Predição da qualidade da carne de animais abatidos no programa Precoce MS

Inara Santana Ortiz

Predição da qualidade da carne de animais abatidos no programa Precoce MS

Inara Santana Ortiz

Orientador: *Profº Drº Renato Porfirio Ishii*

Dissertação entregue a Faculdade de Computação da Universidade Federal de Mato Grosso do Sul - FACOM-UFMS para o exame de Qualificação do Mestrado Profissional em Computação Aplicada.

UFMS - Campo Grande
janeiro/2021

Abstract

In order to encourage producers in the state of Mato Grosso do Sul to produce quality beef and thus reach the criteria of an increasingly demanding market, the carcass bonus program was created, which benefits cattle producers and adds value to your product, in addition to supplying meat of higher quality to the slaughterhouse industries. The Precoce MS program encourages the slaughter of animals at a young age, which meet the criteria of carcass quality, through a more sustainable means of production. However, producers face the challenge of meeting the criteria of the bonus programs and innovations in the agricultural sector that need to emerge to support them in decision making in the quality meat production process. This work aims to build a classification model to predict the quality of the carcass through Data Mining techniques, using different data sources that contain information on animal slaughter according to the classification criteria as early steer, the productive process, the climate, the geographic location and vegetation indexes.

Keywords: Data Mining, Machine learning, Supervised Algorithms, Precision Livestock.

Resumo

Com o intuito de incentivar os produtores sul-mato-grossenses na produção de carne bovina de qualidade e, assim, atender aos critérios de um mercado cada vez mais exigente, foi criado o programa de bonificação de carcaça, que beneficia pecuaristas e agrega valor ao seu produto, além de fornecer carne de maior qualidade às indústrias frigoríficas. O programa Precoce MS estimula o abate de animais em idade jovem, que atendam aos critérios de qualidade de carcaça, através de um meio de produção mais sustentável. No entanto, os produtores enfrentam o desafio de atender aos critérios dos programas de bonificação e inovações no setor agropecuário precisam surgir para apoiá-los na tomada de decisão no processo produtivo de carne de qualidade. Este trabalho tem por objetivo construir um modelo de classificação para predizer a qualidade da carcaça através de técnicas de Mineração de Dados, utilizando diferentes fontes de dados que contém informações do abate animal segundo os critérios de classificação como novilho precoce, do processo produtivo, do clima, da localização geográfica e dos índices de vegetação.

Palavras-chave: Mineração de Dados, Aprendizado de Máquina, Algoritmos Supervisionados, Pecuária de Precisão.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM Aprendizado de Máquina

ABNP Associação Brasileira do Novilho Precoce

ASPNP Associação Sul-Mato-Grossense de Produtores de Novilho Precoce

AM Aprendizado de Máquina

MD Mineração de Dados

IA Inteligência Artificial

ICMS Imposto sobre Circulação de Mercadorias e Serviços

SIF Serviço de Inspeção Federal

Proape Programa de Avanços na pecuária de Mato Grosso do Sul

USDA *United States Department of Agriculture*

FACOM Faculdade de Computação

UFMS Universidade Federal de Mato Grosso do Sul

RFC *Random Forest Classifier*

KNN *K-Nearest Neighbour*

SVM *Support Vector Machines*

CRISP-DM *Cross Industry Standard Process for Data Mining*

SVR *Support Vector Machines for Regression*

DW *Data Warehouse*

CEMTEC Centro de monitoramento do Clima e Tempo do Mato Grosso do Sul

ITU Índice de Temperatura e Umidade

MODIS *MODerate resolution Imaging Spectroradiometer*

SAS *Statistical Analysis System*

CAPÍTULO

1

Introdução

Em 2018, o Brasil foi o maior exportador mundial de carne bovina, fornecendo quase 20% do total das exportações mundiais. Além disso, o *United States Department of Agriculture* (USDA) projeta que o Brasil continuará sua trajetória de crescimento das exportações para a próxima década, atingindo 2,9 milhões de toneladas, ou 23% do total das exportações mundiais de carne bovina, até 2028 [6]. A produção de carne bovina desempenha um papel importante em termos de geração de emprego e riqueza para a economia brasileira [7].

Apesar de ser o líder mundial em tonelagem de carne bovina exportada, o país tem uma renda relativamente baixa, já que não exporta para os mercados de maior valor agregado. A carne brasileira, segundo alguns importadores, não é considerada de boa qualidade [8]. Um dos fatores que influencia a qualidade dos animais é a idade ao abate, que no Brasil gira em torno de 36 a 40 meses, em função dos animais serem criados em sistemas extensivos de produção, ou seja, somente à pasto, levando mais tempo que para chegarem ao ponto de abate se comparado a países que utilizam sistemas mais intensivos.

O conceito sobre a produção de "novilho precoce" surgiu no final da década de 70 e visava aumento da produtividade da pecuária e agregação de valor à carne que seria comercializada. Em agosto de 1977 foi criada a Associação Brasileira do Novilho Precoce (ABNP). Posteriormente, por volta de 1994, começam a surgir associações e núcleos de novilho precoce e os governos de vários estados passaram a conceder devolução escalonada do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) aos pecuaristas, nos demais estados, para fomentar a produção de bovinos jovens [9].

Em 1998 foi fundada a Associação Sul-Mato-Grossense de Produtores de

Novilho Precoce (ASPNP), e desde então, vem proporcionando aos seus associados melhores condições para a negociação de seus produtos, e assegurando que os produtos ofertados estão de acordo com as normas e exigências do mercado nacional e internacional [10].

Em 2003 o Estado de Mato Grosso do Sul (MS), conforme o decreto Nº 11.176 de 11/04/2003, institui o Programa de Avanços na pecuária de Mato Grosso do Sul (Proape), visando à expansão e o fortalecimento da bovinocultura. Posteriormente o programa passa por reestruturação de operacionalização através do decreto Nº 14.526, de 28 de julho de 2016 e da resolução conjunta SEFAZ/SEPAF Nº 69 de 30 de agosto de 2016 e é instituído o subprograma Precoce MS. Neste, a qualidade da carne é avaliada por três elementos, sendo eles o animal (sexo, maturidade, acabamento e peso), o lote (no mínimo 60% deve ser composto por novilhos precoces) e o processo produtivo (boas práticas agropecuárias, identificação animal, associativismo e sustentabilidade), aos produtores inscritos no programa é concedido isenção do ICMS de até 67%.

Os programas de bonificação são um grande incentivo para a produção de carne bovina de qualidade, no entanto os produtores enfrentam o desafio de atender aos critérios dos programas de bonificação e consequentemente aos critérios de qualidade do mercado interno e externo. Para viabilizar a produção de carne bovina de qualidade, inovações no setor agropecuário precisam surgir para apoiar os produtores na tomada de decisão no processo produtivo de carne de qualidade, e que equilibre os benefícios das bonificações em relação aos níveis de investimentos necessários. Nesse sentido a área da Pecuária de precisão, visa desenvolver esse tipo de soluções tecnológicas, e pode ser definida como uma prática de manejo de rebanhos bovinos na qual a tecnologia da informação e comunicação é usada para garantir boas práticas de produção de carne. Com base em dados específicos de rebanho, áreas de pastoreio georreferenciadas, automação em todas as etapas da indústria de carnes, visa otimizar os custos de produção na obtenção de uma carne de qualidade [11].

Na busca por tais inovações, a Faculdade de Computação (FACOM) da Universidade Federal de Mato Grosso do Sul (UFMS) juntamente com a Embrapa Gado de Corte tem desenvolvido pesquisas na área da Pecuária de Precisão. Os trabalhos são elaborados pelos discentes sob orientação de professores da FACOM e coorientação de pesquisadores da Embrapa. A exemplo temos dois trabalhos cujo o objetivo era elaborar, aplicar e avaliar abordagens de análise dos dados de animais abatidos, extrair conhecimento e utilizá-lo como suporte na tomada de decisão no processo de produção de gado de corte.

No primeiro trabalho [12] o autor aplicou em um conjunto de dados fornecido pela ASPNP, técnicas computacionais voltadas a Business Intelligence

(BI) e Mineração de Dados (MD) com o objetivo de descobrir padrões e relacionamentos vinculados ao grau de acabamento e ao rendimento da carcaça. No entanto o autor encontrou dificuldade para alcançar uma boa precisão, devido a generalização das informações dos animais abatidos. Na seção de trabalhos futuros o autor sugere a criação de um novo conjunto de dados com informações mais precisas do animal e informações do sistema de produção "porteira a dentro", ou seja, dados que informem como esses animais foram criados antes de serem abatido, e dessa forma obter resultados mais assertivos. O segundo [13] foi inspirado nos resultados e propostas de [12] e teve como objetivo principal a construção de um classificador do grau de acabamento da carcaça. O conjunto de dados utilizado era composto pelas informações de abate de bovinos cadastrados no Precoce MS e do processo produtivo de cada estabelecimento rural correspondente, informações mais precisas se comparadas ao conjunto de dados utilizado no anterior. Dessa forma obteve-se resultados de taxa de precisão de aproximadamente 70,45% e 70,11% nas duas melhores abordagens utilizadas, e o desbalanceamento do conjunto de dados de abate foi uma das dificuldades elencadas pelo autor.

Ao observar os resultados dos trabalhos anteriores vimos que, aplicando técnicas de MD é possível gerar informações que possam auxiliar o produtor na tomada de decisão e obter resultados mais precisos. Para a execução deste trabalho técnicas de MD serão aplicadas em um conjunto de dados composto por informações do programa Precoce MS e de outras fontes dentro do contexto da produção animal, tais como: clima, localização geográfica e índices de vegetação, desta forma espera-se otimizar a predição da qualidade da carne.

1.1 Objetivos

O objetivo geral deste trabalho é predizer a qualidade da carcaça através de técnicas de Mineração de Dados (MD), utilizando diferentes fontes de dados que contém informações do abate animal segundo os critérios de classificação como novilho precoce, do processo produtivo, do clima, da localização geográfica e do índices de vegetação. A partir do objetivo geral, define-se os objetivos específicos como:

I Compreender o problema e identificar as reais necessidades do projeto;

II Adquirir e documentar os dados e a partir análise exploratória identificar padrões, obter *insights* e formular hipóteses;

III Tratar os dados para que os algoritmos possam ser aplicados, ou seja, eliminar qualquer tipo de inconsistência que possa interferir nos desempenho do algoritmo aplicado ao conjunto de dados;

IV Aplicar os algoritmos de Aprendizado de Máquina (AM) no conjunto de dados; e

V Selecionar dentre os algoritmos de AM os que apresentarem melhor desempenho.

Para alcançar os objetivos específicos descritos, primeiramente será feita uma busca na literatura por trabalhos que empregam técnicas de Aprendizado de Máquina na Pecuária para descobrir o quê já foi publicado sobre o tema. Após a aquisição do conjunto de dados, deve-se entender a origem de cada informação e descrevê-las, analisar previamente os dados ao aplicar técnicas de análise exploratória, além de descobrir variáveis importantes e detectar valores anômalos. Para que os modelos de AM possam ser aplicados nas bases, deve-se corrigir problemas tais como valores nulos e dados inconsistentes. Após, testes com os modelos de AM serão conduzidos no conjunto de dados e o modelo que apresentar a melhor acurácia, selecionado.

Ao final, espera-se construir um modelo que possa predizer a qualidade da carcaça com alta acurácia, e descobrir quais são as características que mais contribuem para a produção de carne de qualidade. Dessa forma, pretende-se auxiliar os produtores rurais na tomada de decisão, mais do que isso, diminuir o risco de insucesso técnico e econômico do produtor e; principalmente em regiões onde há um maior necessidade, contribuir para a criação de políticas públicas e incentivos a produção de novilho precoce.

1.2 Organização

O texto está organizado como segue. No Capítulo 2 conceitos básicos importantes que fundamentam este trabalho são descritos. Alguns estudos com proposta similar a este projeto são apresentados no Capítulo 3. Por fim, no Capítulo 4, a proposta de trabalho e metodologia são descritos e o cronograma de execução do projeto é apresentado.

CAPÍTULO

2

Fundamentação Teórica

2.1 Mineração de Dados

Nos últimos anos, o aumento da disponibilidade de uma grande quantidade e variedade de dados advindos de diversas fontes, internas e externas às empresas, levou ao aumento do interesse em métodos para extrair informações e conhecimento a partir dos dados [14].

A Mineração de dados, do inglês *Data Mining*, é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grande quantidade de dados armazenados em bancos de dados [15]. Em parte da literatura relacionada, a MD é também tratada como sinônimo para outro termo, a descoberta de conhecimento em bases de dados (KDD, do inglês *Knowledge Discovery in Databases*) [16]. Outros autores consideram a mineração de dados uma etapa no processo de KDD, o qual compreende as seguintes etapas: a seleção, pré-processamento, transformação, mineração dos dados e a interpretação dos resultados [17].

Para facilitar o processo de descoberta de conhecimento, as equipes de cientistas de dados utilizam e desenvolvem métodos que servem como guias durante o desenvolvimento dessas etapas. Atualmente diversos processos definem e padronizam as fases e atividades da MD, de maneira geral todos possuem uma estrutura similar. Neste trabalho, escolhemos o CRISP-DM (*Cross-Industry Standard Process of Data Mining*) como modelo e será apresentado na seção a seguir.

2.2 CRISP-DM

A aplicação de técnicas de MD para extrair conhecimento estava se tornando cada vez mais comum na indústria, algumas empresas e instituições viram a necessidade de unir forças para identificar boas práticas, bem como erros comuns em suas experiências anteriores. O *Cross Industry Standard Process for Data Mining* (CRISP-DM) foi concebido em 1996 em parceria entre três empresas que aplicavam MD em seus negócios: Daimler Chrysler, SPSS e NCR4 [18].

O ciclo de vida apresentado pela metodologia CRISP-DM consiste em seis fases que não precisam ser seguidas rigorosamente, tendo certa flexibilidade para ir e vir retomando as etapas anteriores conforme se faz necessário. A seguir, uma breve descrição das etapas:

I **Entendimento de Negócio (*Business Understanding*):** Esta é a fase inicial do CRISP-DM. Nesta fase, um entendimento do objetivo e dos requisitos do projeto deve ser formado a partir de uma perspectiva de negócio. Esse entendimento será então transformado em uma definição de problemas de mineração de dados, para criar um plano de projeto para atingir os objetivos.

II **Entendimento dos Dados (*Data Understanding*):** Esta fase inicia com uma coleta inicial de dados e, em seguida, prossegue com atividades que permitem que você se familiarize com os dados, identifique problemas de qualidade de dados, descubra os primeiros *insights* sobre os dados, detecte subconjuntos interessantes. Nesta fase é também onde se formam hipóteses em cima do que se aprendeu com os dados.

III **Preparação dos Dados (*Data Preparation*):** esta etapa consiste em preparar os dados para a modelagem. É a construção de um conjunto de dados obtidos dos dados brutos iniciais. As tarefas de preparação de dados provavelmente serão executadas várias vezes e não em qualquer ordem prescrita.

IV **Modelagem (*Modeling*):** Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ideais, as técnicas que serão utilizadas neste trabalho são apresentadas na seção 2.3. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos na forma de dados. Portanto, voltar para a fase de preparação de dados muitas vezes é necessário.

V **Avaliação (*Evaluation*):** Antes que o modelo possa ser implantado, o trabalho realizado precisa ser avaliado minuciosamente e revisar as etapas executadas para criá-lo, para ter certeza de que o modelo atinja os objetivos de negócios adequadamente. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser tomada.

VI **Utilização ou Aplicação (*Deployment*):** Nesta fase, o modelo final é implantado. Dependendo dos requisitos do projeto essa fase pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível em toda a empresa. Nesta fase, é importante que o cliente entenda de quais ações precisam ser realizadas para realmente fazer uso dos modelos criados.

2.3 Aprendizado de Máquina

A mineração de dados usa aprendizagem de máquina AM, do inglês *machine learning*, que é uma subárea da inteligência artificial, onde ferramentas computacionais são capazes de criar por si próprias, a partir da experiência passada, indução de uma hipótese, ou função (também chamada de modelo). Podemos encontrar diversas definições de AM na literatura, das quais destacamos a seguir:

> "A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência [19]."

Embora AM seja naturalmente associado a Inteligência Artificial (IA), outras áreas de pesquisa são importantes e tem contribuições diretas e significativas no avanço do AM, como Probabilidade e Estatística, Teoria da Computação, Neurociência, Teoria da Informação, para citar algumas. AM e uma das áreas de pesquisa da computação que mais tem crescido nos últimos anos [16].

A principal característica dessa área de estudo é a utilização de algoritmos específicos que recebem dados de entrada (dados de treinamento), com o objetivo de aprender com experiência passada. Após o final do processo de aprendizagem, em tarefas de previsão, o modelo treinado poderá prever um rótulo ou valor que caracterize um novo exemplo.

Os modelos nos quais existe um conjunto de dados rotulado previamente para a aprendizagem, onde a saída correta já é conhecida, são chamados de *supervisionados*. E modelos nos quais não existem esses rótulos, são chamados de *não supervisionados*. Neste trabalho apenas modelos de aprendizado de máquina supervisionado serão utilizados..

2.4 Algoritmos de Aprendizado de Máquina

Um algoritmo de AM, para modelo supervisionados, é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. O rótulo ou etiqueta toma valores num domínio conhecido. Se esse domínio for um conjunto de valores nominais, tem-se um problema de classificação e o estimador gerado é um classificador [16]. Uma definição formal seria, dado um conjunto de observações de pares

$$D = \{(xi, f(xi)), i = 1, ..., n\}$$

em que f representa uma função desconhecida, um algoritmo de AM preditivo aprende uma aproximação f' da função desconhecida f. Essa função aproximada, f', permite estimar o valor de f para novas observações de x. Para a classificação, $yi = f(xi) \in \{c1, ..., cm\}$, ou seja, $f(xi)$ assume valores em um conjunto discreto, não ordenado[16].

Os algoritmos de aprendizado de máquina que serão utilizados neste trabalho são: *Naive Bayes*, *K-Nearest Neighbor* ou K-Vizinho Mais Próximo, *Random Forest Classifier* ou Floresta Aleatória, *Support Vector Machines* ou Máquinas de Vetores de Suporte e *Deep learning* ou Aprendizagem Profunda .

2.4.1 Naive Bayes

Outra forma de lidar com tarefas preditivas em AM, principalmente quando as informações disponíveis são incompletas ou imprecisas, é por meio do uso de algoritmos baseados no teorema de Bayes, os métodos probabilísticos bayesianos. Os métodos probabilísticos bayesianos assumem que a probabilidade de um evento A, que pode ser uma classe, dado um evento B, que pode ser um valor para um atributo de entrada, não depende apenas da relação entre A e B, mas também da probabilidade de observar A independentemente de observar B [19].

2.4.2 K-Nearest Neighbour

Algoritmo *K-Nearest Neighbour* (KNN) considera a proximidade entre os dados na realização de predições. A hipótese base é que dados similares tendem a estar concentrados em uma mesma região no espaço de entrada. Seu funcionamento é relativamente simples: ao se classificar uma nova instância, o algoritmo busca as k instâncias que possuem a menor distância em relação à nova instância. Uma grande desvantagem deste algoritmo (além da lentidão ao se predizer novas instâncias), é que todas as características de uma instância possuem o mesmo peso ao se calcular a distância, portanto deve ser utilizado

com cautela. Ou seja, caso uma característica tenha uma importância maior do que as outras, na hora de se classificar uma instância, esta importância acaba sendo descartada por este algoritmo [20].

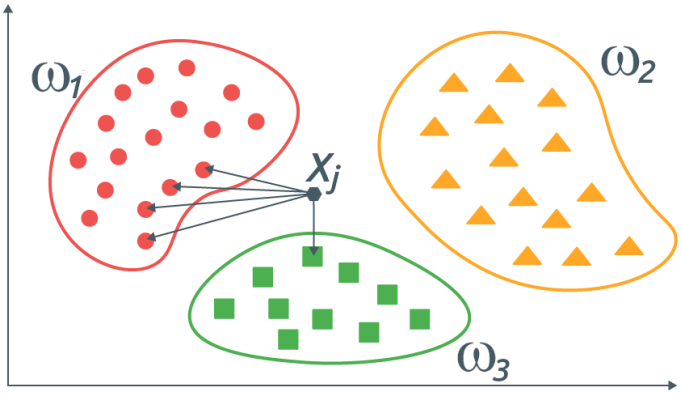


Figura 2.1: O KNN é um algoritmo simples que prevê pontos de dados desconhecidos de acordo com os seus vizinhos mais próximos. Imagem adaptada de [1].

2.4.3 Random Forest Classifier

Uma *árvore de decisão* é a decomposição de um problema muito complexo em subproblemas menos complexos. A ideia é a utilização da divisão para a conquista e, assim, de forma recursiva a mesma técnica é aplicada nos subproblemas. Tal capacidade de discriminar a árvore consiste em divisão em subespaços com a utilização de atributos e a cada subespaço se faz uma associação de uma classe [21].

O algoritmo *Random Forest Classifier* (RFC) cria várias subamostras de árvores de decisão a partir do subconjunto do conjunto de treinamento selecionado aleatoriamente. Em seguida, agrega os votos de diferentes árvores de decisão para decidir a classe final do objeto de teste usando a média para melhorar a precisão preditiva e controlar o ajuste excessivo [22]. Seu funcionamento é ilustrado da figura 2.2.

2.4.4 Support Vector Machines

As Máquinas de Vetores de Suporte, do Inglês *Support Vector Machines* (SVM) é uma das técnicas que recorre à otimização de uma função em seu treinamento [16]. Basicamente o funcionamento de uma SVM pode ser descrito da seguinte forma: dadas duas classes e um conjunto de pontos que pertencem a essas classes, uma SVM determina o hiperplano que separa os pontos de forma a colocar o maior número de portos da mesmo classe do mesmo lado, enquanto maximiza a distância de cada classe a esse hiperplano.

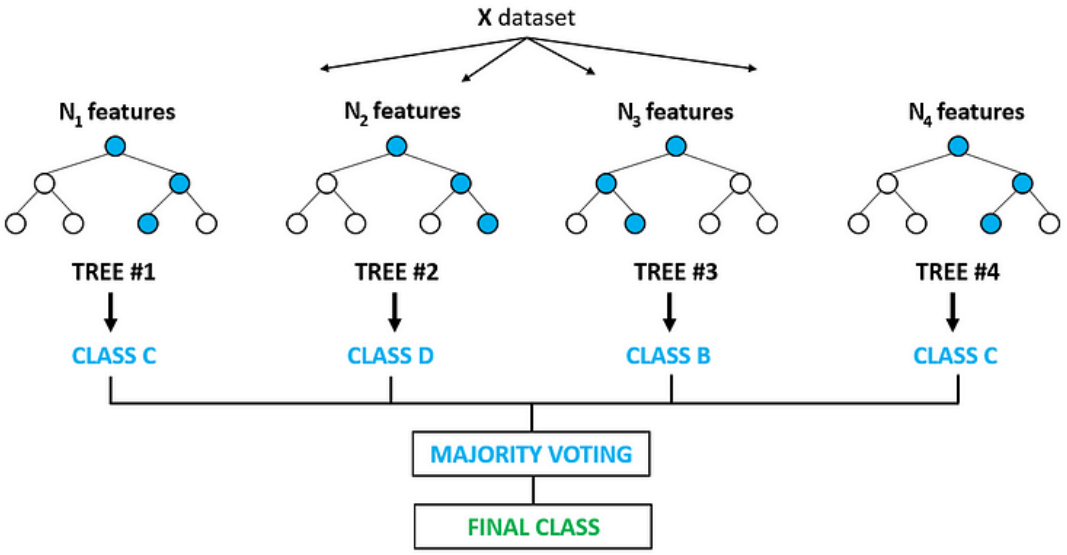


Figura 2.2: O diagrama acima mostra a estrutura do RFC. De um determinado conjunto de dados *Dataset* são selecionadas amostras aleatórias. Uma árvore de decisão é construída para cada amostra e obtém-se um resultado de previsão para cada uma. A previsão com mais votos *Majority Voting* é selecionada como previsão final [2].

Esta técnica originalmente desenvolvida para classificação binária, busca a construção de um hiperplano como superfície de decisão, de tal forma que a separação entre exemplos seja máxima. Isso considerando padrões linearmente separáveis. Já para padrões não-linearmente separáveis, busca-se uma função de mapeamento apropriada para tornar o conjunto mapeado linearmente separável [23].

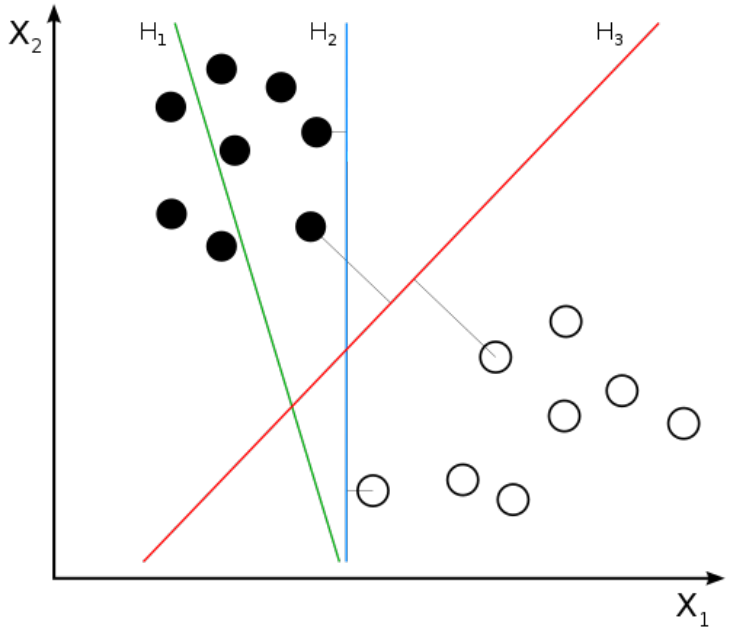


Figura 2.3: No diagrama acima, a linha vermelha é a melhor linha, pois tem a maior distância dos pontos mais próximos [3].

2.4.5 Deep Learning

Deep learning ou Aprendizagem Profunda em português é um tipo de Aprendizado de máquina que treina computadores para realizar tarefas como seres humanos. Os modelos apresentados anteriormente são chamados de "superficiais" ou "rasos" (o termo utilizado em inglês, é *shallow*), e buscam diretamente por uma única função que possa, a partir de um conjunto de parâmetros, gerar o resultado desejado. Por outro lado, em aprendizagem profunda temos métodos que aprendem por meio da composição de funções, cada função tem seu próprio conjunto de parâmetros e sua saída será passada para a próxima função [24].

Os autores de [25] afirmam que *Deep Learning* permite que os computadores aprendam seguindo uma hierarquia de conceitos, partindo de conceitos mais simples. Nessa hierarquia os conceitos são construídos um sobre o outro, no qual se colocados em um gráfico será possível observar que o gráfico ficará com muitas camadas ou profundo, diante disso que se considera a denominação de aprendizado profundo.

Uma *Deep Learning* busca construir conceitos complexos partindo de conceitos mais simples, utilizando Rede Neural Artificial (RNA), em que o aprendizado é dividido em hierarquias ou camadas, o que permite que o aprendizado ocorra em várias etapas. Quanto maior a quantidade de etapas ou profundidade, maior a quantidade de instruções que podem ser executadas em sequência [25].

2.5 Medidas de Avaliação

Para que seja possível avaliar o desempenho da classificação de algoritmos de AM, ou seja, verificar o quanto está acertando, podemos aplicar métricas de desempenho. Dentre as métricas mais usadas para avaliar modelos, para este trabalho serão utilizadas: matriz de confusão, *precision*, *recall* e *f1 score*.

A **matriz de confusão** ilustra o número de predições corretas e incorretas em cada classe. Para um determinado conjunto de dados, as linhas dessa matriz representam as classes verdadeiras, e as colunas, as classes preditas pelo classificador. Logo, cada elemento mij de uma matriz de confusão M apresenta o número de exemplos da classe i classificados como pertencentes à classe j. Para k classes, M tem então dimensão $k \times k$. A diagonal apresenta os acertos do classificador, enquanto os outros elementos correspondem aos erros cometidos nas suas predições. Por meio do exame dessa matriz, têm-se medidas quantitativas de quais classes o algoritmo de aprendizado tem maior dificuldade. Na Tabela 2.1 é apresentado um exemplo de matriz de confusão para um problema com três classes.

Tabela 2.1: Exemplo de uma matriz de confusão com três classes.

		Classe predita		
		A	**B**	**C**
Classe verdadeira	**A**	6	2	0
	B	3	4	2
	C	0	1	11

A forma mais comum de analisar uma matriz de confusão pode ser observada na Tabela 2.2. Usualmente, uma classe é denotada positiva **P** e a outra é denominada negativa **N**.

Tabela 2.2: Matriz de confusão para um problema com duas classe.

		Classe predita	
		P	**N**
Classe verdadeira	**P**	VP	FN
	N	FP	VN

- **Verdadeiro Positivo - VP:** corresponde ao número de verdadeiros positivos, ou seja, o número de exemplos da classe positiva classificados corretamente.

- **Verdadeiro Negativo - VN:** corresponde ao número de verdadeiros negativos, ou seja, o número de exemplos da classe negativa classificados corretamente.

- **Falso Positivo - FP:** corresponde ao número de falsos positivos, ou seja, o número de exemplos cuja classe verdadeira é negativa mas que foram classificados incorretamente como pertencendo à classe positiva.

- **Falso Negativo - FN:** corresponde ao número de falsos negativos, ou seja, o número de exemplos pertencentes originalmente à classe positiva que foram incorretamente preditos como da classe negativa.

A partir da matriz de confusão, uma série de outras medidas de desempenho pode ser derivada. Entre elas, temos [26]:

- **Precisão ou *Precision*:** proporção de exemplos positivos classificados corretamente entre todos aqueles preditos como positivos, conforme apresentado pela equação 2.1;

$$Precision = \frac{VP}{VP + FP} \tag{2.1}$$

- **Sensitividade ou *Recall*:** corresponde à taxa de acerto na classe positiva. Também é chamada de taxa de verdadeiros positivos (TVP), conforme apresentado pela equação 2.2;

$$Recall = \frac{VP}{VP + FN} \tag{2.2}$$

- **Acurácia ou *Accuracy*:** é calculada a partir do total de acertos (VP e VN) e dividida pelo conjunto total dos dados (VP, FN, FP e VN). Representa a taxa de exemplos positivos e negativos classificados corretamente, conforme apresentado pela equação 2.3;

$$Accuracy = \frac{VP + VN}{(VP + FN + FP + VN)} \tag{2.3}$$

No geral, as medidas precisão e sensitividade são combinadas para medir o desempenho do modelo, gerando uma equação conhecida como ***f1-score***, que é a média harmônica entre ambas as métricas. O cálculo de ***f1-score*** é mostrado na equação 2.4;

$$f1\text{-}score = \frac{(2 \times precision \times recall)}{(precision \times recall)} \tag{2.4}$$

2.6 Considerações Finais

A fundamentação teórica apresentou os conceitos e tecnologias que serão utilizadas no desenvolvimento do processo de descoberta de conhecimento, na base de dados objeto de estudo deste trabalho. A mineração de dados é um recurso utilizado para reconhecer padrões e regras que possam auxiliar na tomada de decisão, e para tal tarefa são aplicados à base de dados os modelos de Aprendizado de Máquina. Na seção 2.5 foram apresentadas as medidas de avaliação, relacionadas ao desempenho obtido nas predições realizadas, as quais serão utilizadas neste trabalho para avaliar o desempenho dos modelos testados.

CAPÍTULO

3

Trabalhos Relacionados

Neste capítulo serão apresentados trabalhos correlatos que tem como tema o uso de técnicas de Aprendizado de máquina para auxiliar a produção animal. Foi feita a busca por pesquisas com aplicação de técnicas de AM dentro do contexto da pecuária e com palavras-chave como, qualidade da carne ou carcaça, peso de animais ou carcaça. Dentre as pesquisas encontradas foram selecionadas aquelas dedicadas ao estudo de previsão e estimativa de características animais que são importantes para a otimização dos resultados econômicos, com exceção daquelas que utilizam imagens em seu conjunto de dados, uma vez que não é a proposta deste trabalho.

Na pecuária de corte, a lucratividade dos produtores geralmente está intimamente relacionada a algumas características específicas dos animais que indicam a qualidade da carne, principalmente o peso, dessa forma encontramos na literatura algumas abordagens nesse contexto. No trabalho de [27] foi apresentado um classificador SVM para estimar os o peso de cada animal, e para isso os autores utilizaram o peso dos animais em diferentes idades. O objetivo também estava em mostrar um método que supera a regressão separada de cada animal quando há apenas alguns pesos disponíveis e busca-se previsões com pesagem superior a 100 dias.

Em [28], utilizando modelos *Support Vector Machines for Regression* (SVR), os autores apresentam uma função para predizer o peso da carcaça de bovinos de corte da raça *Asturiana de los Valles* utilizando algumas medidas morfológicas dos animais tomadas dias antes do abate, os resultados obtidos mostram que é possível prever o peso da carcaça 150 dias antes do abate com um erro médio absoluto de 4,27%.

Além do peso, outros trabalhos tinham como objetivo prever outras carac-

terísticas que são potencialmente influenciadores na determinação do preço da carne. No estudo de [29] cinco tipos diferentes de algoritmo de AM, *Deep Learning* (DL), *Gradient Boosting Tree* (GBT), *K-Nearest Neighbor* (KNN), *Model Tree* (MT) e *Random Forest* (RF) foram empregados para prever cinco características morfológicas de ovelhas. O conjunto de dados era composto pelas informações do peso vivo e de carcaça, valores morfológicos e registros ambientais, totalizando 101 atributo preditivo. O autores demonstraram que os métodos de aprendizado de máquina são eficazes para prever características de carcaça em ovinos, principalmente *Random Forest* que superou a todos os outros.

Na Coréia, o peso corporal, espessura da gordura posterior e o marmoreio, são os principais determinantes do grau de carcaça e, consequentemente, seu preço. No trabalho de [30] foram aplicados quatro modelos, *multilayer perceptron* (MLP), *model tree* (MT), *random forest* (RF) e SVM utilizando Otimização Sequencial Mínima (SMO), para prever seis características de carcaça, inclusive as características citadas anteriormente. Na maioria dos cenários, o SMO e MT apresentaram desempenho relativamente melhor do que os outros métodos.

A autora de [31] comparou a eficácia dos métodos tradicionais de regressão linear e regressão linear generalizada das abordagens de AM (RF e Redes Neurais Multicamadas) para prever o peso da carcaça bovina, a maturidade, o acabamento e a qualidade de carcaça. O conjunto de dados continha informações sobre mais de 4 milhões de bovinos de corte de 5.204 fazendas correspondendo a 4,3% da produção do Brasil, além das informações, de participação da fazenda em programa de aconselhamento técnico, produtos nutricionais utilizados, variáveis econômicas relacionadas à produção de carne bovina, fertilidade do solo e classificação climática. O RF foi o melhor modelo para prever o peso da carcaça e maturidade, enquanto a regressão linear generalizada foi o melhor para acabamento e qualidade de carcaça. Mais do que a previsão, o estudo extraiu informações relevantes; padrões relacionadas ao campo, como a nutrição usada pela fazenda se mostraram importantes como preditores de acabamento; qualidade do solo e clima para prever a maturidade; participação em um programa de orientação técnica para a previsão de todas as variáveis de produção e qualidade da carne, indicando a importância de conhecimento especializado.

A base de dados do Programa Precoce MS foi objeto de estudo em outros trabalhos, o primeiro [12] aplicou em um conjunto de dados, fornecido pela Associação Sul-mato grossense de Produtores de Novilho Precoce (ASPNP), meios computacionais como armazém de dados DW, consultas analíticas online (OLAP) e mineração de dados, com o objetivo de descobrir padrões e rela-

cionamentos vinculados ao grau de acabamento e ao rendimento da carcaça. Como resultado foi construído um portal Web para acesso aos dados armazenados no DW de forma fácil e eficaz por meio de consultas OLAP, painéis gráficos e relatório. No entanto o autor encontrou dificuldade para alcançar uma boa precisão para prever o rendimento da carcaça, pois o atributo mais relevante identificado nos experimentos, o peso do animal vivo, era armazenado no conjunto de dados na forma de média de todos os animais de um lote abatido e não de forma individualizada, sendo este o atributo essencial para o cálculo de rendimento de carcaça e que também influenciava diretamente no grau de acabamento. Como solução foi realizado um estudo de caso com um novo conjunto de dados obtidos de estudos científicos controlados da Embrapa para simular um cenário ideal, ou seja, que o peso vivo seja armazenado de forma individual por animal, resultando na melhora da taxa de acerto médio se comparado aos resultados obtidos utilizando um conjunto de dados da ASPNP. Na seção de trabalhos futuros o autor sugere a criação de um novo conjunto de dados com informações mais precisas do animal e informações do sistema de produção "porteira a dentro", ou seja, dados que informem como esses animais foram criados antes de serem abatido, e dessa forma obter resultados de mineração de dados mais assertivos.

O segundo trabalho [13], inspirado nos resultados e propostas do trabalho de [12] teve como objetivo principal a construção de um classificador do grau de acabamento da carcaça, por meio de técnicas de MD, uma vez que o produtor é bonificado apenas pelas carcaças classificadas com grau de acabamento 2, 3 e 4. O conjunto de dados utilizado era composto pelas informações de abate de bovinos, cadastrados no programa estadual Precoce MS, e do processo produtivo de cada estabelecimento rural correspondente, informações mais precisas se comparado ao dados utilizados por Mota. Dessa forma obteve-se resultados de taxa de precisão de aproximadamente 70,45% e 70,11% para os algoritmos RFC e SVM. O autor encontrou como características mais relevantes para o sucesso da bonificação: o peso da carcaça, o sexo, o mês de abate, a maturidade do animal e a localização da fazenda. O desbalanceamento do conjunto de dados de abate foi uma das dificuldades elencadas pelo autor, pois as classes 1 e 5 representam 0,48% e 0,02% dos dados respectivamente, o que afetou o desempenho dos algoritmos utilizados para a predição, fazendo-se necessário um estudo e comparação sobre técnicas de balanceamento de dados. Como trabalho futuro propõe otimizar os resultados com a aplicação de AM Hierárquico e *Deep Learning*.

Os resultados obtidos nos trabalhos mostram que é possível utilizar técnicas de MD para prever características dos animais. A maioria dos trabalhos obtiveram bons resultados com os algoritmos SVM e RFC e desta forma foram

escolhidos para serem aplicados a este trabalho.

3.1 Considerações Finais

Os resultados obtidos nos trabalhos mostram que é possível utilizar técnicas de mineração de dados para prever características dos animais. A maioria dos trabalhos obtiveram bons resultados com os algoritmos SVM e RFC e desta forma foram escolhidos para serem aplicados a este trabalho.

Os autores, em sua maioria, buscam a predição de valores de características que trazem rentabilidade aos produtores, o objetivo deste trabalho é considerar os valores dessas características de forma unificada e predizer em qual categoria de qualidade o animal será classificado. Apenas os trabalhos de [31] [29] incluíram em seu conjunto de dados informações complementares aos dados exclusivos do animal, tais como, dados do climáticos e dados do processo produtivo. Espera-se que além de prever a qualidade da carne este trabalho possa extrair informações relevantes, como foi apresentado em [31], que contribuam com a produção de carne sul-mato-grossenses.

CAPÍTULO

4

Proposta de Trabalho

Hoje não há uma metodologia adequada para todos projetos de MD, mas alguns são mais comumente empregados do que outros. De acordo com as pesquisas realizadas por [32], em suma, foi verificado que a maioria das empresas utilizam o CRISP-DM ou usam métodos e processos próprios da organização. Para fortalecer essa visão, no período de 2015 a 2019, podemos encontrar um grande número de estudos que aplicam ou fazem adaptações do CRISP-DM e em domínios diferentes: saúde [33] [34], engenharia [35] [36], educação [37],produção [38] e análises de dados governamentais abertos [39].

Este trabalho, que tem como área de estudo a pecuária, visa obter resultados que possam auxiliar produtores a atender critérios de qualidade da carne e, consequentemente, obter benefícios econômicos. Uma vez que o CRISP-DM nasceu de uma iniciativa das indústrias e tem como proposta entender processos produtivos e guiar técnicas de Mineração de dados para que a tomada de decisão traga bons resultados econômicos, este modelo foi considerado um bom método a ser seguido. Nas próximas seções serão descritas como cada fase da metodologia CRISP-DM será aplicada na execução deste projeto.

4.1 Compreensão do domínio

A primeira fase da estrutura CRISP-DM é a compreensão do problema, o que envolve entender o programa de bonificação de carcaças do estado de Mato Grosso do Sul, denominado Precoce MS. O programa beneficia os produtores e os estimulam a produzir animais de qualidade, de acordo com às exigências do mercado consumidor, resultando ao produtor maior remuneração pelo animal classificado como novilho precoce.

Os produtores aprovados no programa recebem isenção do Imposto Sobre Circulação de Mercadorias e Serviços ICMS. Para a concessão do incentivo são utilizados critérios para avaliar o processo produtivo do estabelecimento rural, o animal abatido e a padronização do lote, ou seja, pelo menos 60% do lote a ser abatido deve ser composto por novilhos precoces. A Tabela 4.1 simplificada pode ser usada para exemplificar as regras do programa. As três primeiras colunas mostram os dados do animal abatido e representam até 70% do valor do incentivo. As colunas restantes referem-se aos dados da propriedade avaliada e representam até 30% do valor do incentivo. Sendo assim, o valor do incentivo a ser retornado pelo frigorífico ao produtor é a porcentagem que ele atingir na tabela (até o máximo de 67%) sobre o valor do ICMS. O restante do valor do ICMS é pago ao Estado pelo frigorífico.

Tabela 4.1: Esquema simplificado da classificação de carcaças bonificadas pelo programa Precoce MS [5]. Tipificação: F = fêmea, C = macho castrado e M = macho inteiro; Maturidade: J0 = apenas dentes de leite, J2 = dois dentes incisivos permanentes e J4 = quatro dentes incisivos permanentes.

70% Produto			**30% Processo Produtivo**		
Tipificação	**Maturidade**	**Acabamento**	**Avançado 30%**	**Intermediário 26%**	**Intermediário 26%**
M,C, F	J0	3,4	67	64	61
M,C, F	J2	3,4	62	59	56
C, F	J4	3,4	48	45	42
M, C, F	J0	2	62	59	56
M,C , F	J2	2	39	36	33
C, F	J4	2	22	19	16

Os processos produtivos de uma propriedade serão avaliados por meio de quatro critérios:

I Utilizem ferramentas que permitam a gestão sanitária individual de bovinos;

II Apliquem regras e conceitos de boas práticas agropecuárias;

III Apliquem tecnologias que promovam a sustentabilidade do sistema produtivo, em particular aquelas que visem à mitigação da emissão de carbono por meio de práticas de agropecuária de baixo carbono; e

IV Participem de associações de produtores visando à produção comercial sistematizada e organizada conforme padrões pré-estabelecidos para atendimento de acordos comerciais.

E dessa forma é possível categorizar os estabelecimentos em Simples se atendem a nenhum ou pelo menos um dos critérios, em Intermediário se atendem pelo menos dois critérios e em Avançado se atendem no mínimo três critérios. Compete ao responsável técnico pela propriedade, devidamente cadas-

trado no programa, prestar as informações necessárias à avaliação e classificação do estabelecimento, bem como promover a atualização das informações [5].

A classificação de um animal como novilho precoce é realizada por meio dos seguintes parâmetros: gênero (F = fêmea; C = macho castrado; e M = macho inteiro[1]) e a maturidade dentária conforme apresentada na Figura 4.1. Ainda que atendam aos demais critérios, não serão classificados para efeito do incentivo fiscal os animais com mais de quatro dentes.

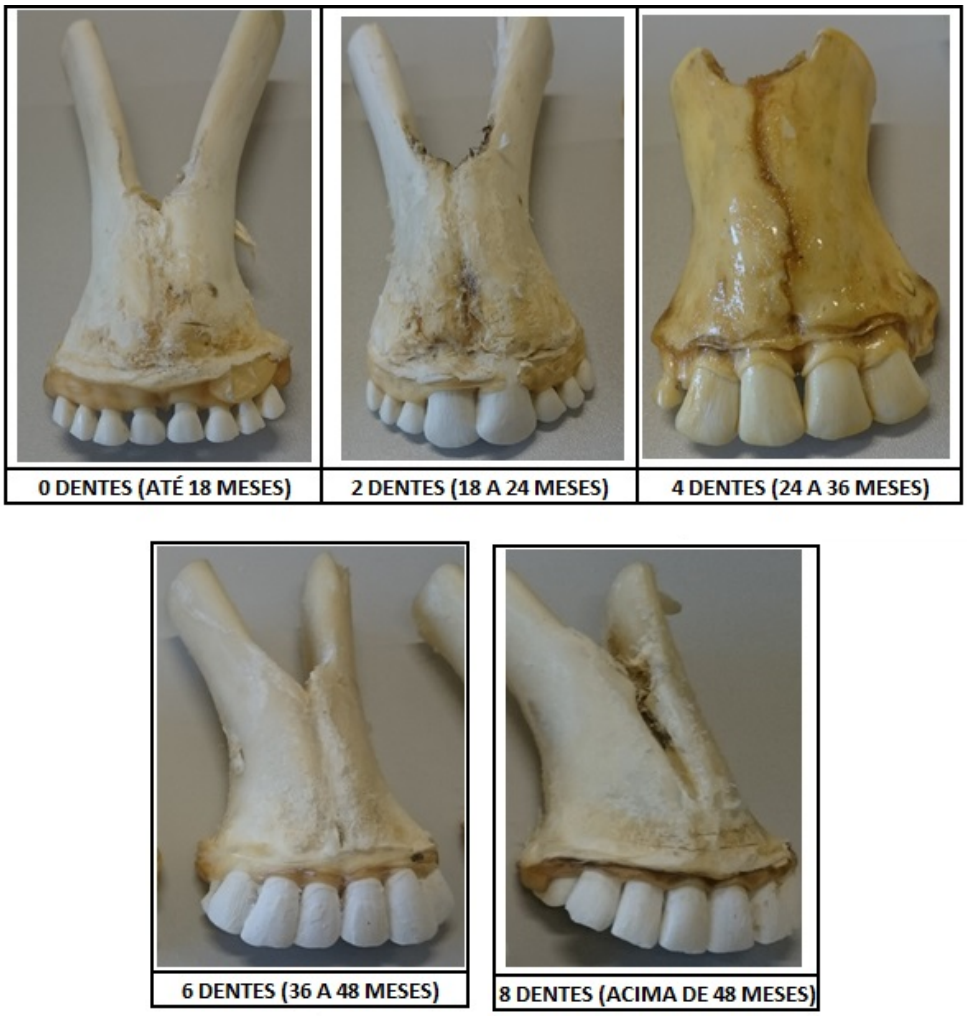


Figura 4.1: Maturidade dentária de novilhos precoces: J0 = 0 dentes, apenas dentes de leite; J2 = dois dentes incisivos permanentes; e J4 = quatro dentes incisivos permanentes; J6 = seis dentes permanentes; J8 = 8 dentes permanentes. Imagem adaptada de [4].

A fim de determinar a tipificação da carcaça, quanto ao seu acabamento, é adotado como parâmetro a medida da espessura da gordura subcutânea, denominando classes da seguinte forma: 1 = ausência total de gordura ou menos de 1 mm; 2 = com 1 a 3 mm; 3 = 3 a 6 mm; 4 = 6 e até 10 mm; 5 = acima de 10 mm. Ainda que atendam aos demais critérios o produtor rural

[1]Machos inteiros são considerados aqueles que apresentem testículos e não tenham sido submetidos a qualquer meio de castração regularmente admitidos [5]

não será bonificado se o animal for classificado com os acabamentos 1 e 5. Por fim, o peso do animal antes do abate deve ser no mínimo 180 quilogramas para fêmeas e 225 quilogramas para machos (inteiros ou castrados) [5].

Neste trabalho, o que pretende-se buscar é o desenvolvimento de um modelo para predizer a qualidade da carne através de algoritmos de AM supervisionados, e desta forma classificar a qualidade da carne em relação ao resultado da bonificação obtida. Sendo assim, a classe alvo será a classificação do animal conforme Tabela 4.2.

Tabela 4.2: Classificação da qualidade da carne em relação ao resultado da bonificação obtida.

Sexo	**Peso Macho (MIN)**	**Peso Fêmea (MIN)**	**Maturidade**	**Acabamento**	**Incentivo**	**Classificação**
M,C,F	225	180	J0	3 OU 4	67%	AAA
M,C,F	225	180	J2	3 OU 4	62%	AA
M,C,F	225	180	J0	2	62%	AA
C,F	225	180	J4	3 OU 4	48%	BBB
M,C,F	225	180	J2	2	39%	BB
C,F	225	180	J4	2	22%	C
M,C,F	>225	180	J6/J8	3 OU 4	0%	D
M,C,F	>225	180	J6/J8	>10	0%	D
M,C,F	>225	180	J6/J8	2	0%	D

O processo de MD poderá nos fornecer informações sobre o conjunto de dados em estudo, que podem ser relevantes para o processo de produção de carne. Algumas questões de negócios já foram levantadas pelos especialistas internos e o processo de MD será direcionado para tentar respondê-las:

- Quais são as características que mais influenciam na qualidade da carcaça?
- Quais são os fatores determinantes para diferenciar os produtores em relação a programas de bonificação?
- Quais são as características dos sistemas de produção preponderantes para a produção de carcaças de alta qualidade?
- Quais as características do sistema de produção / região do Estado / clima / fazem com que o produtor consiga obter as melhores bonificações?
- Qual região do Estado produz animais de melhor qualidade?

Este trabalho será realizado em parceria com a Embrapa Gado de Corte e conta com a colaboração de pesquisadores da Embrapa atuando como especialistas do negócio. O conjunto de dados utilizado neste trabalho será

proveniente da integralização de várias fontes de informação, realizada pelos pesquisadores da Embrapa.

O modelo será implementado utilizando a linguagem *Python*[2], que dispõe da biblioteca gratuita e de código aberto *Scikit-learn* que contém a implementação dos algoritmos de otimização de hiperparâmetros[3], bem como das principais técnicas de aprendizado de máquina [42].

4.2 Entendimento dos Dados

Esta fase é caracterizada pela coleta e exploração mais detalhada dos dados para análise. O conjunto de dados que será utilizado na execução deste trabalho foi integrado a partir de diferentes fontes para fornecer uma visão abrangente do contexto da produção de gado e serão descritas nas seções seguintes.

4.2.1 Dados do programa Precoce MS

Os dados referentes ao programa estadual Precoce MS foram entregues pela Superintendência de Gestão da Informação SGI, mediante solicitação à Secretaria de Estado de Meio Ambiente, Desenvolvimento Econômico, Produção e Agricultura Familiar (SEMAGRO), em dois conjuntos de dados diferentes. O primeiro conjunto de dados compreende os dados dos estabelecimentos rurais cadastrados no programa e seus respectivos processos produtivos. Neste conjunto de dados constou 1.595 estabelecimentos rurais cadastrados.

O processo produtivo de um estabelecimento rural é definido por meio de um questionário online. O preenchimento desses dados é feito por responsáveis técnicos, que devem ter formação como médico veterinário, engenheiro agrônomo ou zootecnista e são corresponsáveis por essas informações [5]. A Tabela 4.3 contém uma descrição de cada atributo e do seu formato para que seja possível perceber mais detalhadamente o que cada atributo representa.

O segundo conjunto de dados engloba todos os abates individuais de bovinos, no período de 09/02/2017 até 31/12/2018, com as respectivas classificações das carcaças obtidas. Neste conjunto de dados consta 1.107.798 animais abatidos e 21 características. A Tabela 4.4 contém uma descrição de cada atributo e do seu formato do conjunto de abates bovinos.

[2] *Python* é uma linguagem de programação interpretada, orientada a objetos e de alto nível com semântica dinâmica [40].

[3] Em *machine learning* há dois tipos de parâmetros a serem estimados: os parâmetros usuais de um algoritmo; e os parâmetros de ajuste, ou hiperparâmetros, que busca minimizar a probabilidade de erros de predição para dados futuros, reduzindo o viés indutivo, assim como reduzindo a variância do mesmo quando da previsão de novos casos [41].

Tabela 4.3: Descrição dos atributos contidas no banco de dados do Precoce MS.

Característica	Descrição	Formato
EstabelecimentoMunicipio	Município de localização da propriedade	Nominal
PossuiOutrosIncentivos	Existem outros incentivos? Se o produtor tem outros incentivos além do Precoce MS. Resposta Sim ou Não	Nominal
RecuperacaoPastagem	Pratica recuperação de pastagem? Caso resposta sim, as opções são fertirrigação, ILP, ILPF, IPF ou null	Nominal
reguaManejo	Faz controle de pastejo que atende aos limites mínimos de altura para cada uma das forrageiras ou cultivares exploradas, tendo como parâmetro a régua de manejo instituída pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa)? Resposta Sim ou Não	Nominal
BPA	O Estabelecimento rural apresenta atestado de Programas de Controle de Qualidade (Boas Práticas Agropecuárias - BPA/BOVINOS ou qualquer outro programa com exigências similares ou superiores ao BPA)? Resposta Sim ou Não	Nominal
aliancas_mercadologica	O Estabelecimento rural está envolvido com alguma organização que utiliza-se de mecanismos similares a aliança mercadológica para a comercialização do seu produto? Resposta Sim ou Não	Nominal
boaCoberturaVegetal	A área manejada apresenta boa cobertura vegetal, com baixa presença de invasoras e sem manchas de solo descoberto em, no mínimo, 80% da área total de pastagens (nativas ou cultivadas)? Resposta Sim ou Não	Nominal
erosaoLaminar	A área manejada apresenta sinais de erosão laminar ou em sulco igual ou superior a 20% da área total de pastagens (nativas ou cultivadas)? Resposta Sim ou Não	Nominal
rastreamentoSISBOV	Executa o rastreamento SISBOV? Resposta Sim ou Não	Nominal
ListaTrace	Faz parte da Lista Trace? Resposta Sim ou Não	Nominal

4.2.2 Dados climáticos

Os dados de clima foram obtidos no site do Centro de monitoramento do Clima e Tempo do Mato Grosso do Sul (CEMTEC)[43]. O CEMTEC disponibiliza até o quinto dia útil de cada mês, em planilhas eletrônicas o banco de dados meteorológico do mês anterior, constando dados de precipitação, umidade relativa, vento (intensidade e direção) e temperaturas provenientes dos sensores das 45 Estações Meteorológicas que existem em Mato Grosso do Sul. Os dados obtidos são diários e referentes ao período de abate dos animais no programa precoce, de 01/01/2017 a 31/12/2018.

Os boletins meteorológicos são disponibilizados na internet mensalmente, onde cada aba da planilha corresponde a uma variável climática, e cada linha

Tabela 4.4: Descrição dos atributos contidas no banco de dados do Precoce MS sobe os animais.

Característica	**Descrição**	**Formato**
data abate	Data do abate	Data
identificação propriedade	Identificador único para a propriedade rural	Numérico
cidade	Município de uma dada propriedade	Nominal
UF	Estado de uma dada propriedade	Nominal
tipificação	Ao abater um animal, o frigorífico registra a tipificação (Macho INTEIRO, Macho CASTRADO ou Fêmea)	Nominal
maturidade	Maturidade (Dente de leite, Dois dentes, Quatro dentes, Seis dentes ou Oito dentes)	Nominal
peso	O peso da carcaça (em kg)	Numérico
acabamento	grau de gordura da carcaça Magra - Gordura ausente, Gordura Escassa - 1 a 3 mm de espessura, Gordura Mediana - acima de 3 a até 6 mm de espessura, Gordura Uniforme - acima de 6 e até 10 mm de espessura ou Gordura Excessiva - acima de 10 mm de espessura)	Nominal

da planilha a uma estação meteorológica (Município), foi necessário fazer uma transposição dos dados e integração de todas as planilhas (21 planilhas, com 11 abas cada) para um arquivo único com todas as variáveis, datas e estações meteorológicas (Figura 4.2). A integração das planilhas em um só arquivo no formato .txt foi feita por meio de código criado na linguagem *Python*. Um primeiro passo foi fazer um comando *melt* para que cada planilha ficasse com apenas uma aba contendo todas as informações necessárias, posteriormente foi realizada a junção com os demais arquivos de cada mês e, finalmente, uma transposição dos dados de forma que o banco de dados final contivesse a lista de municípios (estações meteorológicas, data, e variáveis) conforme Figura 4.3.

Após a transformação dos dados e com base neles, foi calculado o Índice de Temperatura e Umidade (ITU) que é baseado na relação entre a temperatura ambiente e a umidade relativa do ar e é utilizado como uma medida de conforto animal. O cálculo do ITU foi feito por meio da fórmula desenvolvida por [44]:

$$ITU = 0,8\mathrm{Ta} + \mathrm{UR}(\mathrm{Ta} - 14,3)/100 + 46,3 \tag{4.1}$$

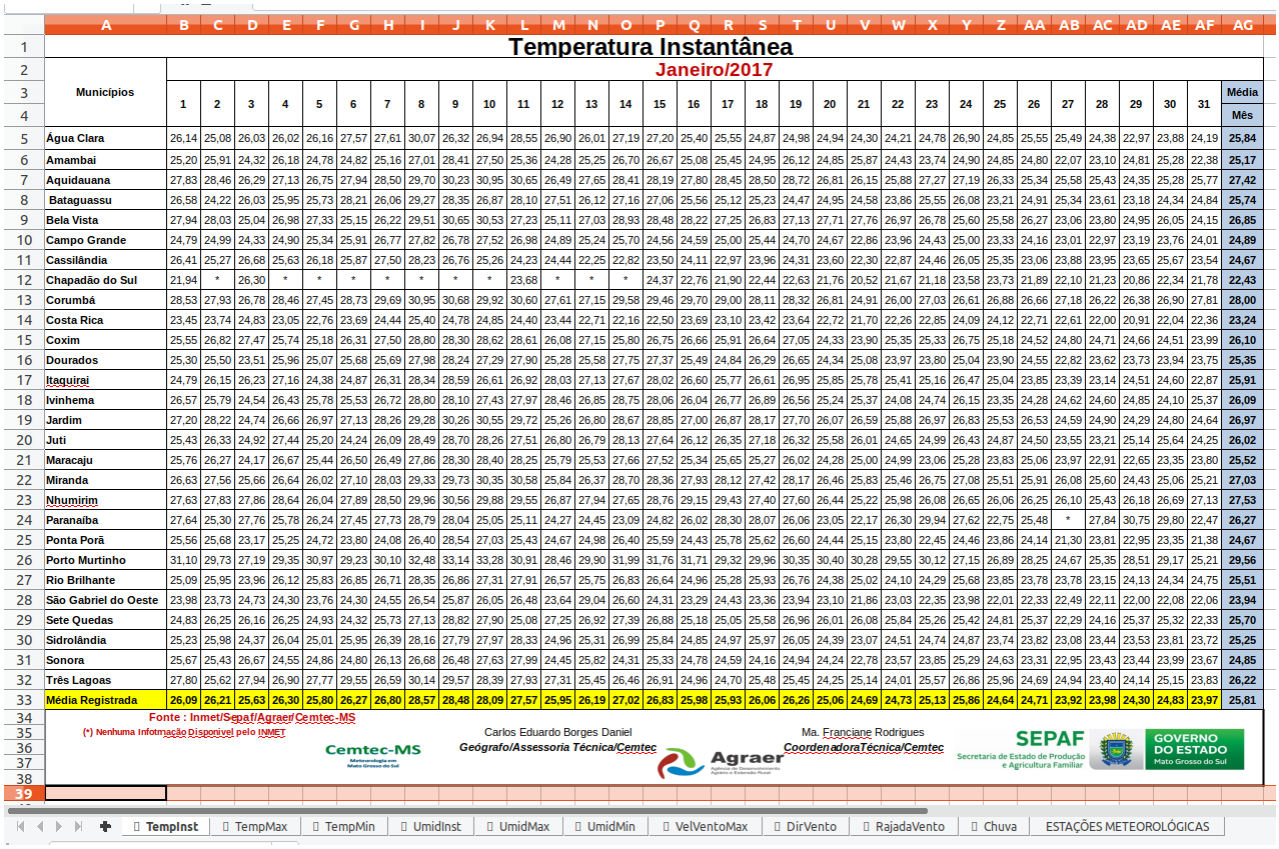
Onde:

ITU - valor médio diário do índice de temperatura e umidade;

Ta - temperatura do ar (ºC); e

UR - umidade relativa do ar (%).

A partir da fórmula citada anteriormente, foram criadas novas variáveis para o conjunto de dados. Utilizando a temperatura e a umidade máxima do dia foi criada a variável formITUmax e a utilizando a temperatura e umidade instantânea do dia, a variável formITUinst. Os resultados o obtidos do cálculo o ITU foram divididos em classes relacionadas ao conforto térmico dos animais



Temperatura Instantânea

Janeiro/2017

Municípios	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	Média Mês
Água Clara	26,14	25,08	26,03	26,02	26,16	27,57	27,61	30,07	26,32	26,94	28,55	26,90	26,01	27,19	27,20	25,40	25,55	24,87	24,98	24,94	24,30	24,21	24,78	26,90	24,85	25,55	25,49	24,38	22,97	23,88	24,19	25,84
Amambai	25,20	25,91	24,32	26,18	24,78	24,82	25,16	27,01	28,41	27,50	25,36	24,28	25,25	26,70	26,67	25,08	25,45	24,95	26,12	24,85	25,87	24,43	23,74	24,90	24,85	24,80	22,07	23,10	24,81	25,28	22,38	25,17
Aquidauana	27,83	28,46	26,29	27,13	26,75	27,94	28,50	29,70	30,23	30,95	30,65	26,49	27,65	28,41	28,19	27,80	28,45	28,50	28,72	26,81	26,15	25,88	27,27	27,19	26,33	25,34	25,58	25,43	24,35	25,28	25,77	27,42
Bataguassu	26,58	24,22	26,03	25,95	25,73	28,21	26,06	29,27	28,35	26,87	28,10	27,51	26,12	27,16	27,06	25,56	25,12	25,23	24,47	24,95	24,58	23,86	25,55	26,08	23,21	24,91	25,34	23,61	23,18	24,34	24,84	25,74
Bela Vista	27,94	28,03	25,04	26,98	27,33	25,15	26,22	29,51	30,65	30,53	27,23	25,11	27,03	28,93	28,48	28,22	27,25	26,83	27,13	27,71	27,76	26,97	26,78	25,60	25,58	26,27	23,06	23,80	24,95	26,05	24,15	26,85
Campo Grande	24,79	24,99	24,33	24,90	25,34	25,91	26,77	27,82	26,78	27,52	26,98	24,89	25,24	25,70	24,56	24,59	25,00	25,44	24,70	24,67	22,86	23,96	24,43	25,00	23,33	24,16	23,01	22,97	23,19	23,76	24,01	24,89
Cassilândia	26,41	25,27	26,68	25,63	26,18	25,87	27,50	28,23	26,76	25,26	24,23	24,44	22,25	22,82	23,50	24,11	22,97	23,96	24,31	23,60	22,30	22,87	24,46	26,05	25,35	23,06	23,88	23,95	23,65	25,67	23,54	24,67
Chapadão do Sul	21,94	*	26,30	*	*	*	*	*	*	*	23,68	*	*	*	24,37	22,76	21,90	22,44	22,63	21,76	20,52	21,67	21,18	23,58	23,73	21,89	22,10	21,23	20,86	22,34	21,78	22,43
Corumbá	28,53	27,93	26,78	28,46	27,45	28,73	29,69	30,95	30,68	29,92	30,60	27,61	27,15	29,58	29,46	29,70	29,00	28,11	28,32	26,81	24,91	26,00	27,03	26,61	26,88	26,66	27,18	26,22	26,38	26,90	27,81	28,00
Costa Rica	23,45	23,74	24,83	23,05	22,76	23,69	24,44	25,40	24,78	24,85	24,40	23,44	22,71	22,16	22,50	23,69	23,10	23,42	23,64	22,72	21,70	22,26	22,85	24,09	24,12	22,71	22,61	22,00	20,91	22,04	22,36	23,24
Coxim	25,55	26,82	27,47	25,74	25,18	26,31	27,50	28,80	28,30	28,62	28,61	26,08	27,15	25,80	26,75	26,66	25,91	26,64	27,05	24,33	23,90	25,35	25,33	26,75	25,18	24,52	24,80	24,71	24,66	24,51	23,99	26,10
Dourados	25,30	25,50	23,51	25,96	25,07	25,68	25,69	27,98	28,24	27,29	27,90	25,28	25,58	27,75	27,37	25,49	24,84	26,29	26,65	24,34	25,08	23,97	23,80	25,04	23,90	24,55	22,82	23,62	23,73	23,94	23,75	25,35
Itaquiraí	24,79	26,15	26,23	27,16	24,38	24,87	26,31	28,34	28,59	26,61	26,92	28,03	27,13	27,67	28,02	26,60	25,77	26,61	26,95	25,85	25,78	25,41	25,16	26,47	25,04	23,85	23,39	23,14	24,51	24,60	22,87	25,91
Ivinhema	26,57	25,79	24,54	26,43	25,78	25,53	26,72	28,80	28,10	27,43	27,97	28,46	26,85	28,75	28,06	26,04	26,77	26,89	26,56	25,24	25,37	24,08	24,74	26,15	23,35	24,28	24,62	24,60	24,85	24,10	25,37	26,09
Jardim	27,20	28,22	24,74	26,66	26,97	27,13	28,26	29,28	30,26	30,55	29,72	25,26	26,80	28,67	28,85	27,00	26,87	28,17	27,70	26,07	26,59	25,88	26,97	26,83	25,53	26,53	24,59	24,90	24,29	24,80	24,64	26,97
Juti	25,43	26,33	24,92	27,44	25,20	24,24	26,09	28,49	28,70	28,26	27,51	26,80	26,79	28,13	27,64	26,12	26,35	27,18	26,32	25,58	26,01	24,65	24,99	26,43	24,87	24,50	23,55	23,21	25,14	25,64	24,25	26,02
Maracaju	25,76	26,27	24,17	26,67	25,44	26,50	26,49	27,86	28,30	28,40	28,25	25,79	25,53	27,66	27,52	25,34	25,65	25,27	26,02	24,28	25,00	24,99	23,06	25,28	23,83	25,06	23,97	22,91	22,65	23,35	23,80	25,52
Miranda	26,63	27,56	25,66	26,64	26,02	27,10	28,03	29,33	29,73	30,35	30,58	25,84	26,37	28,70	28,36	27,93	28,12	27,42	28,17	26,46	25,83	25,46	26,75	27,08	25,51	25,91	26,08	25,60	24,43	25,06	25,21	27,03
Nhumirim	27,63	27,83	27,86	28,64	26,04	27,89	28,50	29,96	30,56	29,88	29,55	26,87	27,94	27,65	28,76	29,15	29,43	27,40	27,60	26,44	25,22	25,98	26,08	26,65	26,06	26,25	26,10	25,43	26,18	26,69	27,13	27,53
Paranaíba	27,64	25,30	27,76	25,78	26,24	27,45	27,73	28,79	28,04	25,05	25,11	24,27	24,45	23,09	24,82	26,02	28,30	28,07	26,06	23,05	22,17	26,30	29,94	27,62	22,75	25,48	*	27,84	30,75	29,80	22,47	26,27
Ponta Porã	25,56	25,68	23,17	25,25	24,72	23,80	24,08	26,40	28,54	27,03	25,43	24,67	24,98	26,40	25,59	24,43	25,78	25,62	26,60	24,44	25,15	23,80	22,45	24,46	23,86	24,14	21,30	23,81	22,95	23,35	21,38	24,67
Porto Murtinho	31,10	29,73	27,19	29,35	30,97	29,23	30,10	32,48	33,14	33,28	30,91	28,46	29,90	31,99	31,76	31,71	29,32	29,96	30,35	30,40	30,28	29,55	30,12	27,15	26,89	28,25	24,67	25,35	28,51	29,17	25,21	29,56
Rio Brilhante	25,09	25,95	23,96	26,12	25,83	26,85	26,71	28,35	26,86	27,31	27,91	26,57	25,75	26,83	26,64	24,96	25,28	25,93	26,76	24,38	25,02	24,10	24,29	25,68	23,85	23,78	23,78	23,15	24,13	24,34	24,75	25,51
São Gabriel do Oeste	23,98	23,73	24,73	24,30	23,76	24,30	24,55	26,54	25,87	26,05	26,48	23,64	29,04	26,60	24,31	23,29	24,43	23,36	23,94	23,10	21,86	23,03	22,35	23,98	22,01	22,33	22,49	22,11	22,00	22,08	22,06	23,94
Sete Quedas	24,83	26,25	26,16	26,25	24,93	24,32	25,73	27,13	28,82	27,90	25,08	27,25	26,92	27,39	26,88	25,18	25,05	25,58	26,96	26,01	26,08	25,84	25,26	25,42	24,81	25,37	22,29	24,16	25,37	25,32	22,33	25,70
Sidrolândia	25,23	25,98	24,37	26,04	25,01	25,95	26,39	28,16	27,79	27,97	28,33	24,96	25,31	26,99	25,84	24,85	24,97	25,97	26,05	24,39	23,07	24,51	24,74	24,87	23,74	23,82	23,08	23,44	23,53	23,81	23,72	25,25
Sonora	25,67	25,43	26,67	24,55	24,86	24,80	26,13	26,68	26,48	27,63	27,99	24,45	25,82	24,31	25,33	24,78	24,59	24,16	24,94	24,24	22,78	23,57	23,85	25,29	24,63	23,31	22,95	23,43	23,44	23,99	23,67	24,85
Três Lagoas	27,80	25,62	27,94	26,90	27,77	29,55	26,59	30,14	29,57	28,39	27,93	27,31	25,45	26,46	26,91	24,96	24,70	25,48	25,45	24,25	25,14	24,01	25,57	26,86	25,96	24,69	24,94	23,40	24,14	25,15	23,83	26,22
Média Registrada	26,09	26,21	25,63	26,30	25,80	26,27	26,80	28,57	28,48	28,09	27,57	25,95	26,19	27,02	26,83	25,98	25,93	26,06	26,26	25,06	24,69	24,73	25,13	25,86	24,64	24,71	23,92	23,98	24,30	24,83	23,97	25,81

Fonte : Inmet/Sepaf/Agraer/Cemtec-MS

(*) Nenhuma Infotmação Disponivel pelo INMET

Carlos Eduardo Borges Daniel
Geógrafo/Assessoria Técnica/Cemtec

Ma. Franciane Rodrigues
Coordenadora Técnica/Cemtec

Cemtec-MS · Agraer · SEPAF Secretaria de Estado de Produção e Agricultura Familiar · Governo do Estado Mato Grosso do Sul

TempInst | TempMax | TempMin | UmidInst | UmidMax | UmidMin | VelVentoMax | DirVento | RajadaVento | Chuva | ESTAÇÕES METEOROLÓGICAS

Figura 4.2: Exemplo da planilha com os dados meteorológicos.

Tipo_Variavel	Municipio	Data	Chuva	DirVento	RajadaVento	TempInst	TempMax	TempMin	UmidInst	UmidMax	UmidMin	VelVentoMax
0	Amambai	1/Fevereiro/2017	2.8	SO	28.80	22.4292	27.7	18.4	86.1667	98.0	57.0	12.24
1	Amambai	10/Fevereiro/2017	21.4	SO	54.36	23.3917	31.2	20.1	83.6250	98.0	55.0	20.88
2	Amambai	11/Fevereiro/2017	6.4	SO	32.04	24.7125	31.6	21.2	83.6667	97.0	40.0	14.40
3	Amambai	12/Fevereiro/2017	20.8	SO	29.52	24.8792	30.9	21.1	82.5833	97.0	51.0	7.92
4	Amambai	13/Fevereiro/2017	4.8	SO	24.48	25.0917	32.6	20.2	75.0417	95.0	44.0	8.64

Figura 4.3: Amostra do banco de dados de clima após a transformação.

[44], conforme a Tabela 4.5. A variável classifITUmax indica a classe na qual o Índice de temperatura e umidade máximo se encontra e a variável classifITUinst indica a classe na qual o Índice de temperatura e umidade instantânea se encontra.

Tabela 4.5: Classificação do ITU segundo Thom (1959).

Classe	ITU	Conforto térmico dos animais
1	>74	Ambiente confortável
2	Entre 74 e 78	Estresse moderado
3	Entre 78 e 84	Estresse forte
4	Acima de 84	Estresse severo

A Tabela 4.6 contém uma descrição de cada atributo da base final de dados climáticos após processamento.

Tabela 4.6: Descrição das variáveis contidas no banco de dados climáticos utilizados.

Variáveis	**Descrição**
Município	Município onde a estação meteorológica está situada, sendo um total de 45.
Data	Dados diários do dia 01/01/2017 a 31/01/2018
Chuva	Total de precipitação em milímetros durante o dia
DirVento	Direção do vento (norte, sul, leste, oeste)
RajadaVento	Um vento de curta duração, em geral com menos de 20 segundos, que tem velocidade pelo menos 18,5 km/h maior do que a média de velocidade que vinha sendo sendo observada antes dela acontecer
TempInst	Temperatura instantânea do dia medida em graus Celsius
TempMax	Temperatura máxima do dia em graus Celsius
TempMin	Temperatura mínima do dia em graus Celsius
UmidInst	Umidade instantânea do ar em porcentagem
UmidMax	Umidade máxima do ar no dia em porcentagem
UmidMin	Umidade mínima do ar no dia em porcentagem
VelVentoMax	Velocidade do vento máxima em km/h
formITUinst	Índice de temperatura e umidade instantânea
formITUmax	Índice de temperatura e umidade máxima
classif ITUmax	Classe na qual o Índice de temperatura e umidade máximo se encontra
classif ITUinst	Classe na qual o Índice de temperatura e umidade instantânea se encontra

4.2.3 Dados de imagens de satélite

Índices de Vegetação

Os índices NDVI e EVI apresentam correlação com variáveis biofísicas da vegetação, como área foliar e biomassa verde, capazes de indicar a presença e o vigor vegetal em uma determinada área de interesse. As séries temporais desses índices vegetativos permitem que se acompanhe, ao longo do tempo, o comportamento da vegetação nesses locais. Assim, é possível identificar determinados tipos de uso e cobertura da terra, como áreas florestais, culturas agrícolas anuais, culturas agrícolas perenes e semiperenes, pastagens, entre outros, bem como seus processos de transição ao longo do tempo, como desmatamentos, conversão de sistemas agropecuários, intensificações agrícolas, etc. Os dados disponibilizados podem ser utilizados para atividades relacionadas ao monitoramento da produção agrícola e do meio ambiente, podendo, inclusive, apoiar a verificação de perdas agrícolas no âmbito do seguro rural.

NDVI é a abreviação da expressão em inglês para *Normalized Difference Vegetation Index*, o que equivale em português a Índice de Vegetação da Diferença Normalizada. Serve para analisar a condição da vegetação natural ou agrícola nas imagens geradas por sensores remotos. É frequentemente usado para medir a intensidade de atividade clorofiliana, inclusive comparando vários períodos distintos.

A energia captada e absorvida pelas plantas possui diversos espectros. O índice nada mais é que um cálculo realizado em cima dessas bandas espectrais. Esses espectros são captados por sensores, que na maioria dos casos

estão instalados em drones ou satélites, e posteriormente são tratados pela seguinte equação do índice:

$$\textbf{NDVI} = (\text{Infra Vermelho} - \text{Vermelho})/(\text{Infra Vermelho} + \text{Vermelho}) \quad (4.2)$$

A fórmula da equação 4.2 é realizada em cada pixel, respectivamente, nas bandas dos canais vermelho e infravermelho próximo, resultando em um valor final entre -1 e 1. Quanto mais próximo de 1, maior é a atividade vegetativa no local representado pelo pixel. Valores negativos ou próximos de 0 indicam áreas de água, edificações, solo nú, enfim, onde há pouca ou nenhuma atividade clorofiliana. Desta forma, as aplicações dos cálculos de **NDVI** na agricultura são várias, como por exemplo:

I Monitoramento de lavouras;

II Detecção de efeitos de secas;

III Detecção de danos provocados por pragas;

IV Estimativas de produtividade agrícola;

V Modelização hidrológica;

VI Mapeamento de áreas agrícolas.

EVI *(Enhanced Vegetation Index)* que significa Índice de Vegetação Melhorado. É um cálculo que leva em consideração o vermelho e infravermelho como o **NDVI**, mas utiliza a banda do azul para descontar influências atmosféricas no índice.

Com o índice **EVI**, busca-se aperfeiçoar o sinal da vegetação, reduzindo a influência do sinal do solo e da atmosfera sobre a resposta do dossel através da detecção em regiões com maiores densidades de biomassa [45]. A fórmula do **EVI**, segundo Justice et al. (1998) é definida na equação 4.3.

$$\textbf{EVI} = 2{,}5(\text{Infravermelho} - \text{Vermelho})/(L + \text{Infravermelho} + C_1 \times \text{Vermelho} - C_2 \times \text{Azul}) \quad (4.3)$$

Onde: L = Fator de correção para o solo, C_1 e C_2 = Coeficientes de ajuste para o efeito de aerossóis da atmosfera. Segundo Heute et al. (1997) e Justice et al. (1998) os coeficientes adotados pelo algoritmo de cálculo do **EVI** são: $L = 1$; $C_1 = 6$; $C_2 = 7{,}5$.

Os dados dos índices de vegetação (**NDVI** e **EVI**) estão disponíveis em diferentes plataformas tanto nacionais quanto internacionais. A Embrapa Informática, localizada em Campinas possui uma *API* denominada *API SATVeg*,

que é derivada do Sistema de Análise Temporal da Vegetação [46], uma ferramenta *web* desenvolvida pela Embrapa Informática Agropecuária, destinada à geração e visualização de perfis temporais dos índices vegetativos **NDVI** e **EVI** para o Brasil e toda a América do Sul, com o objetivo de apoiar atividades de gestão territorial, monitoramento agrícola e ambiental. Os índices vegetativos são gerados a partir de imagens multiespectrais fornecidas pelo sensor *MODerate resolution Imaging Spectroradiometer* (MODIS), a bordo dos satélites Terra e Aqua, da NASA, contemplando dados produzidos a partir de 2000 até a última data então disponibilizada por seu repositório oficial, com resolução temporal de 16 dias e resolução espacial de 250 metros.

Por meio desta *API*, foi solicitada a Embrapa Informática os índices **NDVI** e **EVI** de todas as áreas de pastagens do Estado de Mato Grosso do Sul para os anos 2017 e 2018, período correspondente aos dados de abate dos animais precoces do Estado.

As áreas de pastagem do Mato Grosso do Sul foram obtidas por meio do site Pastagem.org [47] que contém a série histórica das áreas de pastagens do Brasil, produzida para toda a extensão territorial brasileira, para os últimos últimos 33 anos (1985 a 2017), no âmbito do projeto do Mapbiomas. O *shapefile* da área de pastagem do estado do Mato Grosso do Sul para o ano de 2017 foi utilizado para extrair as séries temporais dos índices de vegetação (**NDVI** e **EVI**) na *API* SatVeg da Embrapa Informática para o período de 2017 e 2018.

O arquivo recebido da Embrapa Agroinformática continha dois arquivos em planilha eletrônica sendo um deles a média do **NDVI** da área de pastagem de cada município do Estado de MS, com intervalo de 8 dias para o período de 2017 e 2018, e o segundo com as médias do **EVI**, da mesma forma, conforme apresentado na Figura 4.4.

	A	B	C	D	E	F	G	H
1	**MUNICIPIO**	**COD_IBGE**	**NUM_PIXELS**	**1/1/2017**	**9/1/2017**	**17/1/2017**	**25/1/2017**	**2/2/2017**
2	AGUA CLARA	500020	71154	0,367679	0,408382	0,440632	0,418384	0,396556
3	ALCINOPOLIS	500025	13526	0,403706	0,438564	0,472092	0,462904	0,447779
4	AMAMBAI	500060	18908	0,412318	0,416978	0,409468	0,386548	0,393458
5	ANASTACIO	500070	8672	0,392357	0,41787	0,438387	0,424579	0,408529
6	ANAURILANDIA	500080	18166	0,418902	0,475237	0,472317	0,460124	0,448704
7	ANGELICA	500085	2260	0,465337	0,446758	0,454075	0,453606	0,434079
8	ANTONIO JOAO	500090	1788	0,427806	0,446056	0,47757	0,441388	0,454221
9	APARECIDA DO TABOADO	500100	9136	0,393042	0,433758	0,48818	0,444427	0,43365

Figura 4.4: Médias do EVI por municípios no período de 2017 a 2019.

4.2.4 Preços das commodities

Os preços das principais *commodities* associadas ao estudo (soja, milho e arroba do boi) foram obtidas por meio do site do CEPEA - Centro de pesquisas econômicas da Escola Superior de Agricultura Luiz de Queiroz (ESALQ), campus da Universidade de São Paulo [48].

Os dados de cotação são médias diárias dos anos de 2017 e 2018. No caso do boi gordo utilizou-se a cotação diária do indicador do boi gordo CEPEA/BR, do milho foi o indicador do milho ESALQ/BM&FBOVESPA e da soja indicador da soja ESALQ/BM&FBOVESPA - Paranaguá.

4.2.5 Integração dos dados

Por meio do Software *Statistical Analysis System* (SAS) todas as bases de dados foram integradas de forma que cada animal abatido contivesse os dados de todas as variáveis relativas à data de abate correspondente para o município em questão. A base de dados final contou com $1,107$ milhões de linhas e 91 colunas contendo todas as variáveis, em formato temporal, sendo que todos os parâmetros foram calculados por meio da média de dos últimos 7 dias, 15 dias, 30 dias, 90 dias, e 180 dias antes da data de abate.

4.2.6 Exploração dos dados

Nesta etapa, atividades de mineração de dados como técnicas de consultas, visualização e relatórios serão aplicadas no conjunto de dados, e dessa forma será possível identificar quais são os dados importantes para a resolução do problema, identificar padrões, obter *insights* e formular hipóteses.

A partir dessa análise exploratória dos dados podemos analisar a qualidade dos dados, a fim de encontrar problemas como, dados ausentes, anomalias e *outliers*, e quando apropriado, incluir gráficos para ilustrar as características dos dados.

4.3 Preparação dos Dados

Nesta etapa será realizada uma análise mais detalhada para preparar os dados para a modelagem. Esta análise vai permitir que seja efetuada uma seleção mais apropriada das variáveis a serem incluídas em cada modelo.

Alguns pontos importantes nesta etapa:

- Tratar valores nulos e dados inconsistentes;
- Criação de dados, se necessário (atributos derivados, atributos construídos a partir de um ou de muitos registros), em geral, tendem a vir de uma operação de agregação;
- Seleção de atributos relevantes para o modelo; e
- Seleção da amostra.

4.4 Modelagem

Nesta fase, serão definidas quais as técnicas de modelagem a serem testadas. Para atingir o objetivo deste trabalho serão utilizados modelos de aprendizado de máquina supervisionado, apresentados na seção 2.4.

De acordo com o modelo investigado pode haver a necessidade de retornar a fase anterior e preparar os dados para atender aos critérios do modelo, como é o caso da maioria dos algoritmos de aprendizado de máquina que requerem que os valores das características sejam numéricos [49], para isso as colunas com valores categóricos devem ser transformadas em valores numéricos.

4.5 Avaliação

Esta etapa avalia o grau em que o modelo atende aos objetivos do trabalho. Podemos analisar os modelos com relação aos critérios de sucesso elencados na primeira fase da metodologia. Os modelos gerados que atendem aos critérios selecionados, tornam-se os modelos aprovados.

Segundo o guia [18], é apropriado que seja feita uma revisão mais completa, a fim de determinar se há qualquer fator ou tarefa importante que de alguma forma não foi alcançada. E dessa forma podemos listar e propor ações futuras que podem aprimorar os resultados obtidos.

4.6 Aplicação

O modelo será apresentado para o especialista do negócio que fará sua avaliação. Deverá ser previsto que o código desenvolvido dos modelos seja de fácil integração à alguma solução web, que será disponibilizada aos produtores.

4.7 Cronograma de Execução

A Tabela 4.7 apresenta o cronograma de execução das atividades previstas para este projeto de pesquisa de mestrado.

Atividade 1 (A1). Cumprimento das disciplinas obrigatórias. Nesta atividade foram concluídas todas as disciplinas do programa de mestrado.

Atividade 2 (A2). Compreender os trabalhos que deram origem à proposta deste trabalho. A base de dados utilizada para este trabalho foi objeto de mais dois estudos em parceria com a Embrapa, mostrando resultados promissores e evidenciaram que ao incluir fatores do contexto da produção da carne no conjunto de dados, como os dados do processo produtivo, obtemos resultados mais precisos.

Tabela 4.7: Cronograma de atividades.

	2019		2020							2021					
	1º Sem.	2º Sem.	1º Sem.	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun
A1	X	X													
A2			X	X											
A3					X	X	X	X	X						
A4			X					X							
A5										X					
A6										X	X				
A7										X	X	X			
A8													X		
A9														X	
A10															X

Atividade 3 (A3). Busca por estudos correlatos e estudo dos conceitos que fundamentam este trabalho. A fim de investigar abordagens que utilizam algoritmos de Aprendizado de Máquina com suporte às atividades na área da pecuária. De acordo com os trabalhos encontrados nesta atividade, vimos que a utilização de Aprendizado de Máquina está em crescimento e gerando resultados promissores, auxiliando na tomada de decisão no processo produtivo. Foi possível identificar os algoritmos de AM mais utilizados e os que apresentaram melhor resultado. Para dar embasamento teórico a este trabalho, foi realizado um estudo pelas técnicas e métodos que serão utilizados.

Atividade 4 (A4) Aquisição do conjunto de dados. Neste trabalho será utilizado um conjunto de dados composto por atributos advindos de diversas fontes de informação, as características de cada fonte foram descritas para documentação e melhor entendimento dos dados.

Atividade 5 (A5) Exame de qualificação.

Atividade 6 (A6) Preparação do conjunto de dados. Nesta atividade os dados serão tratados para que os algoritmos possam ser aplicados, ou seja, eliminar qualquer tipo de inconsistência que possa interferir nos desempenho do algoritmo aplicado ao conjunto de dados. Como ferramenta técnica para realizar esta atividade e as demais, será utilizada a biblioteca *scikit-learn* do *Python*, o estudo desta ferramenta já foi iniciado e tem se demonstrado apropriada para a execução deste trabalho.

Atividade 7 (A7) Modelagem. Nesta atividade os algoritmos de AM escolhidos para serem executados serão aplicados ao conjunto de dados.

Atividade 8 (A8) Análise dos resultados obtidos. Avaliar os resultados obtido na A7, verificar se o modelo apresentou resultado satisfatório tecnicamente e para as necessidades do problema.

Atividade 9 (A9) Escrita da dissertação.

Atividade 10 (A10) Defesa da dissertação.

4.8 Considerações finais

O presente trabalho está em desenvolvimento, orientado pela metodologia CRISP-DM, cuja a primeira fase compreende o entendimento do problema, e para isso foram realizados estudos na literatura e houve suporte especializado de um pesquisador da Embrapa. Desta forma foi possível compreender o programa Precoce MS, seus critérios para garantir a carne e qualidade e sua importância para o cenário econômico. A base de dados que será utilizada para a aplicação de técnicas de MD, resultado do trabalho de integração de dados dos pesquisadores da Embrapa, foi analisada para que se pudesse compreender e descrever os atributos que a compõe. Partindo deste ponto de compreender o problema e definir objetivo que se deseja alcançar, as próximas etapas compreendem em aplicar as diversas técnicas de MD, considerando algumas abordagens encontradas nos trabalhos relacionados e as fases da metodologia definida para este trabalho.

Bibliografia

[1] knn classifier. Acessado em: 02/12/2020. Citado nas páginas vii e 9.

[2] Random forest classifier and regressor with python | machine learning | kgp talkie. Acessado em: 24/11/2020. Citado nas páginas vii e 10.

[3] Svm separating hyperplanes. Acessado em: 02/12/2020. Citado nas páginas vii e 10.

[4] Importância da idade de bovinos e a bonificação da indústria frigorífica. Acessado em: 02/12/2020. Citado nas páginas vii e 20.

[5] Resolução conjunta sefaz/sepaf no 69 de 30 de agosto de 2016. Acessado em: 02/11/2020. Citado nas páginas viii, 19, 20, 21, e 22.

[6] Usda.united states department of agriculture. brazil once again becomes the world's largest beef exporter.2019. Acessado em: 28/06/2020. Citado na página 1.

[7] Danilo Domingues Millen, Rodrigo Dias Lauritano Pacheco, Paula M Meyer, Paulo H Mazza Rodrigues, and Mario De Beni Arrigoni. Current outlook and future perspectives of beef production in brazil. *Animal frontiers*, 1(2):46–52, 2011. Citado na página 1.

[8] Albino Luchiari Filho et al. Produção de carne bovina no brasil qualidade, quantidade ou ambas. *Simpósio Sobre Desafios E Novas Tecnologias Na Bovinocultura De Corte-Simboi*, 2:2006, 2006. Citado na página 1.

[9] Pedro Eduardo de Felício. Sistemas de qualidade assegurada na cadeia de carne bovina: a experiência brasileira. 2001. Citado na página 1.

[10] Quem somos. Acessado em: 16/09/2020. Citado na página 2.

[11] Edson N Cáceres, Hemerson Pistori, Marcelo Augusto Santos Turine, Pedro P Pires, Cleber O Soares, and Camilo Carromeu. Computational precision livestock-position paper. In *II Workshop of the Brazilian Institute for Web Science Research*, number 02-03, page 9, 2011. Citado na página 2.

[12] Fernando Maia da Mota. Uma abordagem de análises olap e de mineração de dados para suporte à tomada de decisão no setor da pecuária de corte do brasil. Master's thesis, Universidade Federal de Mato Grosso do Sul, 2016. Citado nas páginas 2, 3, 15, e 16.

[13] Higor Henrique Picoli Nucci. Classificação do grau de acabamento de gordura da carcaça de bovinos de corte usando aprendizado de máquina. Master's thesis, Universidade Federal de Mato Grosso do Sul, 2019. Citado nas páginas 3 e 16.

[14] Foster PROVOST and Tom FAWCETT. Data science para negócios. *Tradução de Marina Boscatto*, 2016. Citado na página 5.

[15] Bhavani Thuraisingham. *Data Mining*. CRC Press, 1999. Citado na página 5.

[16] Ana Carolina Lorena, João Gama, and Katti Faceli. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC, 2000. Citado nas páginas 5, 7, 8, e 9.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996. Citado na página 5.

[18] P Chapman, J Clinton, R Kerber, T Khabaza, T Reinartz, and C Wirth Shearer. Crisp-dm 1.0. step-by-step data mining guide. Citado nas páginas 6 e 30.

[19] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Education, mar 1997. Citado nas páginas 7 e 8.

[20] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002. Citado na página 9.

[21] João Gama. Árvores de decisão. *Palestra ministrada no Núcleo da Ciência de Computação da Universidade do Porto, Porto*, 2002. Citado na página 9.

[22] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the

effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012. Citado na página 9.

[23] André Ricardo Gonçalves. Máquina de vetores suporte. *Acesso em*, 21, 2015. Citado na página 10.

[24] Moacir Antonelli Ponti and Gabriel B Paranhos da Costa. Como funciona o deep learning. *arXiv preprint arXiv:1806.07908*, 2018. Citado na página 11.

[25] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. Citado na página 11.

[26] José Augusto Baranauskas and Maria Carolina Monard. Reviewing some machine learning concepts and methods. *Relatórios Técnicos do ICMC/USP*, 102, 2000. Citado na página 12.

[27] Jaime Alonso, Alfonso Villa, and Antonio Bahamonde. Improved estimation of bovine weight trajectories using support vector machine classification. *Computers and electronics in agriculture*, 110:36–41, 2015. Citado na página 14.

[28] Jaime Alonso, Ángel Rodríguez Castañón, and Antonio Bahamonde. Support vector regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and electronics in agriculture*, 91:116–120, 2013. Citado na página 14.

[29] Saleh Shahinfar, Khama Kelman, and Lewis Kahn. Prediction of sheep carcass traits from early-life records using machine learning. *COMPUTERS AND ELECTRONICS IN AGRICULTURE*, 156:159–177, JAN 2019. Citado nas páginas 15 e 17.

[30] Saleh Shahinfar, Hawlader A. Al-Mamun, Byoungho Park, Sidong Kim, and Cedric Gondro. Prediction of marbling score and carcass traits in Korean Hanwoo beef cattle using machine learning methods and synthetic minority oversampling technique. *MEAT SCIENCE*, 161, MAR 2020. Citado na página 15.

[31] Vera Cardoso Ferreira Aiken. *Large Scale Integrated Analysis of Beef Production and Quality in Brazil*. The University of Wisconsin-Madison, 2019. Citado nas páginas 15 e 17.

[32] What main methodology are you using for your analytics, data mining, or data science projects? Acessado em: 02/11/2020. Citado na página 18.

[33] Olegas Niaksu. Crisp data mining methodology extension for medical domain. *Baltic Journal of Modern Computing*, 3(2):92, 2015. Citado na página 18.

[34] Nicholas M Njiru. *Clustering and visualizing the status of child health in kenya: A data mining approach*. PhD thesis, University of Nairobi, 2015. Citado na página 18.

[35] Michał Rogalewicz and Robert Sika. Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review*, 7(4):97–108, 2016. Citado na página 18.

[36] Steffen Huber, Hajo Wiemer, Dorothea Schneider, and Steffen Ihlenfeldt. Dmme: Data mining methodology for engineering applications–a holistic extension to the crisp-dm model. *Procedia Cirp*, 79:403–408, 2019. Citado na página 18.

[37] Layth Almahadeen, Murat Akkaya, and Arif Sari. Mining student data using crisp-dm model. *International Journal of Computer Science and Information Security*, 15(2):305, 2017. Citado na página 18.

[38] Franziska Schäfer, Christian Zeiselmair, Jonas Becker, and Heiner Otten. Synthesizing crisp-dm and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 190–195. IEEE, 2018. Citado na página 18.

[39] Paulo Henrique Cardoso et al. Ciência de dados aplicada a dados governamentais abertos sob a ótica da ciência da informação. 2019. Citado na página 18.

[40] What is python? executive summary. Acessado em: 02/12/2020. Citado na página 22.

[41] André Filipe de Moraes Batista and Alexandre Dias Porto Chiavegatto Filho. Machine learning aplicado à saúde. *Sociedade Brasileira de Computação*, 2019. Citado na página 22.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. Citado na página 22.

[43] Centro de monitoramento de tempo, do clima e dos recursos hídricos de mato grosso do sul (cemtec). Acessado em: 02/12/2020. Citado na página 23.

[44] Earl Crabill Thom. The discomfort index. *Weatherwise*, 12(2):57–61, 1959. Citado nas páginas 24 e 25.

[45] FJ Ponzoni and YE Shimabukuro. Sensoriamento remoto no estudo da vegetação. 2007. *São José dos Campos: Parêntese*, 2010. Citado na página 27.

[46] Satveg - sistema de análise temporal da vegetação. Acessado em: 02/12/2020. Citado na página 28.

[47] portal pastagem.org. Acessado em: 02/12/2020. Citado na página 28.

[48] Cepea - centro de estudos avançados em economia aplicada. Acessado em: 02/12/2020. Citado na página 28.

[49] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003. Citado na página 30.