**Formulário de documentação para ações de Mineração de dados**

**O Algoritmo Árvore de Decisão**

O algoritmo de Árvore de Decisão (no original em inglês – ***Decision Tree***), é um algoritmo de classificação e regressão fornecido pelo *Microsoft Analysis Sevices* para uso em modelagens de previsão de atributos discretos e contínuos.

Em relação aos atributos discretos, o algoritmo faz previsões baseadas nas colunas de entrada em um conjunto de dados. Ele usa valores, conhecidos como estados, dessas colunas para prever o estado de uma coluna que foi definida como previsível. Dito de outra maneira, o algoritmo identifica as colunas de entrada que são correlacionadas com a coluna previsível.

No que diz respeito aos atributos contínuos, o algoritmo usa a regressão linear para determinar onde uma árvore de decisão se divide.

**Como funciona o algoritmo**

O algoritmo Árvore de Decisão, funciona da seguinte maneira. Ele gera um modelo de Mineração de Dados criando uma série de divisões de árvore. Essas divisões, por sua vez, são representadas como nós. O algoritmo adiciona um nó ao modelo, toda vez que uma coluna de entrada é considerada correlacionada a uma coluna previsível. A determinação da divisão, ou seja, de um nó, depende do fato do algoritmo está prevendo uma coluna continua ou discreta.

**Configurando os Dados necessários para o modelo de árvore de decisão**

Os requisitos para um modelo de árvore de decisão, são os seguintes:

* **Uma única coluna key**: cada modelo deve conter apenas uma coluna chave, que pode ser numérica ou de texto. No entanto, não são permitidas chaves compostas.
* **Uma coluna previsível**: é necessário que tenha pelo menos uma coluna previsível e, o atributo previsível pode ser de diferentes tipos, tantos numéricos como discretos. No entanto, deve-se observa que, o aumento no número de atributos previsíveis pode aumenta o tempo de processamento.
* **Colunas de entrada**: requer colunas de entrada que podem ser continuas ou discretas. O aumento do número de colunas de entrada, também afetam o tempo de processamento.

Depois que o modelo é processado, os resultados são armazenados como um conjunto de padrões e estatísticas. Esse conjunto pode ser usado para explorar relações e fazer previsões.

**Estudo de Casos 1**: O Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte - **IFRN**, enfrenta nos atuais dias, um problema muito sério, que vem se agravando, ao longo dos anos, que são os relacionados aos altos índices de evasão escolar e reprovações. Portanto, é de interesse da comunidade do I**FRN**, como um todo, analisar, as suas bases de dados, para tentar descobrir, nessa massa de dados, algum comportamento que venha indicar as causas para esses problemas. A Tabele 1, apresenta de forma sucinta as atividades a serem desenvolvidas no processo de Mineração de dados (**MD**).

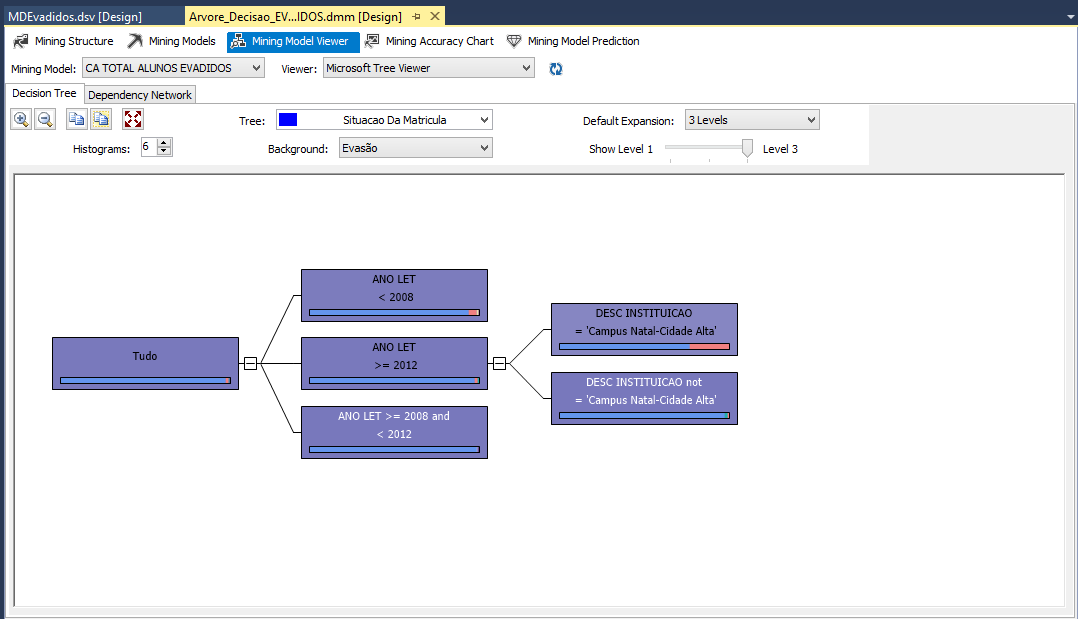
**Tabela 1: Tarefas de Mineração de Dados para o estudo de Caso 1.**

|  |  |  |
| --- | --- | --- |
| **Aplicação**: Sistema de apoio educacional do IFRN | | |
| **Estudo de Caso**: 1  **Objetivo**: o principal objetivo desse estudo de caso é tentar responder as seguintes perguntas:   * Quais são os fatores que mais implicam na evasão, reprovação e desistência dos alunos do IFRN.   **Tarefa de MD**: Árvore de Decisão | **Resumo**: Este estudo de caso tem como objetivo extrair modelos que, baseados no comportamento dos alunos, tendo como base, a base de dados do sistema acadêmico, identificar os fatores que, mais estão influenciando a evasão escolar em nossa instituição. | |
| **Expectativas quanto ao Modelo de Conhecimento**:   * Transparência do modelo; * Representação em regras | | |
| **Plano de Ação**:  Partição do Banco de dados em dados para treino e teste e Preenchimento de valores ausentes. | | |
| **Ciclo nº**: 1 |  |  |
| **Métodos** | **Obs. Sobre parâmetros** | **Resultados** |
| Preenchimento de valores ausentes |  |  |
| Codificação numérica – categórica (Discretização) | **Atributos**:  AlunosEvadidosID (chave)  ANO\_Letivo (entrada)  Desc Curso (entrada)  Desc Instituicao (entrada)  Periodo Letivo (previsão)  QTD Alunos (entrada)  Sigla Curso (entrada)  Situacao Matricula (previsão) |  |
| **Partição do BD em treino e teste** | 30% treino  70% teste |  |
|  |  |  |
|  |  |  |

Resultados obtidos, com o modelo, árvore de decisão aplicado sobre a tabela CA\_TOTAL\_ALUNOS\_EVADIDOS, do banco de dados **DDS Educacional**, modelo multidimensional do sistema **EDUCASE**.

A Figura 1 mostra a árvore de decisão gerada para a entrada fornecida. Como foi dito, o algoritmo de árvore de decisão, crie nós para cada entrada correlacionada ao atributo de previsão. Percebe-se, na figura 1 que alguns nós são mais escuros de que outros. Os nós mais escuros são os que concentram maior número de casos de evasão, no nosso estudo de caso. Veja que para o ano letivo maior do que 2012 e campus Natal-Cidade Alta, representam o maior índice de evasão escolar.

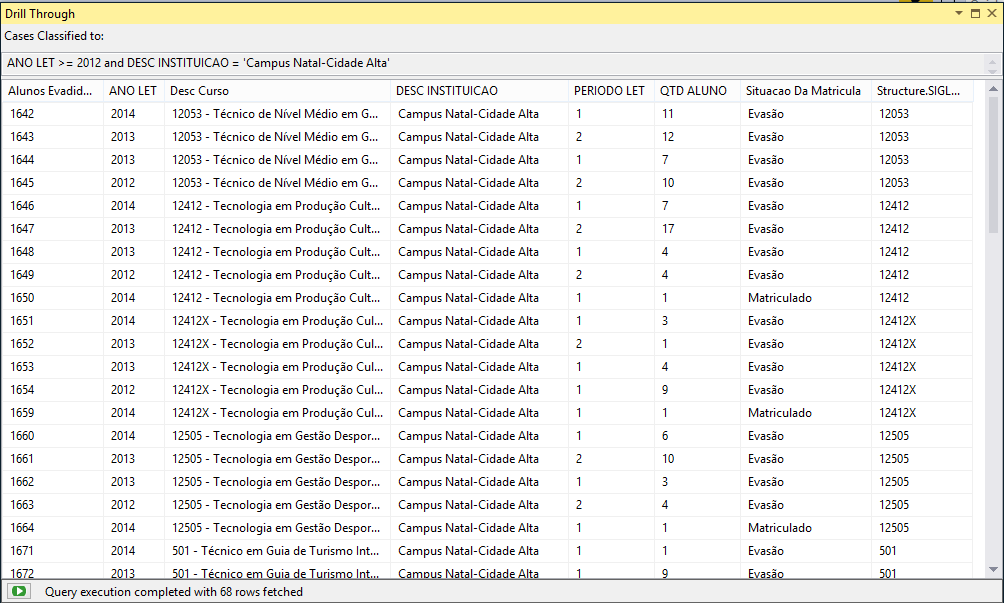
A figura 2, mostra-se um detalhe das informações, que estão representadas no nó Campus Natal-Cidade Alta.



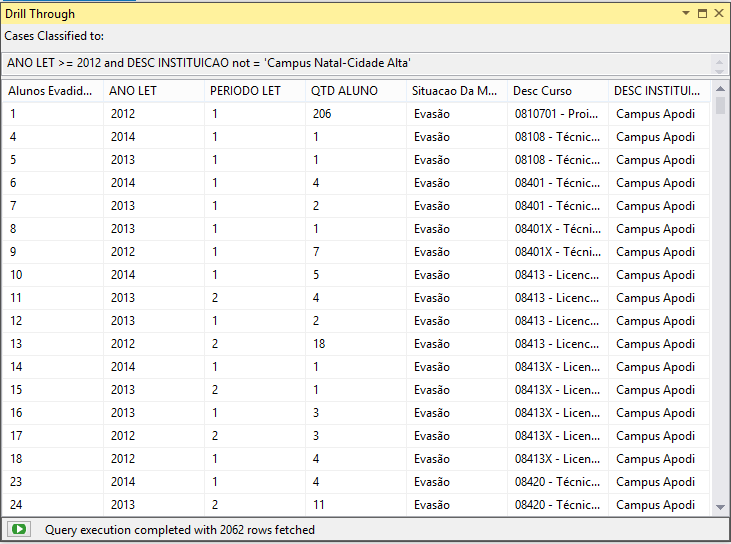
**Figura 1. A Árvore de Decisão criada para a análise de evasão do IFRN, como base nos dados acadêmicos. Fonte: autor**.

A figura 3, mostra-se um detalhe das informações, que estão representadas no nó que não é Campus Natal-Cidade Alta, ou seja, nos nós restantes.

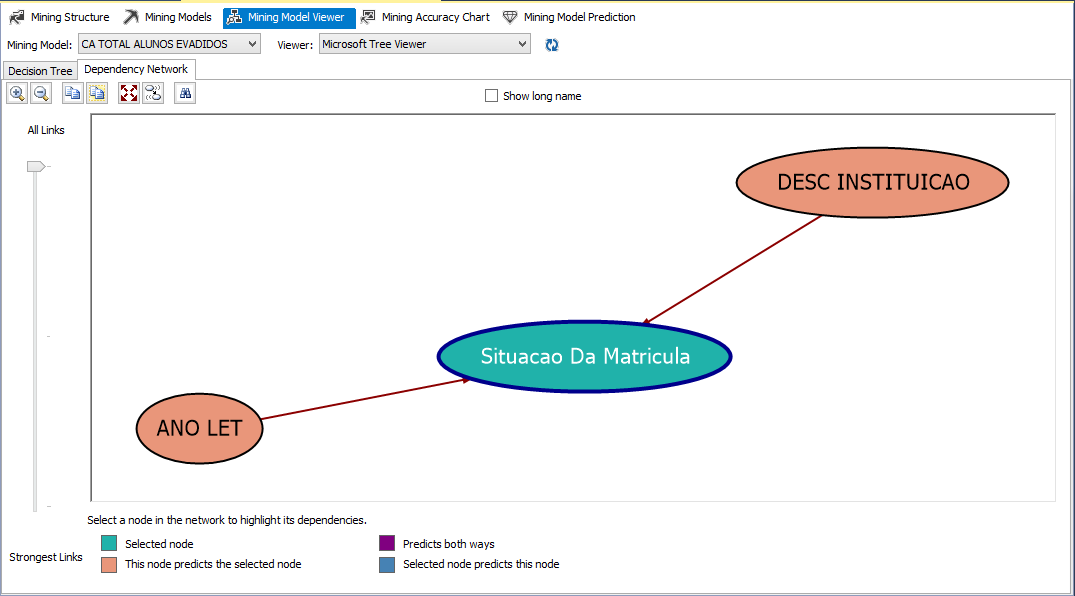
A figura 4, mostra a rede de dependência entre os atributos de entrada e atributo de previsão. Percebe-se que a correlação mais forte entre o atributo de previsão (Situação da Matricula) e os atributos de entrada, são os campos ano letivo e instituição, uma vez que a análise foi realizada, com o intuito de saber qual é os campis onde têm os maiores índices de evasão escolar.



**Figura 2 Dados detalhados para o nó Campus Natal-Cidade Alta**. **Fonte: autor**



**Figura 3 Dados detalhados para os outros campis. Fonte: autor**

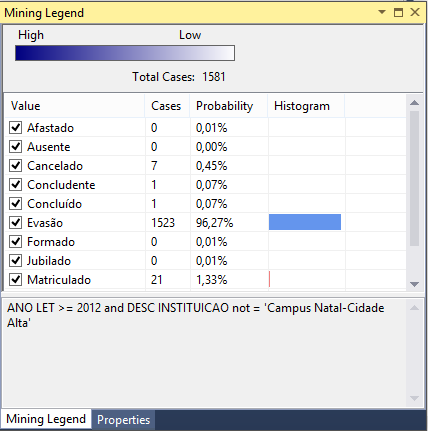


**Figura 4. Rede de dependência entre os atributos de entrada e o atributo de previsão**.

**Fonte: autor**.

A figura 5 exibe a legenda de mineração, onde são exibidas as seguintes informações, quando se clica em um nó da árvore de decisão. A figura em questão exibe as informações do nó Campus não Natal-Cidade Alta.

* O número de casos no nó, divididos em estados do atributo previsível.
* A probabilidade de cada caso do atributo previsível do no.
* Um histograma que inclui uma conta para cada estado do atributo previsível.
* As condições que são exigidas para alcançar um nó específico, também conhecido como caminho do nó.
* E a formula de regressão linear, é claro, para os modelos de regressão linear.



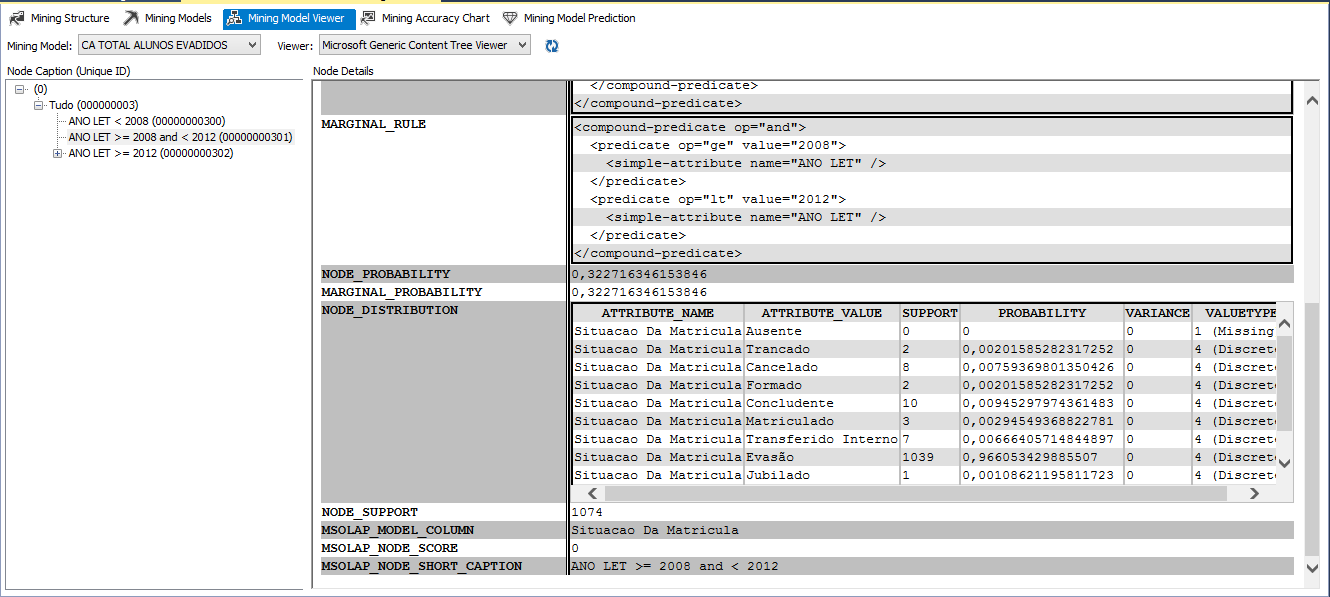
**Figura 5. Informações do nó clicado (Desc Instituicao not = campus Natal-Cidade Alta).**

**Fonte: autor**

A figura 6 mostra o conteúdo do modelo para um nó da árvore de decisão. Onde,

* **MARGINAL\_RULE**: uma descrição XML da regra que descreve o caminho do nó do pai do modelo para o atual.
* **NODE\_PROBABILITY**: a probabilidade associada a este nó.
* **MARGINAL\_PROBABILITY**: a probabilidade de que o nó seja alcançado a partir do nó pai.
* **NODE\_DISTRIBUTION**: uma tabela que contém o histograma de probabilidade do nó.
* **NODE\_SUPPORT**: o número de casos com suporte para este nó.
* **MOLAP\_MODEL\_COLUMN**: indica a coluna que contém o atributo previsível.
* **MOLAP\_NODE\_SCORE**: exibe uma pontuação associada ao nó.
* **MOLAP\_NODE\_SHORT\_CAPTION**: um rótulo usado para exibição.

Observação: Um modelo árvores de decisão não contém um nó separado que armazena estatística para todo modelo, diferentemente do nó de estatística marginais encontrados em um modelo de **rede Naive Bayes** ou **neural**. Em vez disso, o modelo cria uma árvore separada para cada atributo previsível, com um nó (Tudo) na parte superior da árvore.



**Figura 6. Conteúdo do modelo para um nó (ANO Let >= 2008 and < 2012). Fonte:** autor