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

**O Algoritmo Árvore de Decisão**

O algoritmo de Árvore de Decisão da Microsoft (no original em inglês – ***Decision Tree***), é um hibrido que incorpora métodos diferentes para a criação de uma árvore e dá suporte a várias tarefas analíticas, incluindo regressão, classificação e associação e, dá suporte a 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.**

**Estudo de Caso** 1: 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.

**Objetivo**: o principal objetivo desse estudo de caso é tentar responder a seguinte pergunta:

* Quais são os campis e os cursos desses campis que têm os maiores índices de evasão, reprovação, desistência, etc., de alunos do IFRN.

**Tarefa de MD**: Árvore de Decisão

**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:**

* Preenchimento de valores ausentes
* Codificação numérica – categórica (Discretização)

**Atributos**:

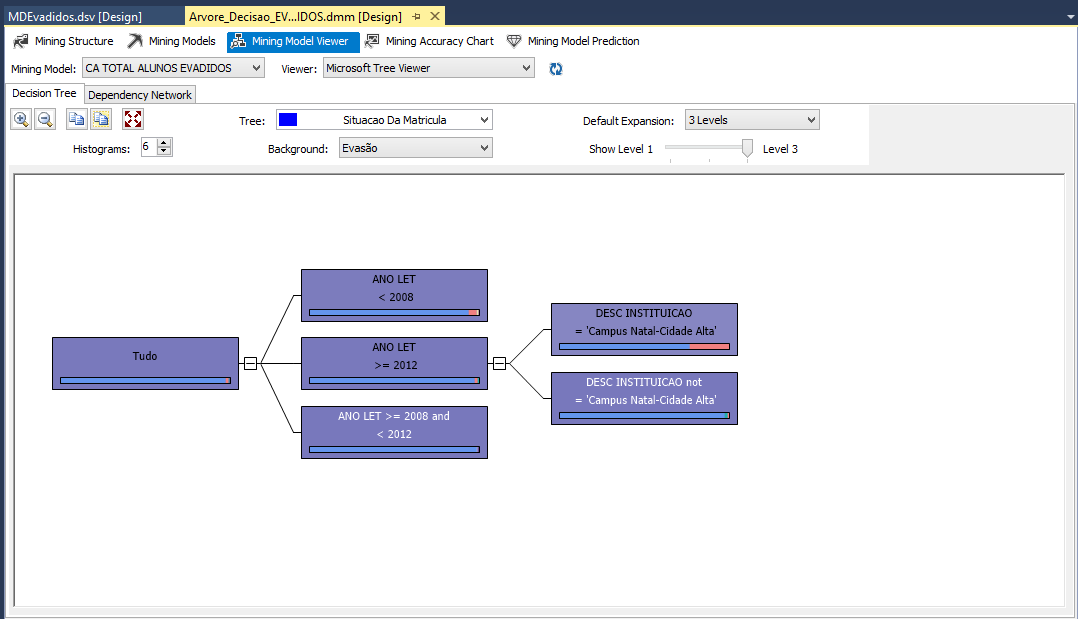
* AlunosEvadidosID (chave)
* ANO\_Letivo (entrada)
* Desc Curso (entrada)
* Desc Instituicao (entrada)
* Periodo Letivo (entrada)
* QTD Alunos (entrada)
* Sigla Curso (entrada)
* Situacao Matricula (previsão)

**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**.

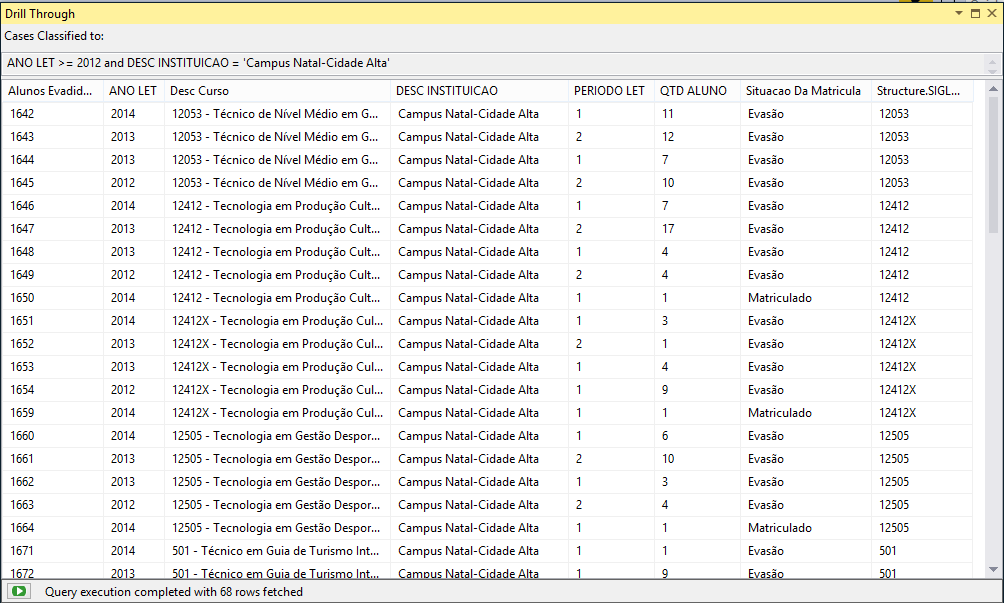
A Figura 1 mostra a árvore de decisão gerada para a entrada fornecida. Como foi dito, o algoritmo de árvore de decisão, cria nós para cada entrada correlacionada ao atributo de previsão. Percebe-se, na figura 1 que alguns nós são mais escuros de que outros. Os nós mais escuros são os que concentram maior 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, representa o maior índice de evasão escolar. Na figura 2, mostra-se um detalhe das informações, que estão representadas no nó Campus Natal-Cidade Alta.



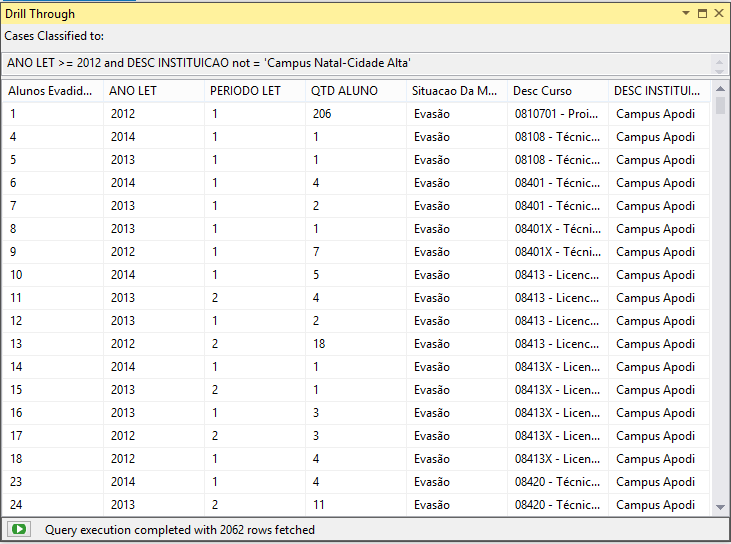
**Figura 1. A Árvore de Decisão criada para a análise de evasão do IFRN, como base nos dados acadêmicos. Fonte: autor**.

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.

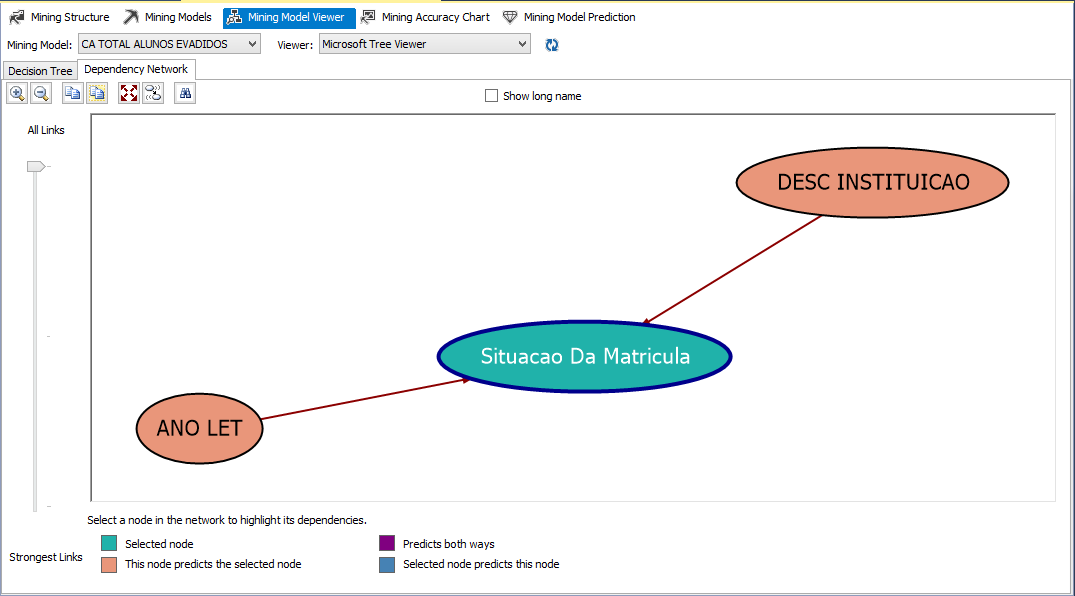
A figura 4, mostra a rede de dependência entre os atributos de entrada e atributo de previsão. Percebe-se que a correlação mais forte entre o atributo de previsão (Situação da Matricula) e os atributos de entrada, são os campos ano letivo e instituição, uma vez que a análise foi realizada, com o intuito 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 estrutura de mineração criada inicialmente, contém um único modelo de mineração de dados que se baseia no algoritmo Árvores de Decisão. Pode-se usar este modelo para identificar quais os campis onde acontecem os maiores índices de evasão escolar, bem outros índices. No entanto, para se ter uma análise mais completa, vou criar modelos relacionados usando algoritmos diferentes e comparar seus resultados. Dessa forma, terei diferentes visões dos resultados.

Para este estudo de caso estarei usando os algoritmos ***Clustering*** e ***Naive Bayes***, respectivamente.

**Algoritmo Microsoft Clustering**

Este algoritmo, fornecido pelo *Analisys Services da Microsoft*, é um algoritmo de segmentação. Segundo a documentação, ele usa técnicas iterativas para agrupar casos em um conjunto de dados em clusters que contenham características semelhantes. Com esses agrupamentos é possível explorar dados, identificando anomalias nos mesmos e criar previsões.

A diferença do algoritmo *clustering* para o algoritmo árvores de decisão é que ele não precisa designar uma coluna previsível para poder criar um modelo de *clusterring*. Ele treina o modelo estritamente a partir das relações existentes nos dados e a partir dos *cluters* que o algoritmo identificar.

**Configurando os dados para agrupar modelos**

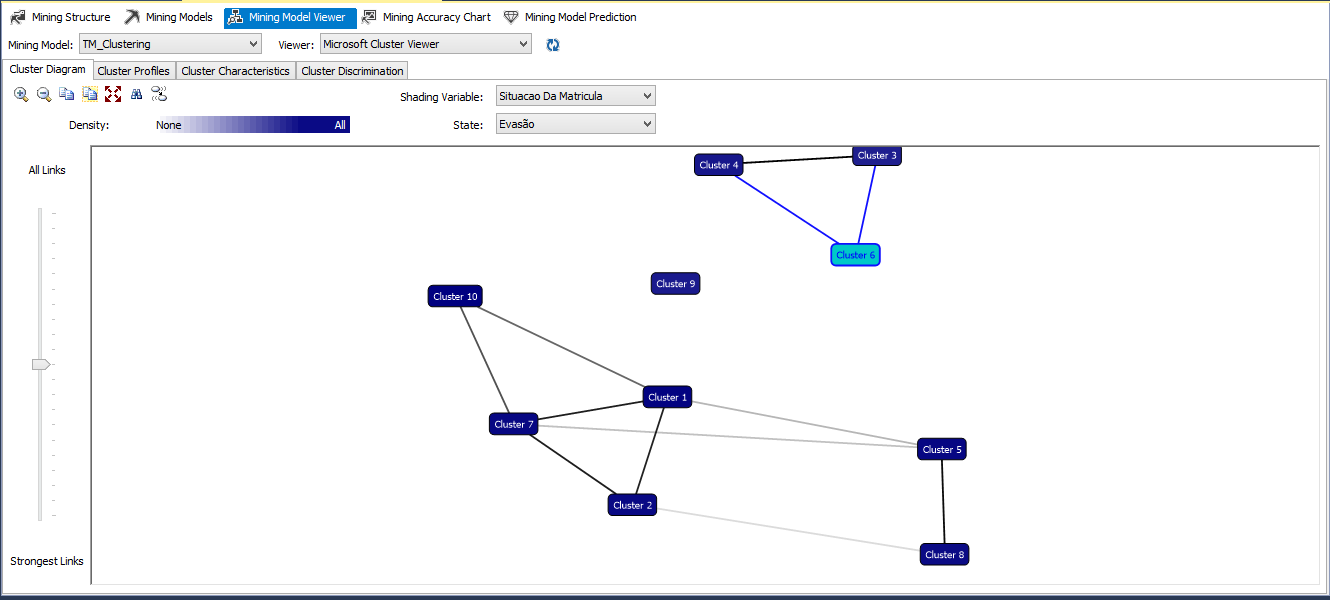
Os requisitos de um modelo clustering são os seguintes:

* Uma única coluna key: Cada modelo deve conter uma coluna chave (key), que tanto pode ser texto ou numérica. Não pode conter chave composta.
* Colunas de entrada: Cada modelo deve conter pelo menos uma coluna de entrada (input) contendo os valores que serão usados para criar os clusters.
* Coluna previsível opcional: O algoritmo não precisa de coluna previsível para criar o modelo. No entanto, você pode adicionar uma coluna previsível de qualquer tipo de dados. A coluna previsível pode ser tratada como entrada para o modelo de *clustering* ou você pode especificar que ela será usada somente para previsão.

Resultados obtidos, com o modelo, clustering aplicado sobre a mesma estrutura usada para o algoritmo árvores de decisão, sobre a tabela CA\_TOTAL\_ALUNOS\_EVADIDOS, do banco de dados **DDS Educacional**, modelo multidimensional do sistema **EDUCASE**.

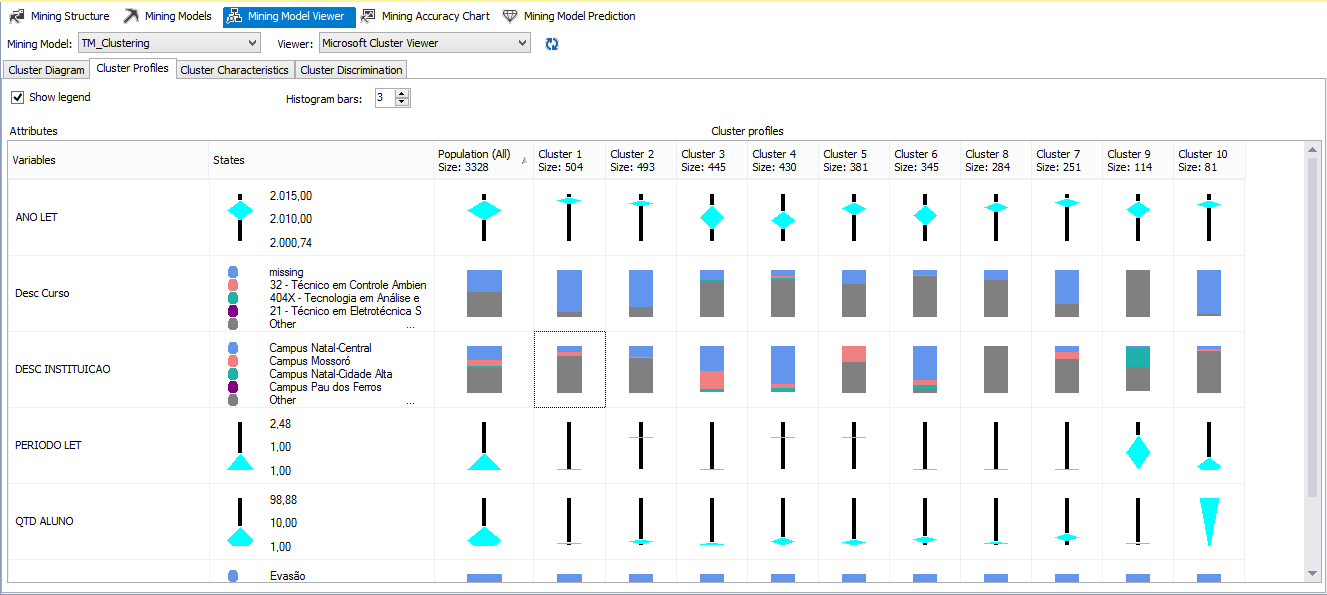
A figura 5 mostra o diagrama de cluster criado pelo algoritmo de clustering. A legenda densidade descreve densidade do par de estados do atributo selecionado. Neste diagrama, o sombreamento mais escuro, indica que tem uma maior percentagem de evasão escolar, naquele nó do diagrama.

Neste diagrama, quando um cluster é selecionado, as linhas que conectam esse cluster a outros clusters são realçadas para que você possa facilmente ver todas as relações desse cluster. Ausência de linhas ligando um cluster a outros clusters, indica que esses clusters não são semelhantes, ou são pouco semelhantes, para perceber isto, basta verificar a intensidade da linha que liga um cluster a outro cluster.



**Figura 5. Diagrama que mostra os clusters agrupados por semelhança**. **Fonte**: autor

Uma outra maneira de analisar os clusters é analisando os perfis dos clusters, para ter uma visão geral do modelo. A figura 6 mostra os perfis dos clusters, onde, a primeira coluna lista os atributos associados a pelo menos um cluster. O resto do visualizador contém a distribuição dos estados de um atributo para cada cluster.

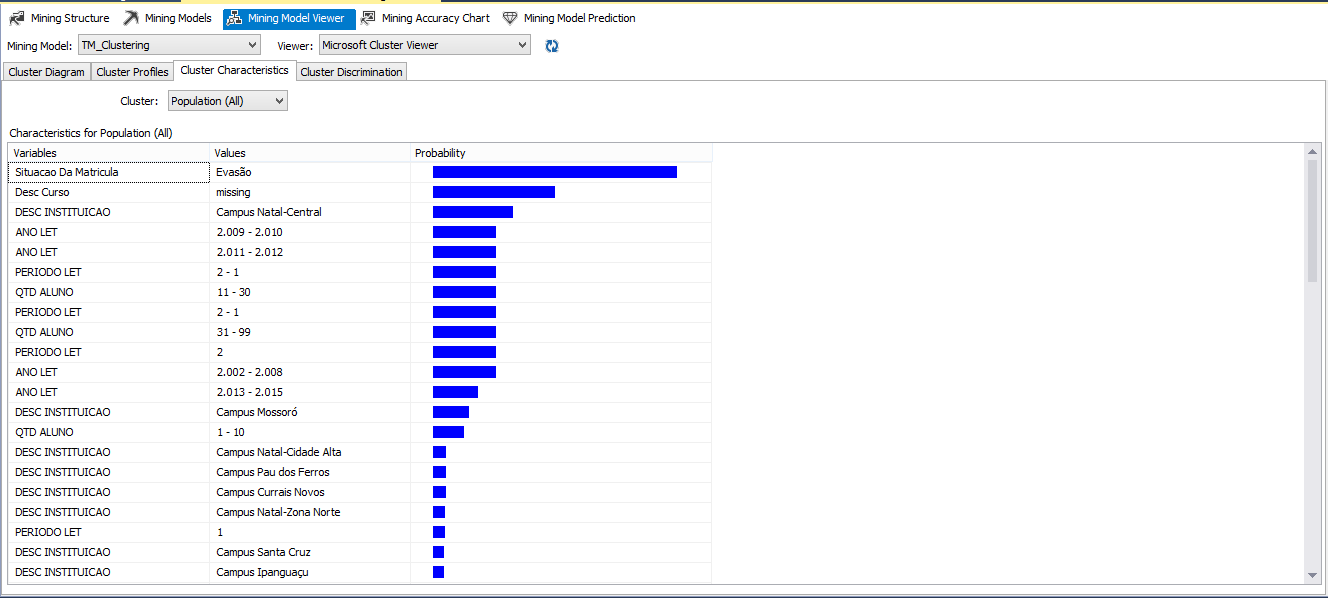


**Figura 6. Diagrama de perfis de clusters**. **Fonte**: autor

Na figura 6, a distribuição de variáveis discretas é mostrada como uma barra colorida com o número máximo de barras exibidas na lista barras de histograma. Os atributos contínuos são exibidos com um gráfico de diamante que representa o desvio médio e padrão em cada cluster.

Analisando a figura 6, percebe-se que no cluster 1 está mostrando que o campus Natal-Central e o Campus Mossoró têm uma certa semelhança. Enquanto, o cluster 2 indica que o Campus Natal\_central é muito diferente do campus Mossoró. As características desses clusters está sendo definida em função dos atributos Ano Letivo, Curso, período letivo e situação da matrícula.

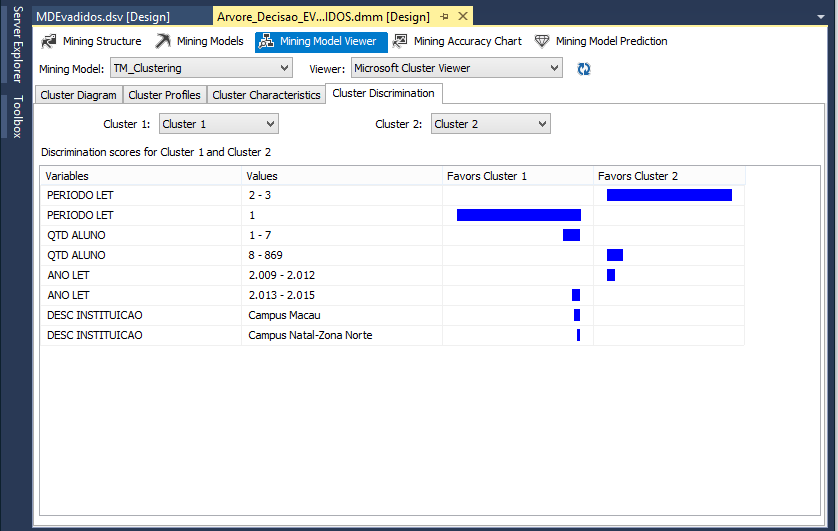
Através do diagrama de características dos clusters, confirmamos o que foi dito no paragráfo anterior, como mostra a figura 7.



**Figura 7**. **Diagrama que exibe as características de cada cluster**. **Fonte**: autor

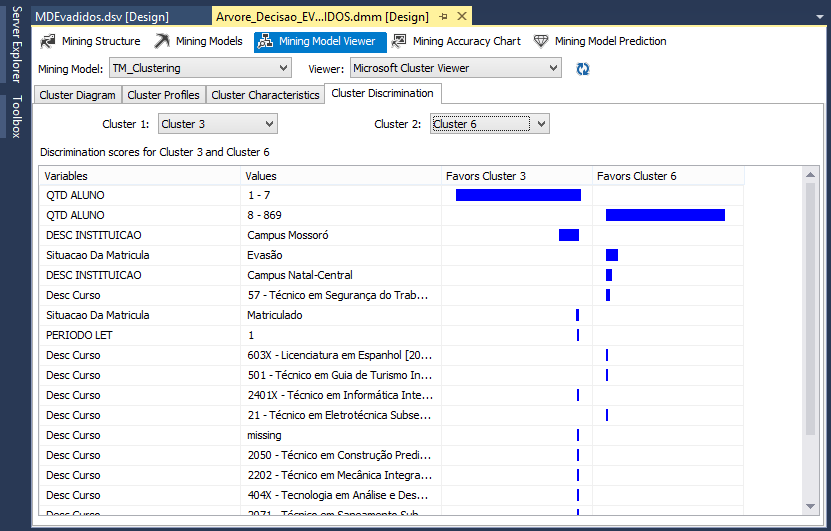
Temos ainda uma maneira de vermos o que distingue um cluster de outro cluster. Através do diagrama de Distinção de Cluster, pode-se explorar as características que distinguem um cluster de outro. Seleciona-se dois clusters e o visualizador calculará as diferenças entre os clusters e exibirá uma lista de atributos que mais distinguem os clusters selecionados. A figura 8 exibe as diferenças entre os clusters 1 e 2 e, a figura 9 exibe as diferenças entre os clusters 3 e 6.

Como mostra a figura 8, os atributos que mais distingue o cluster 2 do cluster 3 é o atributo Período letivo. Enquanto a figura 9 mostra os atributos que mais distinguem os clusters 3 e 6, que no caso é o atributo quantidade de alunos.



**Figura 8. Diagrama que exibe as características que mais distinguem os clusters 2 e 3**.

**Fonte**: autor



**Figura 9**. **Diagrama que exibe as características que mais distinguem os clusters 3 e 6**.

**Fonte**: autor

**Algoritmo Naïve Bayes**

O algoritmo **Naïve Bayes** é um algoritmo de classificação baseado no teorema de Bayes e fornecido pelo Analysis Services para uso em modelagem de previsão.

Esse algoritmo é útil para gerar modelos de mineração de dados rapidamente para descobrir as relações entre as colunas de entrada e as colunas previsíveis.

O algoritmo **Naïve Bayes** calcula a probabilidade do estado de cada coluna de entrada, proporcionando cada possível estado da coluna previsível.

Os requisitos para um modelo **Naïve Bayes** são os seguintes:

* Uma única coluna chave: Cada modelo deve conter uma coluna de texto ou numérica que identifique unicamente cada registro. Não são permitidas chaves compostas.
* Colunas de entrada: Em um modelo Naïve Bayes, todas as colunas são discretas ou diferenciadas.

Quando você usa o modelo para previsão, é importante assegurar que os atributos de entrada sejam independentes uns dos outros.

* Pelos menos uma coluna previsível: O atributo previsível deve conter valores discretos ou diferenciados.