SQL Magazine 134 - Índice

Mineração de dados na prática – Parte 1

Este artigo apresenta um exemplo do uso da mineração de dados para análise do problema de evasão em cursos universitários. Será abordado neste primeiro artigo como podemos preparar a base de dados para a aplicação das técnicas de mineração.

Artigo no estilo Curso

Fique por dentro

A mineração de dados apoia a descoberta de informações úteis que normalmente estão ocultas em bases de dados com grande quantidade de registros.

Neste artigo, apresentaremos um caso prático do uso de técnicas de mineração para análise do problema de evasão em cursos universitários.

Focaremos, na primeira parte deste artigo, na preparação da base de dados para aplicação das técnicas de mineração. Ao fazer isso, iremos comparar também a execução das atividades com e sem o apoio de um modelo de dados preparado especificamente para as atividades de mineração.

Autores: Péricles Magalhães e Rodrigo Spinola

Este artigo apresenta um exemplo prático do uso da mineração de dados. Para isso serão definidos dois cenários. No primeiro deles, a estratégia de mineração será definida considerando como fonte de dados a base original da organização.

Já no segundo cenário, partiremos de uma base de dados definida apenas para apoiar as atividades de mineração. Ambos os cenários serão realizados no contexto de um sistema para análise de dados sobre evasão escolar em universidades comparando o esforço envolvido nas atividades de mineração de ambos.

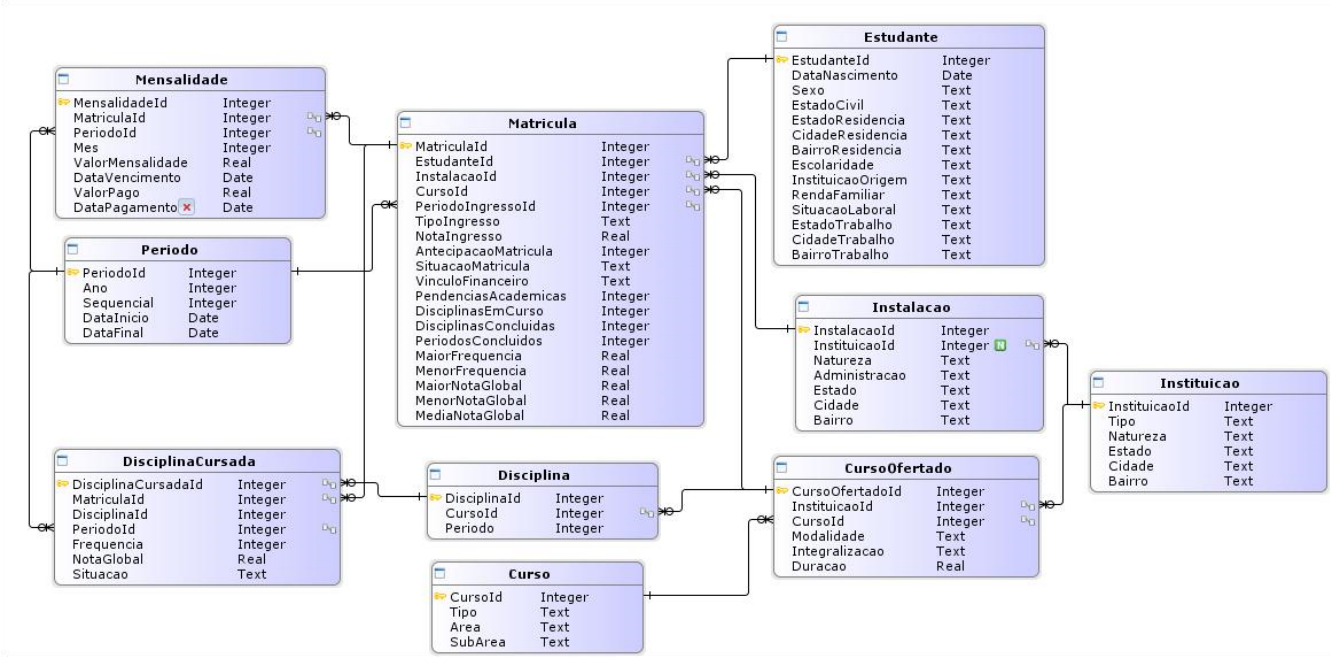
O modelo de dados preparado para mineração que será utilizado foi discutido em detalhes no artigo Modelo de dados para análise de informações educacionais publicado na edição 130 da SQL Magazine.

O estudo de caso descrito neste artigo parte de um cenário real em uma instituição de ensino superior, e procura demonstrar, não apenas a aplicabilidade de um modelo de dados para apoiar a mineração, como também as vantagens em utilizá-lo em detrimento à opção de partir sem uma estrutura específica para estudos de mineração de dados educacionais.

Este artigo é dividido em duas partes. Nesta primeira, faremos a análise até a fase de preparação dos dados.

Estudo de caso

O sucesso de uma aplicação de mineração de dados depende, além da escolha correta do conjunto amostral dos dados a analisar, da correta identificação dos atributos a serem investigados no processo. O modelo de dados apresentado na **Figura 1** reduz o esforço necessário à identificação e seleção de atributos relevantes à identificação de indícios de evasão de estudantes.



abrir imagem em nova janela

Figura 1. Modelo de dados para estudos com mineração de dados educacionais.

Para avaliação do modelo de dados, foram elaborados dois conjuntos de documentos contendo, no primeiro conjunto, o modelo de dados proposto com descrições sobre suas entidades e atributos, enquanto que o segundo conjunto apresentava um diagrama simplificado da base de dados do sistema de processo seletivo da instituição analisada.

Foram convidados seis analistas de sistemas sem conhecimento prévio dos modelos de dados e, aleatoriamente, um dos conjuntos foi entregue para cada participante com o objetivo de que de posse unicamente do material apresentado, pudessem identificar os atributos mais relevantes ao tema da evasão de estudantes.

Todos os analistas de sistemas que participaram da atividade possuem experiência em desenvolvimento de software e análise de diagramas de entidade e relacionamento. Trata-se de um grupo entre 25 e 44 anos de idade com pelo menos um ano de experiência em desenvolvimento de sistemas e modelagem de dados. O participante mais experiente possui mais de 25 anos de atividade profissional.

Cada participante trabalhou exclusivamente com o material disponibilizado, sem que houvesse qualquer limitação ao tempo da experiência. Quatro participantes analisaram o material com o modelo proposto enquanto que três pesquisadores analisaram apenas o modelo original. Um participante analisou ambos os modelos, totalizando sete análises do conjunto.

O tempo médio das análises do material com o modelo da base de dados original foi de 33 minutos, maior que o tempo para a análise do modelo proposto, de 25 minutos. Os tempos máximo e mínimo das análises também foram menores sobre o material relativo ao modelo proposto, conforme assinalado na **Tabela 1**.

	Modelo Proposto	Modelo Original
Tempo mínimo	00:15	00:25
Tempo médio	00:25	00:33
Tempo máximo	00:43	00:48

Tabela 1. Comparativos dos tempos mínimo, médio e máximo das análises dos modelos

Vale ressaltar que não foi identificada nenhuma correlação entre os perfis dos participantes, nos aspectos escolaridade, idade ou tempo de experiência, com os tempos de análises obtidos.

Além disso, todos os participantes alegaram dificuldades na compreensão dos significados das entidades e atributos do modelo original e consequentemente na análise de sua relevância sobre o tema, situação reforçada pelo fato de apenas uma das três análises desse documento ter completado o conjunto de dez atributos solicitados.

Chama a atenção também a baixa reincidência dos atributos desse grupo nos trabalhos. Do total de vinte e quatro atributos identificados pelos analistas como relevantes, apenas dois deles aparecem em mais de uma lista.

Os analistas que trabalharam com o modelo de dados proposto identificaram vinte e seis atributos distintos considerados como relevantes na análise de evasão de estudantes. Desse conjunto, dez atributos aparecem em mais de uma lista, sendo que um deles consta em todas as quatro listas geradas.

A experiência realizada indica que mesmo havendo uma variação do tempo decorrido de acordo com a vivência e dedicação de cada analista, a utilização do modelo proposto na análise do tema sugerido reduziu o esforço para a identificação e seleção dos atributos.

Uma constatação colateral encontrada aponta que a utilização do modelo proposto resultou numa maior homogeneidade dos atributos encontrados o que, por sua vez, indica mais facilidade nas análises. Além disso, todos os participantes que analisaram o modelo original alegaram dificuldades na identificação pela falta de maiores esclarecimentos.

Preparação de dados para a mineração

Uma vez identificado o conjunto de atributos pertinentes à análise do problema a ser investigado com a mineração, é necessário construir um data set contendo os dados de entrada para os algoritmos utilizados.

Esses dados geralmente são provenientes dos bancos de dados de uma ou mais instituições analisadas podendo inclusive ser oriundos de diferentes plataformas tecnológicas (Oracle, SQL Server, PostGreSQL, etc.).

A etapa de preparação de dados para a mineração consiste, justamente, na construção desse conjunto de dados, devidamente tratado, que irá alimentar o algoritmo de mineração utilizado na sua análise.

Para a aplicação de técnicas de mineração, uma amostra de dados reais será definida a partir da qual serão realizadas quatro extrações e tratamentos de dados com a geração, em cada uma, de um data set para mineração.

Duas extrações serão realizadas sem a utilização do modelo de dados proposto,

gerando um arquivo como entrada para um algoritmo de classificação e um arquivo para um algoritmo de clusterização.

As outras duas extrações, realizadas a partir do modelo de dados proposto, deverá gerar arquivos idênticos aos anteriores. O impacto da utilização do modelo proposto será aferido com base nos tempos de cada processo.

Para os casos estudados, foram utilizados dados de uma única instituição de ensino superior com oferta de cursos na modalidade EAD.

A instituição em questão possui um conjunto de aplicações para as diferentes finalidades acadêmico-administrativas do processo educacional – processo seletivo, gestão acadêmica dos estudantes, gestão financeira, além do seu Sistema de Gerenciamento de Aprendizagem, que produz dados sobre a participação do estudante nos seus cursos.

Cada aplicação possui seu sistema de gerenciamento de banco de dados específico, apesar de existirem pontos de integração indireta (por processamento batch) entre os mesmos.

Além disso, não há um padrão geral para nomenclatura de entidades, atributos e relacionamentos, nem documentações institucionais atualizadas para a orientação de desenvolvedores, uma vez que as documentações existentes são pontuais e específicas, construídas pelos analistas de sistemas responsáveis por cada aplicação.

A falta de padronização nos bancos de dados se dá, principalmente, devido à ausência de um administrador de dados dedicado à função e à utilização de aplicações comerciais em conjunto com as aplicações desenvolvidas internamente.

As aplicações institucionais são complementadas com sistemas de menor porte desenvolvidos, muitas vezes, pelas equipes setoriais com a utilização de dados

provenientes de mais de uma base de dados como fonte. A partir de necessidades identificadas pelos setores, são geradas, também, consultas que integram dados de diferentes fontes.

Análise dos dados obtidos

Na realização do estudo de caso, foram analisadas as matrículas de calouros e veteranos de cursos EAD da instituição do primeiro período letivo do ano de 2011 ao último período letivo de 2012 e, sempre que identificado que uma matrícula de um período letivo não constava no período imediatamente seguinte, caracterizava-se uma evasão. Assim, foram identificadas 3.958 situações de evasão divididas nos períodos letivos de acordo com a representação da **Tabela 2**.

Período Letivo	Quantidade
2011.1 – 2011.2	1.127
2011.2 – 2012.1	1.089
2012.1 – 2012.2	1.742

Tabela 2. Distribuição dos evadidos identificados na amostra, por período letivo

Todos os dados de identificação da amostra foram armazenados em um único repositório (Microsoft Access) e iniciou-se, então, a busca pelos demais dados desses estudantes. As **Tabelas 3** a **6** apontam a taxa de sucesso na obtenção dos dados tendo como referência os 3.958 registros (100%).

Atributo	**Taxa de Sucesso**
Tipo de Ingresso	100%

Nota Obtida	100%
Antecipação de Matrícula	68,47%
Opção do Curso	100%

Tabela 3. Taxas de Sucesso na obtenção dos dados de Ingresso dos estudantes

Atributo	Taxa de Sucesso
Sexo	100%
Estado Civil	0%
Idade	100%
Escolaridade	0%
Instituição de Origem	35,45%
Renda Familiar	0%
Situação Laboral	0%
Cidade Residencial	0%
Bairro Residencial	0%
Cidade Comercial	0%
Bairro Comercial	0%
Cidade Polo	100%
Bairro Polo	100%

Tabela 4. Taxas de Sucesso na obtenção dos dados Socioeconômicos dos estudantes

Atributo	Taxa de Sucesso
Vínculo Financeiro	0%
Antecipação Média de Mensalidade.	0%
Pendências Financeiras	0%
Índice de Débito	0%

Tabela 5. Taxas de Sucesso na obtenção dos dados Financeiros dos estudantes

Atributo	Taxa de Sucesso
Tipo do Curso	100%
Modalidade do Curso	100%
Escola do Curso	100%
Pendências Acadêmicas	0%
Disciplinas Cursadas	0%
Períodos Concluídos	100%
Maior Frequência	0%
Menor Frequência	0%
Maior Nota Global	0%
Menor Nota Global	0%
Nota Média Global	0%

Tabela 6. Taxas de Sucesso na obtenção dos dados Acadêmicos dos estudantes

Diante da dificuldade na obtenção dos dados e considerando que a maior concentração de evasão se encontra no primeiro período (semestre) dos cursos, o conjunto de dados foi restrito a estudantes de primeiro semestre, com 1.866 registros que atendem a essa condição (aproximadamente 47% do total de evadidos).

Do ponto de vista dos períodos letivos, conforme a **Figura 2**, os dados da amostra estão distribuídos de maneira uniforme, em consonância com o conjunto total dos registros. Essa constatação aponta uma baixa relação entre o período letivo do estudante e o seu período de ingresso.

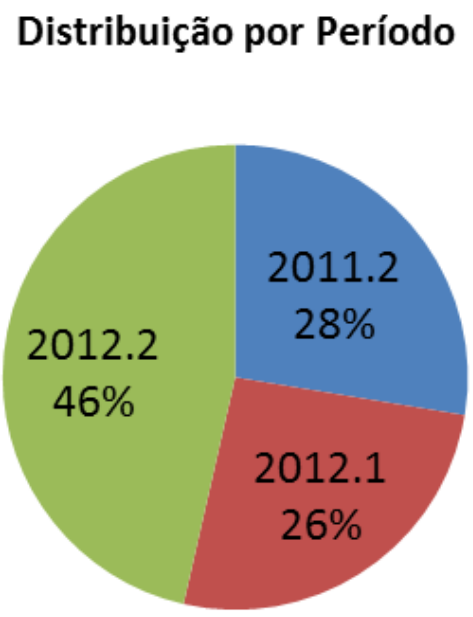


Figura 2. Distribuição da amostra de evadidos por período de ingresso

Analisando os aspectos de sexo e idade dos indivíduos da amostra, conforme demonstrado nas **Figuras 3 e 4**, constata-se também uma baixa relação entre esses atributos e o fenômeno da evasão, uma vez que a maioria de ambos os conjuntos é composta de mulheres (63%) e de pessoas com mais de 25 anos (75%).

Encontra-se uma leve tendência a evadir de estudantes do sexo masculino que, no conjunto de matriculados correspondem a cerca de 29% e, no grupo de evadidos tem sua proporção elevada para 37%.

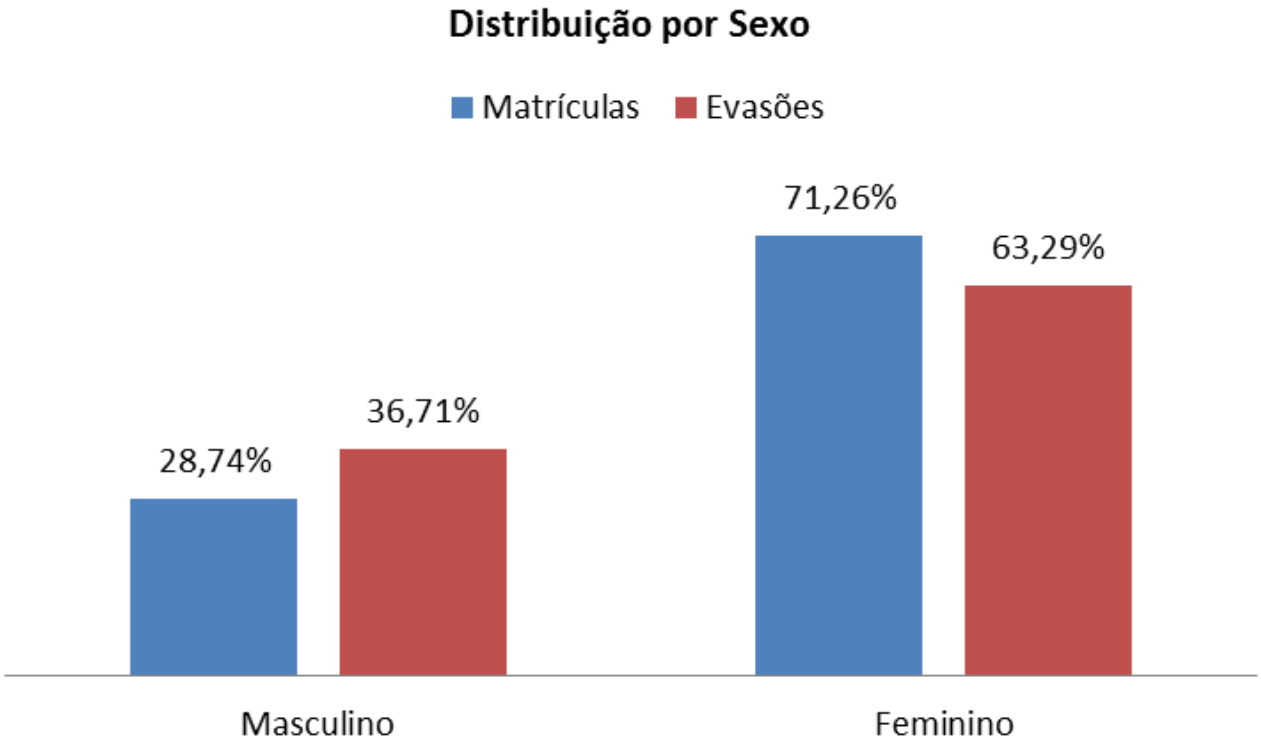


Figura 3. Comparação das distribuições por sexo

A **Figura 4** indica que seguindo o que acontece em relação a sexo, estudantes com mais de 35 anos têm sua proporção aumentada em cerca de 3 pontos percentuais quando comparamos a amostra de evadidos com a população de matriculados.

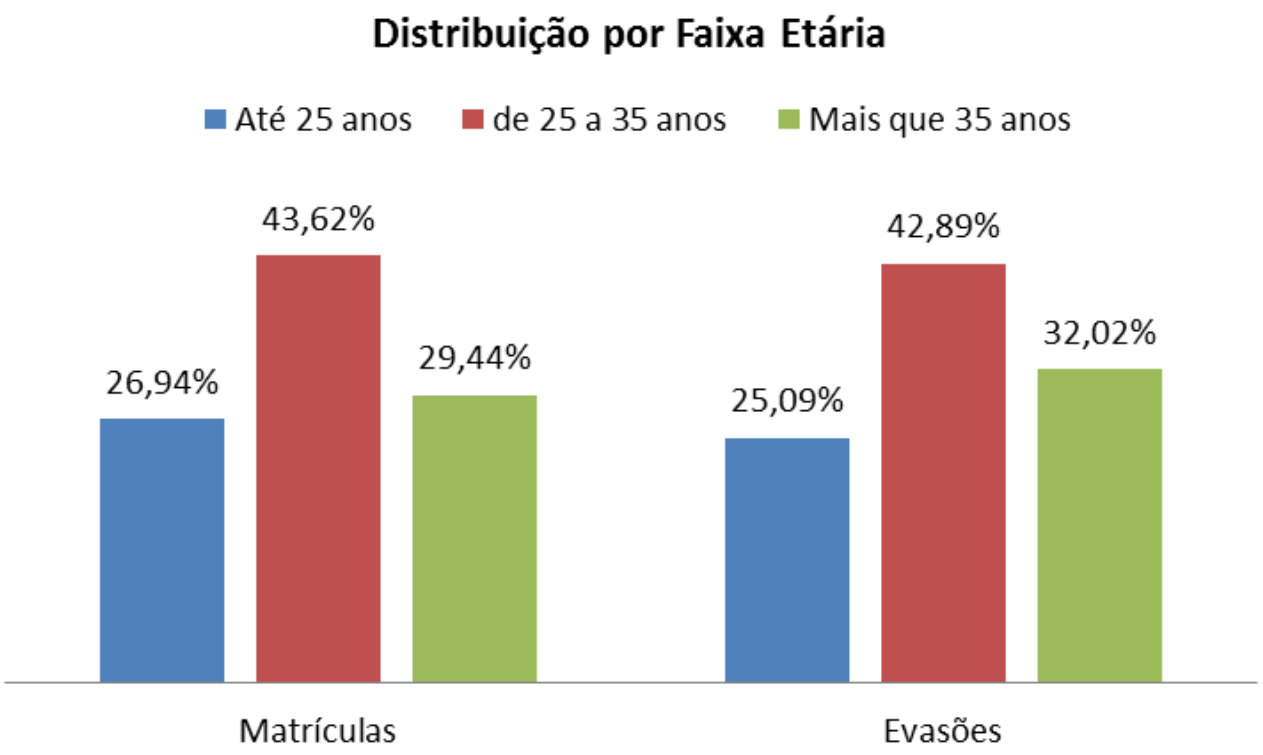


Figura 4. Comparação das distribuições por faixa etária

Quando comparamos os dados por tipo de curso, verificamos uma pequena tendência a evadir em estudantes de cursos de Graduação Tecnológica, como pode ser verificado na **Figura 5**.

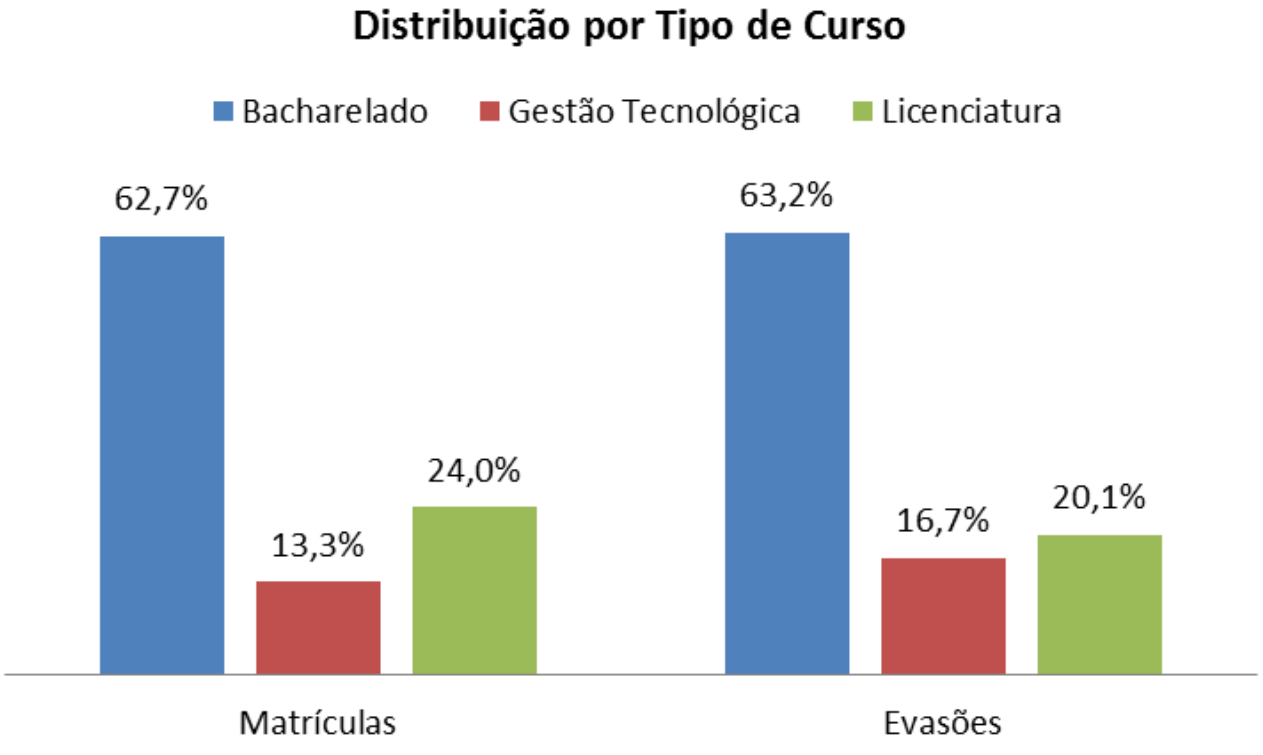


Figura 5. Distribuição da amostra de evadidos por Tipo de Curso

Os dados de antecipação de matrícula, demonstrados na **Figura 6**, já nos revelam uma maior tendência a evadir de estudantes que realizam suas matrículas com mais de uma semana de antecipação (soma das ocorrências de matrícula antecipadas), uma vez que sua proporção na amostra de evadidos aumenta em quase 16 pontos percentuais (de 57,1% para 73,2%) em relação ao montante de matriculados.

Por outro lado, estudantes que realizam suas matrículas no prazo (em até uma semana de antecipação) apresentam uma redução equivalente de participação no grupo de evadidos.

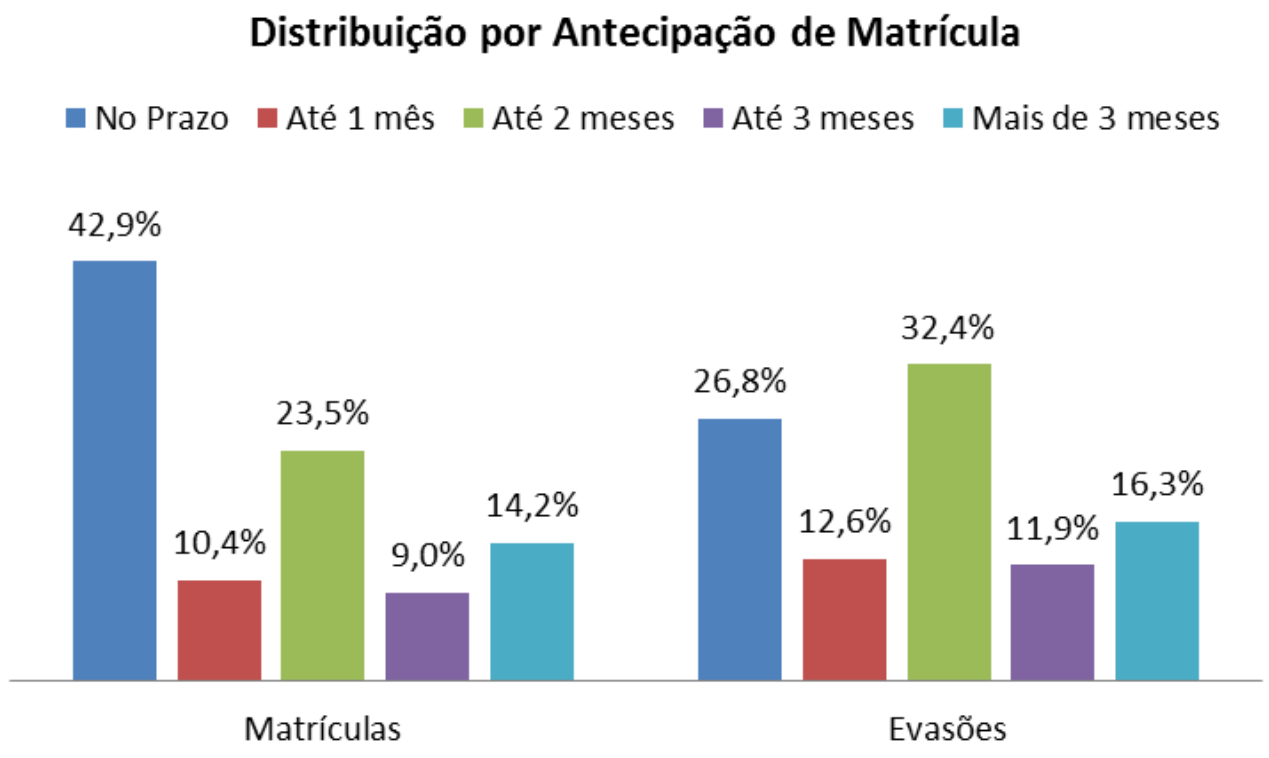


Figura 6. Distribuição da amostra de evadidos por Antecipação de Matrícula

Pelo aspecto da forma de ingresso, demonstrado na **Figura 7**, destaca-se uma considerável tendência a evadir em estudantes provenientes de processos de matrícula especial (portadores de diploma superior) e transferência externa, que têm uma participação de 0,19% na população dos matriculados analisados e, na amostra de

evadidos, passa a contribuir com quase 10% do total.

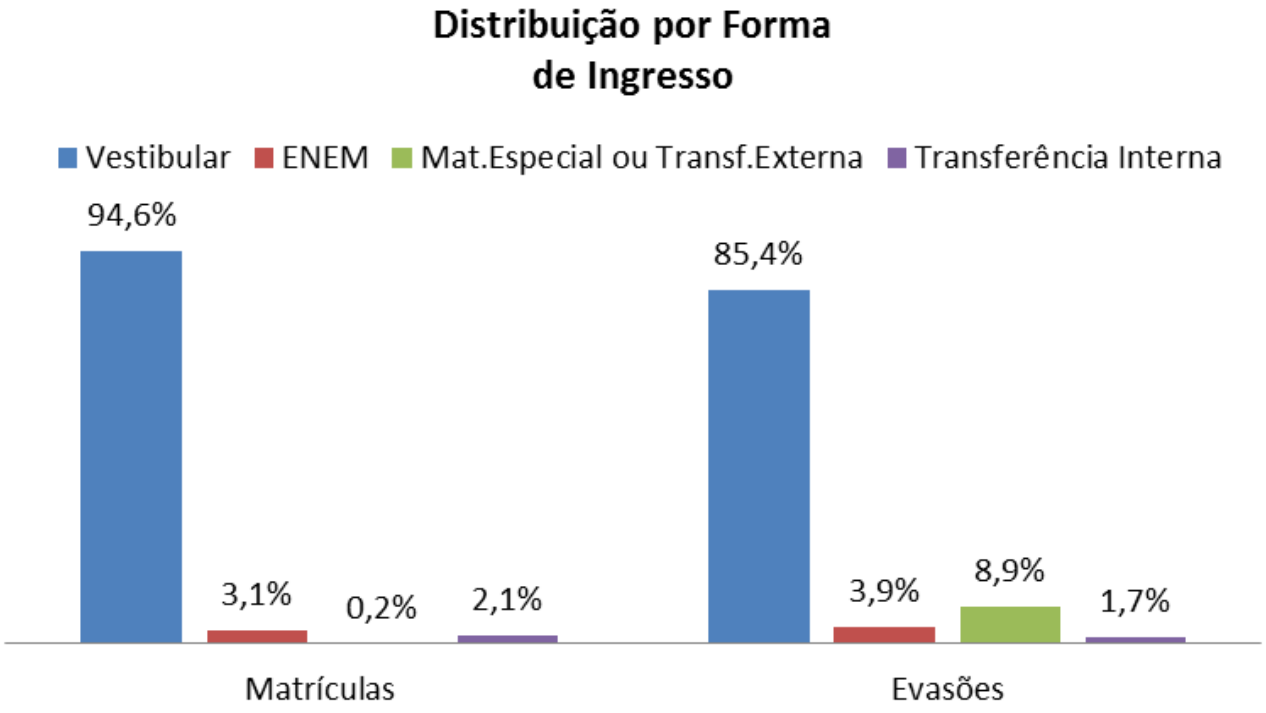


Figura 7. Distribuição da amostra de evadidos por Forma de Ingresso

Apesar do contingente de estudantes que ingressaram por prova (Vestibular ou ENEM) com nota de até 50% do total representar quase três quartos do total de evadidos que ingressaram desta forma, constatamos que ingressar no curso através desta maneira, com um resultado apenas suficiente para a aprovação não constitui indicativo de evasão, uma vez que a população total de matriculados analisada possui praticamente a mesma proporção (**Figura 8**).

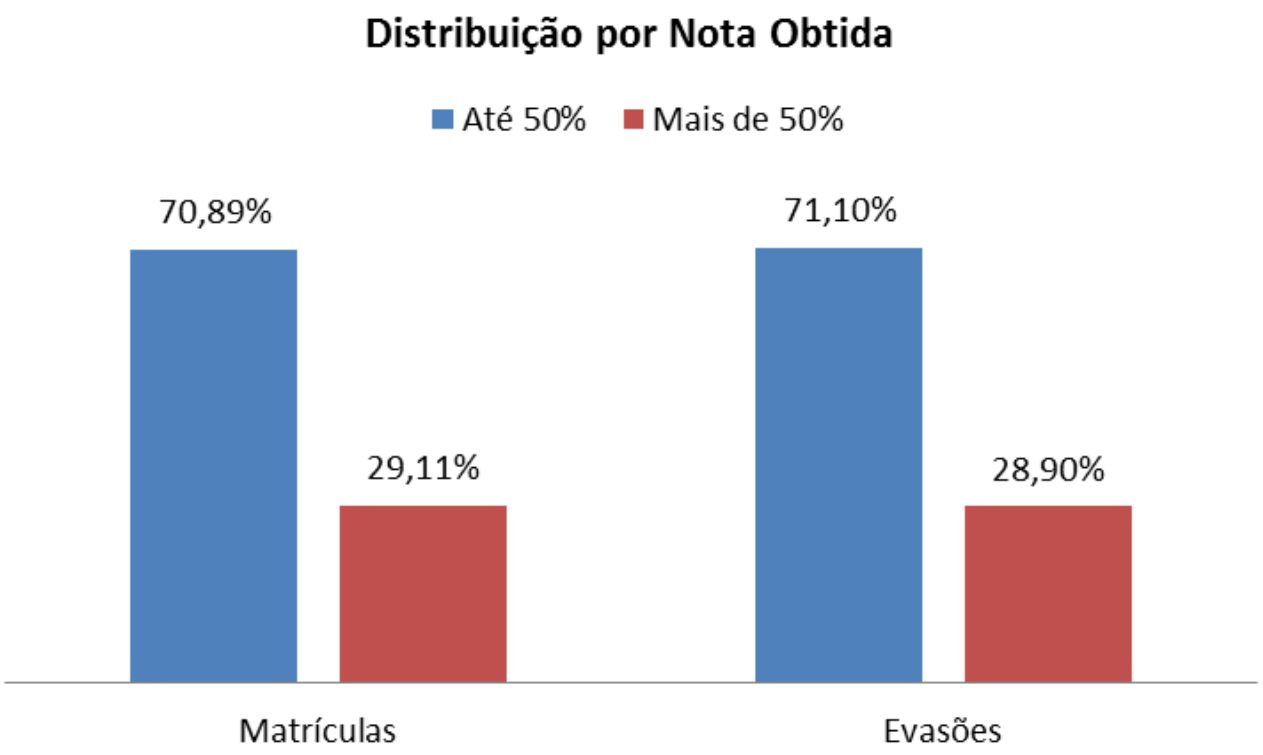


Figura 8. Distribuição da amostra de evadidos que ingressaram através de prova por Nota Obtida

A análise estatística realizada na amostra revela alguns atributos que trazem indícios de evasão tendo como o mais relevante, a antecipação da efetivação da matrícula. Mesmo em atributos onde, pelo estudo realizado, não foram encontradas diferenças significativas entre a amostra e o montante de matriculados, é possível encontrar indícios de evasão.

Analisaremos a partir de agora como este comportamento pode ser percebido através da utilização de técnicas de mineração.

Padrão de entrada para mineração de dados

Para a execução das técnicas de mineração de dados nos casos analisados, foi utilizada a ferramenta Weka (ver seção **Links**). Ela é uma ferramenta gratuita e de código aberto utilizada para minerar dados e transformá-los em conhecimento para o

apoio em tomadas de decisão.

O ambiente disponibiliza recursos para análise de pré-processamento dos dados, para treinamento dos algoritmos e a sua aplicação em si, com uma grande variação de opções de algoritmos para classificação, agrupamento e associação.

O Weka requer como entrada para seus processos arquivos no formato “Atributo-Relação” ou simplesmente arff (Attribute-Relation File Format). São arquivos texto, obedecendo ao padrão ASCII, que descrevem uma lista de elementos que obedecem a um conjunto de atributos pré-estabelecido.

Cada arquivo arff é composto de duas seções distintas: um cabeçalho e uma lista de dados. O cabeçalho deve conter um nome para a relação em questão e a lista dos atributos utilizados, com seus respectivos tipos - numéricos, textuais (strings), datas ou rótulos nominais. A seção de dados apresenta a listagem de instâncias dos registros a serem analisados pelos algoritmos. A **Figura 9** apresenta um esquema geral para um arquivo arff.

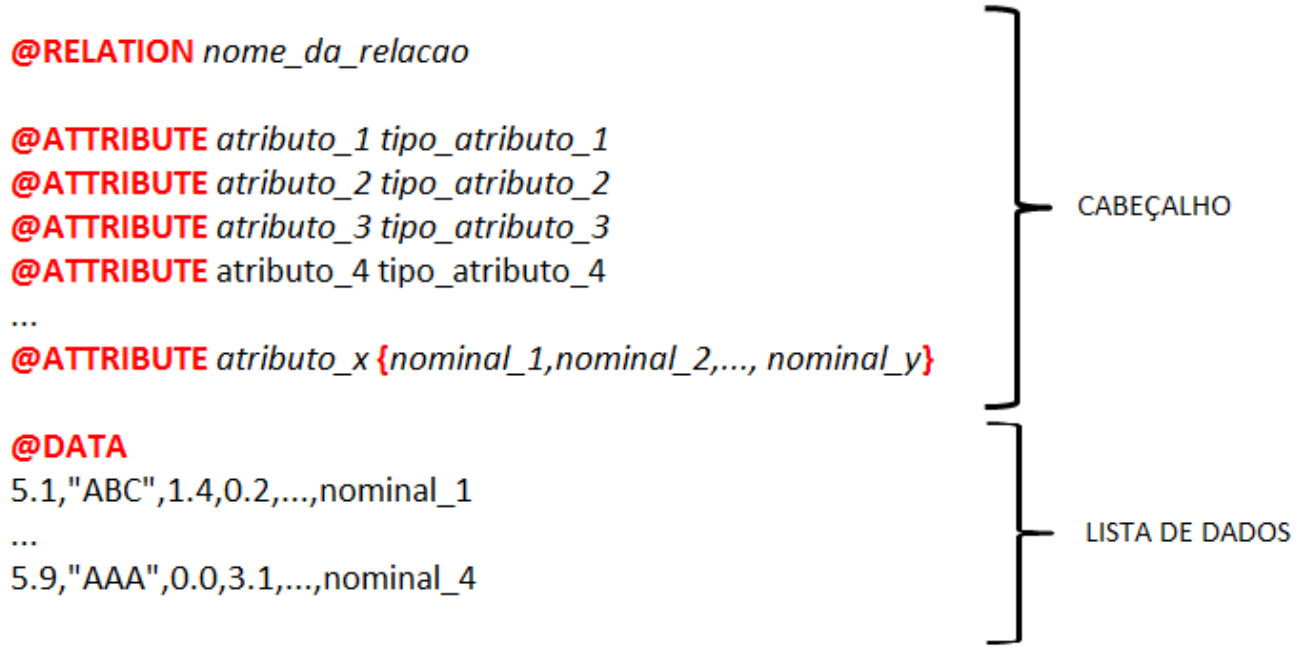


Figura 9. Esquema padrão de um arquivo arff

Em vermelho, encontram-se as palavras de uso reservado @RELATION para identificar a relação, @ATTRIBUTE, para identificar cada atributo e @DATA para apontar o início da seção de dados. Atributos do tipo “valores nominais”, necessariamente, devem apresentar na sua declaração o seu conjunto de valores possíveis relacionados entre chaves.

Tarefas de associação e classificação trabalham, predominantemente, com valores nominais e tarefas de clustering com atributos numéricos. Dessa forma, decidimos trabalhar com arquivos distintos para cada um dos casos analisados.

Em ambos os casos, os dados para os mesmos atributos (Periodo, Curso, TipoIngresso, IngressoNota, IngressoAntecipaçãoMatrícula, Sexo, Idade, Cidade, BairroEstudo, PendenciasAcademicas, DisciplinasCursadas, PeriodosConcluidos e IndicadorEvasão) são utilizados para as tarefas de mineração, variando a forma com que são utilizados e, consequentemente, os tipos dos atributos.

Os atributos para o agrupamento (Caso 1) foram declarados conforme a relação apresentada na **Listagem 1**.

Listagem 1. Atributos para o agrupamento – caso 1

```
@attribute Periodo {20111,20112,20121}
@attribute Curso {LICENCIATURA,BACHARELADO,TECNOLOGICO}
@attribute TipoIngresso {VEST,ENEM,ME,TE,TI,PROUNI}
@attribute IngressoNota real
@attribute IngressoAntecipaçãoMatricula real
@attribute Sexo{1,2}
@attribute Idade real
@attribute Cidade {CAPITAL,INTERIOR}
@attribute BairroEstudo {CENTRO,OUTRO}
@attribute PendenciasAcademicas real
@attribute DisciplinasCursadas real
@attribute PeriodosConcluidos real
@attribute IndicadorEvasao {SIM,NAO}
```

Analogamente, os atributos para a classificação (Caso 2) foram declarados conforme a relação definida na **Listagem 2**.

Listagem 2. Atributos para o agrupamento – caso 2

```
@attribute Periodo {20111,20112,20121}
@attribute Curso {LICENCIATURA,BACHARELADO,TECNOLOGICO}
@attribute TipoIngresso {VEST,ENEM,ME,TE,TI,PROUNI}
@attribute IngressoNota {SemNota,<6,>=6}
@attribute IngressoAntecipaçãoMatricula {SemAntecipacao,AtéUmaSemana,AtéDuasSemana
@attribute Sexo{1,2}
@attribute Idade {Até25anos,de25a35anos,Maisque35anos}
@attribute Cidade {CAPITAL,INTERIOR}
@attribute BairroEstudo {CENTRO,OUTRO}
@attribute PendenciasAcademicas {0}
@attribute DisciplinasCursadas {0,1,2,3,4,5,6,7}
@attribute PeriodosConcluidos {0,1,2,3,4,5,6,7,8}
@attribute IndicadorEvasao {SIM,NAO}
```

Para gerar as listas de valores nominais dos atributos IngressoNota, IngressoAntecipacaoMatricula e Idade, foram determinadas faixas de valores indicando os cortes desejados para os respectivos atributos. Os demais valores nominais foram obtidos através de consultas diretas nas bases de dados relacionando, de forma distinta, os valores existentes.

Extração de Dados sem o modelo proposto

A extração de dados para mineração, quando realizada diretamente nos bancos de dados dos sistemas em produção, demanda um conhecimento prévio de suas estruturas e relações ou, pelo menos, uma boa documentação que facilite a sua análise. Conhecer a estrutura dos bancos de dados em detalhes permite ao analista identificar mais rapidamente a sua origem e definir a melhor estratégia para a obtenção dos dados.

Os dados dos atributos relacionados aos estudos de caso possuem origem em bancos de dados distintos na instituição analisada.

No sistema acadêmico estão os dados cadastrais dos estudantes, assim como dados sobre o seu seguimento acadêmico no curso, enquanto que dados sobre o ingresso do estudante no curso encontram-se no sistema específico para os processos seletivos. A **Tabela 7** divide os atributos selecionados para a mineração de dados por sua origem.

Banco de Dados Acadêmico	Banco de Dados de Processos Seletivos
Periodo	Periodo
Curso	TipoIngresso
Sexo	IngressoNota
Idade	IngressoAntecipacaoMatricula
NomeCidade	
BairroEstudo	
PendenciasAcademicas	
DisciplinasEmCurso	
DisciplinasConcluidas	
PeriodosConcluidos	
IndicadorEvasao	

Tabela 7. Atributos para a mineração de dados por origem nos sistemas legados

A dificuldade em reunir diretamente dados provenientes de bancos de dados distintos demanda, de imediato, a criação de um repositório intermediário para o tratamento e preparação dos dados extraídos. Esse repositório deverá ser formado pela união dos atributos das duas fontes de dados, acrescida de um atributo de identificação única do estudante. A **Tabela 8** apresenta a estrutura necessária para a tabela temporária.

IdEstudante
Periodo
Curso
Sexo
Idade
Cidade
BairroEstudo
PendenciasAcademicas
DisciplinasEmCurso
DisciplinasConcluidas
PeriodosConcluidos
IndicadorEvasao
TipoIngresso
IngressoNota
IngressoAntecipacaoMatricula

Tabela 8. Atributos da tabela temporária para a preparação de dados

Alguns atributos como Sexo, Curso, Cidade, BairroEstudo e TipoIngresso serão armazenados nesse banco intermediário da forma como estão organizados originalmente, sem nenhum tratamento.

A seleção direta dos dados de suas bases originais pode ser realizada com a utilização simples de consultas SQL diretamente nos seus bancos de dados, tendo seus resultados armazenados no repositório intermediário.

A consulta apresentada na **Listagem 3** foi utilizada para extrair os dados referentes aos atributos Periodo, Sexo, Curso, Cidade e BairroEstudo da base acadêmica e inserir na tabela temporária. Na consulta são filtrados os dados de estudantes recém

ingressados (Al.Id_PeriodoLetivo = Al.Id_PeriodoLetivoIngresso) dos períodos letivos desejados (Al.Id_PeriodoLetivo IN (1, 2, 3)).

Os valores alfanuméricos referentes ao atributo Sexo (“F” ou “M”) são transformados em numéricos (1 ou 2) através da expressão IIF(Pes.Sexo ="F",1,2). Os valores dos demais atributos são inseridos na tabela temporária com valor nulo até que as demais consultas os preencham.

Listagem 3. Extração dos dados referentes aos atributos Periodo, Sexo, Curso, Cidade e BairroEstudo

```
  INSERT INTO Temporaria
  (IdEstudante, Periodo, Curso, Sexo, Idade, NomeCidade, BairroEstudo, PendenciasAca
DiscplinasConcluidas, PeriodosConcluidos, TipoIngresso, IngressoNota, IngressoAnteci
  SELECT Al.Nu_Matricula AS IdEstudante, PL.Sigla_PeriodoLetivo AS Periodo, Cur.Nom_
IIF(Pes.Sexo ="F",1,2) As Sexo, NULL, Cid.Nom_Cidade AS NomeCidade, Cam.Bairro_Campu
  FROM Campus Cam INNER JOIN (
    (Cidade Cid INNER JOIN Pessoa Pes
      ON Cid.Id_Cidade = Pes.Id_Cidade) INNER JOIN
     (Periodo_Letivo PL INNER JOIN
      (Curso Cur INNER JOIN Aluno Al
         ON Cur.Id_Curso = Al.Id_Curso)
      ON PL.Id_PeriodoLetivo = Al.Id_PeriodoLetivo)
    ON Pes.Id_Pessoa = Al.ID_Pessoa)
  ON Cam.Id_Campus = Al.Id_Campus
  WHERE (Al.Id_PeriodoLetivo IN (1, 2, 3) AND Al.Id_PeriodoLetivo=Al.Id_PeriodoLetiv
```

Alguns dados da base do sistema de processo seletivo podem ser extraídos também através de consultas simples, conforme a consulta apresentada na **Listagem 4**. Esta seleciona os dados referentes ao atributo TipoIngresso, restritos aos processos seletivos dos períodos letivos desejados (WHERE V.Vest_Id IN (1, 2, 3)).

Listagem 4. Extração dos dados referentes ao atributo TipoIngresso

```
  UPDATE Temporaria T INNER JOIN
```

```
( SELECT C.Cand_Nu_Matricula AS IdEstudante, C.Cand_TipoProva
  AS TipoIngresso, V.Vest_Apelido AS Periodo
  FROM tb_Vestibular AS V INNER JOIN
    tb_Candidato AS C ON V.Vest_Id = C.Vest_Id
  WHERE V.Vest_Id IN (1, 2, 3)
) Aux ON
T.IdEstudante = Aux.IdEstudante AND T.Periodo=Aux.Periodo
SET T.TipoIngresso=Aux.TipoIngresso;
```

Os demais atributos necessários, por outro lado, não existem originalmente na forma desejada, necessitando de um tratamento de seus dados para que atendam ao propósito do estudo de caso:

· Idade – Os dados do atributo Idade são calculados através da diferença, em anos, entre a data vigente e a data de nascimento dos estudantes;

· IndicadorEvasao – Calculado a partir da informação da situação do estudante no curso – estudantes com matrícula “cancelada”, “trancada” ou “em abandono” são considerados evadidos.

A consulta apresentada na **Listagem 5** foi utilizada para extrair os dados referentes aos atributos Idade e IndicadorEvasao da base de dados acadêmica. Além do cálculo da Idade dos estudantes (INT((Now() - P.Dat_Nascimento)/365)), a consulta restringe os valores sobre a situação acadêmica dos alunos a um indicador que informa se ele evadiu ou não (IIF(S.Des_Situacao="ATIVO","NÃO","SIM")).

Listagem 5. Extração dos dados referentes aos atributos Idade e IndicadorEvasao

```
UPDATE Temporaria T INNER JOIN
(
  SELECT A.Nu_Matricula AS IdEstudante, PL.Sigla_PeriodoLetivo
  AS Periodo, INT((Now()-P.Dat_Nascimento)/365) AS Idade,
  IIf(S.Des_Situacao="ATIVO","NÃO","SIM") AS IndicadorEvasao
  FROM
  (    Situacao_Aluno S INNER JOIN
```

```
      (Pessoa P INNER JOIN Aluno A ON P.Id_Pessoa = A.ID_Pessoa
      ) ON S.Id_Situacao = A.Id_SituacaoAluno
  ) INNER JOIN Periodo_Letivo PL ON
     A.Id_PeriodoLetivo = PL.Id_PeriodoLetivo
  WHERE (((A.Id_PeriodoLetivo) IN (1,2,3)
  AND (A.Id_PeriodoLetivo)=[A].[Id_PeriodoLetivoIngresso]))
) AUX
ON T.IdEstudante=Aux.IdEstudante AND T.Periodo=Aux.Periodo
SET T.Idade = Aux.Idade,
T.IndicadorEvasao=Aux.IndicadorEvasao;
```

· IngressoNota – Uma vez que avaliações distintas têm ponderações distintas, os valores armazenados nesse atributo devem conter um percentual representando a pontuação atingida sobre a pontuação máxima possível. Dessa maneira, o desempenho de um estudante em uma avaliação que vale 10 pontos pode ser comparado homogeneamente com o desempenho de outro estudante, de ENEM por exemplo, cuja pontuação máxima é de 1.000 pontos;

· IngressoAntecipacaoMatricula – Com o objetivo de identificar a pontualidade do estudante mediante o prazo final de matrícula, o atributo é calculado pela subtração, em dias, entre a data final determinada para sua matrícula e a data em que a matrícula foi realizada. Dessa maneira, valores próximos a zero indicarão estudantes que tardaram em efetivar suas matrículas, enquanto que valores maiores apontarão para antecipações na formalização.

Os dados referentes aos atributos IngressoNota e IngressoAntecipacaoMatricula são calculados e extraídos através da consulta apresentada na **Listagem 6**. A expressão IIF(C.Cand_TipoProva = "ENEM", C.Cand_Nota/10, C.Cand_Nota / 1000) garante a homogeneidade nos valores das provas de vestibular e ENEM, que possuem ponderações distintas nos seus valores - com a expressão, garantimos que ambos sejam armazenados em valores proporcionais, na faixa de 0% a 100%.

O cálculo da antecipação de matrícula é realizado subtraindo a data de realização da matrícula da data limite prevista para o encerramento das matrículas (M.Marc_DataLimiteMatricula - C.Cand_DataMatricula).

Listagem 6. Extração dos dados referentes aos atributos IngressoNota e IngressoAntecipacaoMatricula

```
UPDATE Temporaria T INNER JOIN
( SELECT C.Cand_Nu_Matricula AS IdEstudante,
  V.Vest_Apelido AS Periodo, IIf(C.Cand_TipoProva="ENEM",
  C.Cand_Nota/10, C.Cand_Nota/1000) AS IngressoNota,
  M.Marc_DataLimiteMatricula-C.Cand_DataMatricula AS
  IngressoAntecipacao FROM
  (  tb_Marcacao M INNER JOIN tb_Candidato C
     ON M.Marc_Id = C.Marc_Id
  ) INNER JOIN tb_Vestibular V ON M.Vest_Id = V.Vest_Id
  WHERE (((C.Vest_Id) IN (1,2,3)))
) AUX ON T.IdEstudante = AUX.IdEstudante
  AND T.Periodo=AUX.Periodo
SET T.IngressoNota = AUX.IngressoNota,
T.IngressoAntecipacaoMatricula = AUX.IngressoAntecipacao;
```

· PendenciasAcademicas – Calculado pela contagem de disciplinas do histórico do estudante que pertencem, na matriz do curso, a um período anterior ao último período cursado e permanecem sem conclusão;

· DisciplinasEmCurso – Calculado pela contagem de disciplinas inconclusas do histórico do estudante que pertencem, na matriz do curso, ao último período cursado;

· DisciplinasConcluidas – Calculado pela contagem de disciplinas do histórico do estudante que pertencem, na matriz do curso, a um período anterior ao último período cursado e já foram concluídas;

· PeriodosConcluidos – Calculado pela contagem de períodos do histórico do

estudante, onde a totalidade de disciplinas esteja com situação concluída.

A geração dos dados calculados para a atualização dos atributos PeriodosConcluidos, DisciplinasEmCurso, DisciplinasConcluidas e PendenciasAcademicas requer a utilização de funções de agregação na sua consulta.

A impossibilidade técnica do sistema gerenciador de banco de dados utilizado em realizar alterações de registros (UPDATE) que possuam funções de agregação demanda a criação de uma segunda tabela temporária para armazenar os resultados calculados e, depois, atualizar a tabela desejada. A **Tabela 9** apresenta a relação de atributos da tabela “temporaria2” criada para armazenar os dados calculados e apoiar na atualização da tabela temporária.

IdEstudante
Periodo
PeriodosConcluidos
DisciplinasEmCurso
DisciplinasConcluidas
PendenciasAcademicas

Tabela 9. Atributos da tabela temporaria2 para a preparação de dados

A consulta apresentada na **Listagem 7** foi utilizada para extrair os valores dos atributos PeriodosConcluidos, DisciplinasEmCurso, DisciplinasConcluidas e PendenciasAcademicas da base de dados acadêmica, referentes aos estudantes recém ingressados nos períodos letivos desejados e inserir na tabela temporaria2. A consulta contabiliza as informações para cada estudante a partir das suas disciplinas cursadas ou em curso.

Listagem 7. Extração dos dados referentes aos atributos PeriodosConcluidos,

DisciplinasEmCurso, DisciplinasConcluidas e PendenciasAcademicas

```
INSERT INTO Temporaria2 (IdEstudante, Periodo, PeriodosConcluidos, DisciplinasEmCu
SELECT A.Nu_Matricula, PL.Sigla_PeriodoLetivo,
A.SemestreAtual-1, SUM(IIf(SM.Des_Situacao="Em Curso",1,0)),
SUM(IIf(SM.Des_Situacao="Concluida",1,0)), SUM(IIf(SM.Des_Situacao<>"Concluida" AN
( Situacao_Matricula SM INNER JOIN
  ( ( Atividade_Curricular AC INNER JOIN Classe C ON
     AC.Id_Atividade_Curricular = C.Id_Atividade_Curricular
     ) INNER JOIN
    ( Aluno A INNER JOIN Matricula M ON
      A.Nu_Matricula = M.Nu_Matricula
    ) ON C.Id_Classe = M.Id_Classe
  ) ON SM.Id_Situacao = M.Id_Situacao_Matricula
) INNER JOIN Periodo_Letivo AS PL ON
A.Id_PeriodoLetivo = PL.Id_PeriodoLetivo
WHERE (((A.Id_PeriodoLetivo) In (1,2,3) And (A.Id_PeriodoLetivo)=[A].[Id_PeriodoLe
GROUP BY A.Nu_Matricula, PL.Sigla_PeriodoLetivo, A.SemestreAtual-1;
```

Os dados referentes à quantidade de períodos concluídos são calculados pela identificação do semestre anterior ao que cada estudante esteja cursando (A.SemestreAtual - 1). Dessa forma, estudantes de 3º semestre têm dois períodos concluídos (o 1º e o 2º semestres), os de 2º semestre têm 1 período concluído e os de 1º semestre não concluíram nenhum período.

Os atributos referentes às quantidades de disciplinas em curso e concluídas têm seus valores calculados através da contagem por estudante das disciplinas com status, respectivamente, “Em Curso” (SUM(IIF(SM.Des_Situacao="Em Curso",1,0))) e “Concluida” (SUM(IIF(SM.Des_Situacao="Concluida",1,0))).

Os valores referentes ao atributo PendenciasAcademicas também são obtidos pela contagem, por estudante, das disciplinas “não concluídas” (IIF(SM.Des_Situacao<>"Concluida")), situadas nas matrizes dos cursos em semestres anteriores ao atualmente cursado pelo estudante

(AC.Semestre_Atividade_Curricular<A.SemestreAtual).

A atualização da tabela temporária a partir da “temporaria2” é realizada através da consulta apresentada na **Listagem 8**.

Listagem 8. Atualização da tabela temporária

```
  UPDATE Temporaria T INNER JOIN (
  SELECT IdEstudante, Periodo, PeriodosConcluidos, DisciplinasEmCurso, DisciplinasCo
  (T.Periodo = AUX.Periodo) AND (T.IdEstudante = AUX.IdEstudante) SET T.PeriodosConc
T.DisciplinasEmCurso = AUX.DisciplinasEmCurso, T.DisciplinasConcluidas = AUX.Discipl
```

A tabela temporária, devidamente preenchida, servirá de base para a geração dos arquivos arff que serão utilizados como entrada de dados para as aplicações de mineração nos estudos de caso.

Para o Caso 1, que analisa a aplicação de um algoritmo de clustering nos dados identificados, utilizamos a consulta apresentada na **Listagem 9** sobre a tabela temporária. As expressões T.Curso="C01", T.Curso="C02" e T.Curso="C03" são simplificações, para fins de apresentação, das listas de cursos dos três grupos – Licenciaturas, Bacharelados e Tecnológicos.

Listagem 9. Consulta para aplicação de um algoritmo de clustering

```
  SELECT T.Periodo, IIF(T.Curso="C01","LICENCIATURA”, IIF(T.Curso= "C02","BACHARELAD
IIF(T.Curso="C03","TECNOLOGICO"))) AS Curso, T.TipoIngresso, T.IngressoNota, T.Ingre
T.Sexo, T.Idade, IIF(T.NomeCidade="SALVADOR","CAPITAL","INTERIOR") AS Cidade,
IIF(T.BairroEstudo="CENTRO","CENTRO","OUTRO") AS BairroEstudo, T.PendenciasAcademica
T.DisciplinasEmCurso, T.PeriodosConcluidos, T.IndicadorEvasao
  FROM Temporaria T;
```

O Caso 2, por outro lado, necessita de valores nominais para a aplicação do algoritmo

de associação. Dessa forma foi utilizada a consulta apresentada na **Listagem 10**. Na consulta, os valores numéricos do atributo IngressoNota são transformados em três valores nominais representando, respectivamente, ingresso sem notas, notas menores do que 60%, e notas maiores ou iguais a 60% (IIF(ISNULL(T.IngressoNota),"SEM NOTA", IIF(T.IngressoNota<0.6,"<60%",">=60%"))).

Listagem 10. Consulta para aplicação de um algoritmo de associação

```
  SELECT T.Periodo, IIF(T.Curso="C01","LICENCIATURA”, IIF(T.Curso= "C02","BACHARELAD
IIF(T.Curso="C03","TECNOLOGICO"))) AS Curso, T.TipoIngresso, IIF(ISNULL(T.IngressoNo
"SEM NOTA", IIF(T.IngressoNota<0.6,"<60%",">=60%")) AS IngressoNota,
  IIf(T.IngressoAntecipacaoMatricula Is Null Or T.IngressoAntecipacaoMatricula=0, "S
      IIf(T.IngressoAntecipacaoMatricula<=7, "AtéUmaSemana", IIf(T.IngressoAntecipac
 IIf(T.IngressoAntecipacaoMatricula<=30,"AtéUmMês","MaisDeUmMês")))) AS IngressoAnte
  IIF(T.Idade<=25,"Até25Anos",IIF(T.Idade<=35,"de25a35Anos","Maisque35anos")) AS Ida
"CAPITAL","INTERIOR") AS Cidade, IIF(T.BairroEstudo="CENTRO","CENTRO","OUTRO") AS Ba
T.PendenciasAcademicas, T.DisciplinasEmCurso, T.PeriodosConcluidos, IIF(T.IndicadorE
  FROM Temporaria AS T;
```

Os valores numéricos referentes ao tempo de antecipação de matrículas dos estudantes são transformados nos valores nominais referentes a cinco faixas – sem antecipação (T.IngressoAntecipacaoMatricula Is Null Or T.IngressoAntecipacaoMatricula=0), até uma semana de antecipação (T.IngressoAntecipacaoMatricula<=7), até duas semanas de antecipação (T.IngressoAntecipacaoMatricula<=15), até um mês (T.IngressoAntecipacaoMatricula<=30) e mais de 1 mês de antecipação (demais valores).

Da mesma forma, os valores referentes a idade dos estudantes são transformados em três faixas de valores – menos de 25 anos (T.Idade<=25), entre 25 e 35 anos (T.Idade<=35) e maiores de 35 anos (demais valores).

A extração dos dados para mineração, tanto para a aplicação de classificação quanto para o algoritmo de agrupamento, quando realizada diretamente a partir dos bancos de dados dos sistemas em produção sem a utilização do modelo proposto, além de demandar um conhecimento prévio das entidades e seus relacionamentos, necessita que seja constituído um data set único, intermediário entre os arquivos arff e as fontes de dados, consolidando as informações das suas diversas origens.

Para a preparação dessa base temporária, conforme apresentado, foram necessárias cinco consultas SQL, sendo duas delas de baixa complexidade (consultas das **Listagens 3 e 4**) e três de média a alta complexidade (consultas das **Listagens 5 a 8**), fazendo uso de recursos de cálculos e agrupamento de registros.

A partir da base intermediária constituída e devidamente preenchida, a extração dos dados para a geração dos arquivos de entrada para os algoritmos de mineração utilizou ainda mais duas consultas simples:

· consulta da **Listagem 9**, para o algoritmo de agrupamento – caso 1;

· consulta da **Listagem 10**, para o algoritmo de classificação – caso 2.

Os dados obtidos dessas consultas foram preparados segundo o formato requerido para o aplicativo de mineração de dados.

Todo o processo, desde as primeiras análises das bases de dados disponíveis até a obtenção dos dois arquivos arff para as aplicações de mineração, descontadas as interrupções, contabilizou aproximadamente duas horas e meia conforme a **Tabela 10**.

Atividade	Tempo Aproximado
Análise dos bancos de dados disponíveis	00:40

Modelagem da tabela temporária	00:10
Listagem 4 – Codificação	00:05
Listagem 4 - Execução e transferência dos dados para a tabela temporária.	00:05
Listagem 5 – Codificação	00:05
Listagem 5 - Execução e transferência dos dados para a tabela temporária.	00:05
Listagem 6 – Codificação	00:10
Listagem 6 - Execução e transferência dos dados para a tabela temporária.	00:05
Listagem 7 – Codificação	00:10
Listagem 7 - Execução e transferência dos dados para a tabela temporária.	00:05
Listagem 8 – Codificação	00:15
Listagem 8 - Execução e transferência dos dados para a tabela temporária.	00:05
Listagem 9 – Codificação	00:05
Listagem 9 - Execução e geração do arff.	00:10
Listagem 10 – Codificação	00:10
Listagem 10 - Execução e geração do arff.	00:10
Tempo Total	**02:35**

Tabela 10. Tempo de extração de dados sem o modelo proposto

Do tempo total executado, somente trinta e cinco minutos foram dedicados à geração dos arquivos arff, enquanto que a preparação da base de dados temporária que foi utilizada como fonte principal para esses arquivos necessitou de duas horas.

Extração de Dados com o modelo proposto

A extração dos dados para os algoritmos de mineração do estudo de caso utilizando o modelo proposto pressupõe a existência de um data set contendo os dados na estrutura do modelo devidamente integrado com as bases de dados da instituição e, por consequência, permanentemente atualizados. Dessa maneira, as consultas sobre esse data set, são suficientes para a geração dos arquivos arff de entrada para a mineração.

A consulta apresentada na **Listagem 11** é utilizada para a extração dos dados para o Caso 1, de aplicação do algoritmo de clustering, a partir do data set constituído considerando o modelo de dados proposto. Na consulta, os dados referentes aos períodos letivos desejados são filtrados através da expressão P.IDPeriodo IN (1, 2, 3) enquanto que a restrição aos recém ingressados é realizada pela expressão M.IdPeriodoMatricula = E.IdPeriodoIngresso. O atributo Idade é calculado pela diferença entre a data vigente e a data de nascimento do estudante (INT((Now() - E.DataNascimento)/365)).

Listagem 11. Extração dos dados para o caso 1 considerando o modelo de dados proposto

```
  SELECT P.Ano & P.Sequencial AS Periodo, C.Tipo AS Curso, M.TipoIngresso,
M.NotaIngresso, M.AntecipacaoMatricula, E.Sexo, INT((Now() - E.DataNascimento)/365)
AS Idade, IIF(I.Cidade="SALVADOR","CAPITAL","INTERIOR")
AS Cidade, IIF(I.Bairro="CENTRO","CENTRO","OUTRO")
AS Bairro, M.PendenciasAcademicas, M.DisciplinasEmCurso,
M.DisciplinasConcluidas, M.PeriodosConcluidos, M.SituacaoMatricula
  FROM Instalacao I INNER JOIN
```

```
    (Curso C INNER JOIN
     (Periodo P INNER JOIN
       (Estudante E INNER JOIN Matricula M
        ON E.IdEstudante = M.IdEstudante)
      ON P.IdPeriodo = M.IdPeriodoMatricula)
    ON C.IdCurso = M.IdCurso)
  ON I.IdInstalacao = M.IdInstalacao
  WHERE P.IDPeriodo IN (1, 2, 3) AND M.IdPeriodoMatricula = E.IdPeriodoIngresso;
```

A consulta da **Listagem 12**, quando aplicada ao data set preparado a partir do modelo proposto, possibilita a criação do arquivo arff de entrada para a aplicação do algoritmo de classificação utilizado no caso 2.

Na consulta foram mantidas as mesmas faixas de valores para os atributos IngressoNota (“SemNota”, “<60%”, >=60%”), IngressoAntecipacaoMatricula (“SemAntecipacao”, “AtéUmaSemana”, “AtéDuasSemanas”, “AtéUmMês”, “MaisdeUmMês”) e Idade (“Até25Anos”, “de25a35Anos”, “Maisque35anos”) para extrair os dados. Da mesma forma, os mesmos filtros foram utilizados garantindo a utilização do mesmo conjunto de dados.

Listagem 12. Extração dos dados para o caso 2 considerando o modelo de dados proposto

```
  SELECT P.Ano & P.Sequencial AS Periodo, C.Tipo AS Curso, M.TipoIngresso,
  IIf(M.NotaIngresso IS Null,"SemNota", IIf(M.NotaIngresso<0.6,"<60%",">=60%")) AS I
  IIf(M.AntecipacaoMatricula Is Null Or M.AntecipacaoMatricula=0, "SemAntecipacao",
  IIf(M.AntecipacaoMatricula<=7, "AtéUmaSemana", IIf(M.AntecipacaoMatricula<=15,"Até
  IIf(M.AntecipacaoMatricula<=30,"AtéUmMês","MaisDeUmMês")))) AS IngressoAntecipacac
  IIF(INT((Now()-E.DataNascimento)/365)<=25,"Até25Anos", IIF(INT((Now()-E.DataNascim
AS Idade, IIF(I.Cidade="SALVADOR","CAPITAL","INTERIOR") AS Cidade, IIF(I.Bairro="CEN
AS Bairro,  M.PendenciasAcademicas, M.DisciplinasEmCurso, M.DisciplinasConcluidas, M
  FROM Instalacao AS I INNER JOIN
    (Curso AS C INNER JOIN
      (Periodo AS P INNER JOIN
        (Estudante AS E INNER JOIN Matricula AS M
        ON [E].IdEstudante=M.IdEstudante)
```

```
     ON P.IdPeriodo = M.IdPeriodoMatricula)
    ON C.IdCurso = M.IdCurso)
ON I.IdInstalacao = M.IdInstalacao
WHERE P.IDPeriodo IN (1, 2, 3)
AND M.IdPeriodoMatricula = E.IdPeriodoIngresso;
```

A geração dos arquivos arff para as aplicações de mineração de dados a partir do data set criado com referência no modelo de dados proposto utilizou apenas as duas consultas SQL apresentadas. O tempo decorrido entre a primeira análise da base de dados com sua documentação e a versão final dos dois arquivos arff para as aplicações de mineração foi de aproximadamente uma hora, conforme demonstrado na **Tabela 11**.

Atividade	Tempo Aproximado
Análise do modelo proposto e sua documentação	00:15
Listagem 11 – Codificação	00:05
Listagem 11 - Execução e geração do arff.	00:10
Listagem 12 – Codificação	00:15
Listagem 12 - Execução e geração do arff.	00:10
Tempo Total	**00:55**

Tabela 11. Demonstração do tempo de extração de dados com o modelo proposto

Vale destacar que apenas quinze minutos foram gastos para análise e planejamento da extração dos dados, correspondente a 27% do tempo total.

Análise dos esforços de extração de dados

A geração dos arquivos de entrada para as aplicações de mineração de dados foi realizada através de dois processos distintos – o primeiro, extraindo os dados diretamente da base de dados real; e o segundo, utilizando um data set construído a partir do modelo de dados da **Figura 1**.

Dessa maneira, quatro arquivos foram gerados ao todo, sendo dois arquivos gerados pelo primeiro processo – um para a aplicação de agrupamento (arquivo A) e um para a aplicação de classificação (arquivo B); e dois arquivos pelo segundo processo, também um arquivo para o agrupamento (arquivo C) e um para a classificação (arquivo D).

Tanto o par de arquivos destinados à aplicação de agrupamento (A e C), quanto os dois arquivos para a aplicação de classificação (B e D) foram produzidos idênticos entre si. Como os processos distintos geraram pares idênticos de arquivos, podemos afirmar que ambos são processos tecnicamente válidos para a preparação dos dados para mineração.

A comparação dos esforços requeridos para a extração de dados pelos dois processos pode ser realizada verificando não apenas o tempo total necessário para as suas execuções, mas também a quantidade de etapas realizadas e sua complexidade.

Do ponto de vista do tempo dedicado, o primeiro processo (sem a utilização do modelo) durou cerca de duas horas e meia para a sua conclusão enquanto que o segundo processo necessitou de apenas cinquenta e cinco minutos - apenas 20% do tempo do primeiro. Apesar de o tempo necessário para a geração dos arquivos ter sido praticamente o mesmo nos dois processos, a preparação dos data sets para a sua geração foi oito vezes menor no processo que utilizou o modelo. A **Tabela 12** demonstra a comparação dos tempos entre os dois processos.

Atividade/Tempo	Sem o modelo proposto	Com o modelo

		proposto
Preparação do data set	02:00	00:15
Geração dos arquivos arff	00:35	00:40
TOTAL	02:35	00:55

Tabela 12. Comparação do tempo de extração de dados nos dois processos realizados

Analisando as etapas realizadas para a extração dos dados e geração dos arquivos para a mineração, identificamos que o primeiro processo necessitou de dezesseis etapas, com a realização de oito operações de banco de dados – duas inserções (INSERT), quatro alterações de dados (UPDATE) e duas consultas (SELECT). A existência de estruturas de bancos de dados distintas foi determinante na complexidade do processo, uma vez que exigiu a criação de duas tabelas temporárias além da manipulação de treze tabelas distintas ao longo das operações de banco de dados.

A extração pelo segundo processo necessitou somente de cinco etapas para a sua realização com a execução de apenas duas operações de consulta (SELECT), uma vez que o data set já se encontrava pré-estruturado para a realização de pesquisas na área de educação, facilitando o planejamento para a preparação dos dados para mineração. A **Tabela 13** apresenta um comparativo dos esforços entre os dois processos.

Atividade/Tempo	Sem o modelo proposto	Com o modelo proposto
Quantidade de Etapas Realizadas	16	5

Quantidade de Consultas Realizadas	8	2
Quantidade de Tabelas distintas envolvidas	15	5

Tabela 13. Comparação dos esforços necessários nos dois processos realizados

A comparação entre os dois processos realizados aponta para vantagens na utilização de um data set atualizado e estruturado segundo o modelo de dados definido na **Figura 1**, principalmente nas etapas de planejamento da preparação de dados para mineração.

A existência de uma estrutura de dados especificamente destinada a análises gerenciais do negócio da educação facilita a identificação das consultas necessárias à extração dos dados e, consequentemente, a geração dos arquivos para a mineração de dados.

A estruturação prévia e sistemática de uma base de dados para análises gerenciais, utilizando o modelo de dados proposto elimina diversas etapas na extração de dados para mineração, reduzindo o tempo necessário para a sua preparação.

Links

Weka

http://www.cs.waikato.ac.nz/ml/weka/

de artigos, dicas, cursos e videoaulas gratuitos e exclusivos para assinantes.

Publicado em 2015

O que você achou deste post?

Gostei (2) (0)

+ Mais conteúdo sobre SQL

Não há comentários

Meus comentarios

Postar dúvida / Comentário

Mais posts

Video aula

Entendendo os tipos de dados para caracteres no MySQL - Curso Completo MySQL - Aula 34

Video aula

Entendendo o funcionamento dos campos de ponto flutuante - Curso Completo MySQL - Aula 33

Video aula

Aprendendo a trabalhar com campos decimais exatos - Curso Completo MySQL - Aula 32

Video aula

Entendendo os limites dos campos inteiros - Curso Completo MySQL - Aula 31

Artigo

Recuperação de bases de dados Oracle

Artigo

Particionamento no Oracle

Listar mais conteúdo