubá, Itajubá, 2006.

MORETTIN, P. A.; TOLOI, C.M.C. **Análise de séries temporais**. São Paulo: Edgard Blucher. 2004. 535p.

NESCARA TECNOLOGIA LTDA. **Agro: Big Data, geolocalização e mobilidade melhoram a produtividade no campo**. PressWorks Assessoria de Imprensa, São Paulo, 26/04/2016. Disponível em: https://www.pressworks.com.br/noticias/agro-big-data-geolocalizacao-e-mobilidade-melhoram-produtividade-no-campo/1006. Acesso em:07 Ago 2016

OLIVEIRA NETO, R.F. DE. **Aprendizagem de máquina**. Pernambuco: Universidade Federal do Vale do São Francisco, 2012. 19 slides, color. Disponível em: <http://www.univasf.edu.br/>. Acesso em: 09 Out 2016.

REUTERS, G. B. **Consultoria vê ganho de RS 24 bilhões para agricultura do Brasil em 5 anos com Big Data**. Estadão, São Paulo, 23/10/2014. Disponível em:<http://economia.estadao.com.br/noticias/geral,consultoria-ve-ganho-de-r-24-bilhoes-para-agricultura-do-brasil-em-5-anos-com-big-data, 1581522/> Acesso em: 07 Ago 2016.

RICARDO, J. **Introdução à tecnologia data warehouse**. 2015. Devmedia. Disponível em: <http://www.devmedia.com.br/introducao-a-tecnologia-data-ware house/27629>. Acesso em: 05 Out. 2016.

SAS. **Big Data: O que é e porque é importante?**. Disponível em: <http:/ /www.sas .com/pt_br/insights/big-data/what-is-big-data.html>. Acesso em 07 Set 2016

SAS. **Análise Preditiva: O que é e porque é importante?.** Disponível em: <http://www.sas.com/pt_br/insights/analytics/analise-preditiva.html>. Acesso em 17 Set 2016

TANAKA, A. K.. **Banco de dados distribuídos e datawarehousing**. Disponível em: <http://www.uniriotec.br>. Acesso em: 05 Out 2016.

Universidade Estadual Paulista Júlio de Mesquita Filho. Grupo de Pesquisa CSME. **Aplicação de big data na agricultura**. São Paulo, 24/01/2016. Disponível em:<http://datastorm.com.br/inovacoes-na-agricultura-conheca-o-big-data> Acesso em 07 Ago 2016

Universidade Estadual Paulista Júlio de Mesquita Filho. **Pesquisadores propõem mapeamento do solo de forma mais precisa.** São Paulo,04/03/2015. Disponível em:<http://unan.unesp.br/destaques/0/16931/Pesquisadores-propoem-mapeamento-do-solo-de-forma-mais-precisa/> Acesso em 07 Ago 2016