Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de Dados Aplicados aos Dados Públicos do Banco Mundial

Ytalo Allexandre Santos Carvalho
Matheus Souza Santana

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Jan Mendonça Correa

Brasília
2017

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof. Dr. Pedro Antônio Dourado de Rezende

Banca examinadora composta por:

Prof. Dr. Jan Mendonça Correa (Orientador) — CIC/UnB
Prof. Dr. Pedro Antônio Dourado de Rezende — CIC/UnB
Gabriel Heleno Gonçalves da Silva — CIC/UnB

CIP — Catalogação Internacional na Publicação

Carvalho, Ytalo Allexandre Santos.

Mineração de Dados Aplicados aos Dados Públicos do Banco Mundial / Ytalo Allexandre Santos Carvalho, Matheus Souza Santana. Brasília : UnB, 2017.
223 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2017.

1. Mineração de Dados, 2. Dados Públicos, 3. Weka, 4. UnB,
5. Ciência da Computação

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de Dados Aplicados aos Dados Públicos do Banco Mundial

Ytalo Allexandre Santos Carvalho
Matheus Souza Santana

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Jan Mendonça Correa (Orientador)
CIC/UnB

Prof. Dr. Pedro Antônio Dourado de Rezende
CIC/UnB

Gabriel Heleno Gonçalves da Silva
CIC/UnB

Prof. Dr. Pedro Antônio Dourado de Rezende
Coordenador do Bacharelado em Ciência da Computação

Brasília, 01 de Junho de 2017

Dedicatória

Dedico esse trabalho a minha família, meus pais, Maria da Conceição e Francisco Santana, meu irmão Lucas Santana, minha esposa Nayara Gaston e meus avós José Santana, Rita França e Rita Campos, por todo o suporte, paciência, carinho e orações ao longo de toda minha formação.

Matheus Souza Santana

Dedicatória

Dedico esse trabalho às mulheres da minha vida, a melhor mãe do mundo, Josiene Araújo dos Santos Carvalho e minha noiva e melhor amiga, Jéssica Fernanda Albuquerque.

Ytalo Allexandre Santos Carvalho

Agradecimentos

Agradeço primeiramente a Deus, pois sem ele nada seria possível.

Agradeço meus pais, Maria da Conceição e Francisco Santana, que sempre me apoiaram, me deram condições e me incentivaram por toda a minha trajetória.

Agradeço a minha querida esposa, Nayara Gaston, que tive o prazer de conhecer ao longo dessa caminhada e desde então se fez presente em cada momento difícil, me apoiando e dando forças até o último dia.

Agradeço aos meus avós, José Santana, Rita França e Rita Campos, que lutam e rezam por mim desde o nascimento e sempre estiveram presentes em minha vida.

Agradeço ao meu amigo e irmão, Lucas Santana, que sempre esteve comigo e me fez mais forte, pela obrigação de me tornar um irmão melhor, referência e exemplo em sua vida.

Agradeço também aos meus tios e tias que sempre acreditaram em mim.

Agradeço ao professor Jan Correa pela paciência na orientação e incentivo que tornaram possível a conclusão dessa etapa.

Agradeço a esta universidade e seu corpo docente por todo conhecimento que me proporcionaram.

Agradeço aos meus amigos Izael Vilela, Hudson Pereira, Rodrigo Lacerda e Claúdio Alves, esse último que estudou comigo durante minha preparação para ingressar nessa universidade, me incentivou e acreditou em mim, enquanto muitos desacreditavam.

Agradeço a minha ex chefe e amiga Denise Inácio que durante os tempos de estágio sempre me incentivou e me deu sábios conselhos para a faculdade e para a vida.

Agradeço ao meu ex chefe e amigo, Osvaldo Andrade que me deu a oportunidade de conciliar os estudos com o trabalho, me incentivou, deu conselhos, me ensinou e elevou meu conhecimento, além de jamais se negar em me ajudar nos momentos difíceis da faculdade.

Por ultimo, mas não menos importante, agradeço aos grandes amigos que fiz dentro da universidade, em especial Ciro Luís, Michael Rodrigues, Thiago Alves e Ytalo Carvalho. Amigos que por sinal possuem qualidades semelhantes, companheirismo, humildade e a forma alegre que levam a vida, que me ajudaram por todos os anos de faculdade.

Matheus Souza Santana

Agradecimentos

Agradeço primeiramente a Deus, que me deu força durante toda minha vida.

Agradeço aos meus pais, Manoel Virginio de Carvalho Neto e Josiene Araújo dos Santos Carvalho, por todo apoio e cuidado durante toda minha trajetória, sempre incentivando meus estudos.

Agradeço a minha maravilhosa noiva, Jéssica Fernanda Albuquerque, pelo carinho, compreensão, amor e por ser minha companheira incondicional. Obrigado por me fazer sentir forte e confiante em todos os momentos difíceis. Amo você!

Agradeço aos meus avós, Vô Vadú e Vó Belinha, por serem referenciais em várias áreas da minha vida, me cercando com carinho e cuidado desde sempre.

Agradeço as minhas irmãs, Bel e Karol, por serem as melhores irmãs que eu poderia ter.

Agradeço a minha tia Itala, que desde os meus primeiros anos de vida se preocupou com minha educação. Agradeço também a todos os outros tios e tias, por todo apoio e carinho.

Agradeço aos meus amigos de infância, Luis Paulo, Fracis, Luan, Luno, Matheus, Alexandre e Ramone, por mostrarem a importância da confiança e da amizade e por me apoiarem sempre.

Agradeço ao professor Jan Correa pela paciência na orientação e incentivo que tornaram possível a conclusão dessa etapa.

Agradeço a esta universidade e seu corpo docente por todo conhecimento que me proporcionaram.

Agradeço a minha ex chefe e amiga Denise Inácio pelo apoio e conselhos durante os dias de trabalho ao seu lado.

Agradeço ao meu ex chefe e amigo, Osvaldo Andrade por ter acreditado no meu potencial e me dado a oportunidade de conciliar os estudos com o trabalho. Agradeço por me incentivar como pessoa, profissional e por me apoiar espiritualmente.

Por fim, agradeço aos meus grandes amigos que fiz durante minha trajetória acadêmica, Ciro Luís, Luís Seabra, Michael Rodrigues, Matheus Souza, Tarcísio Júnior e Thiago Alves. Por todo aprendizado e aventuras vividas juntos durante o curso.

Ytalo Allexandre Santos Carvalho

Resumo

Este trabalho tem como objetivo encontrar e analisar padrões em um extenso volume de dados públicos disponibilizados por um dos maiores bancos de desenvolvimento do mundo, o *World Bank Group*. Os dados são alguns indicadores sociais e econômicos de diversos países do mundo que são organizados e tratados em um *Data Warehouse* a fim de garantir a consistência dos mesmos, para então, aplicar várias técnicas de mineração de dados, visando encontrar a que possui melhor performance para os dados analisados e que permitem encontrar grupos de países semelhantes, regras de associação dos indicadores que permitem a analise mais profunda dos dados. Os resultados mostram que é possível identificar padrões não triviais em alguns indicadores.

Palavras-chave: Mineração de Dados, Dados Públicos, Weka, UnB, Ciência da Computação

Abstract

This paper aims to find and analyze patterns in an extensive volume of public data made available by one of the world's largest development banks, the World Bank Group. These data are some social and economic indicators of several countries of the world that are organized and treated in a Data Warehouse to ensure the data consistency, and then apply several data mining techniques, in order to find the one that has better performance for the analyzed data and allowing groups of similar countries to be found, association rules of indicators that enable deeper analysis of data. The results show that it is possible to identify non-trivial patterns in some indicators.

Keywords: Data Mining, Public Data, Weka, UnB, computer science

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Atualmente, grandes quantidades de dados são geradas a todo momento, obter informações claras sobre esses dados é de grande importância e sabendo disso, empresas de diversas áreas e pesquisadores têm investido tempo e dinheiro no aprimoramento de técnicas que buscam facilitar o entendimento desses dados.

Uma forma para que esses dados sejam transformados em informações úteis é a mineração de dados. Jiawei Han e Micheline Kamber definem mineração como a extração de conhecimento em grandes quantidades de dados [18].

Com este trabalho pretende-se encontrar, com a ajuda de técnicas de Mineração de Dados, informações úteis em uma grande massa de dados, como especificado no Capítulo 3. Estas técnicas automatizam as análises e auxiliam a descoberta de padrões. Através do uso da Mineração de Dados, é possível encontrar padrões de forma automatizada, além disso, o tempo dispendido na análise dos dados poderá ser reduzido, minimizando também a chance de uma análise equivocada.

O objeto de pesquisa deste trabalho é um extenso volume de dados públicos de fácil acesso. A análise proposta utilizará os dados do *World Bank Group*, tendo em vista que atualmente é um dos maiores bancos de desenvolvimento do mundo [3].

O grupo é constituído por uma família de cinco organizações internacionais que têm como objetivo o fim da pobreza extrema e a construção de propriedade partilhada. As cinco organizações membros são: *Bank for Reconstruction and Development* (IBRD), a *International Development Association* (IDA), a *International Finance Corporation* (IFC), a *Multilateral Investment Guarantee Agency* (MIGA) e a *International Centre for Settlement of Investment Disputes* (ICSID).

1.1 Problema

A grande quantidade de dados disponibilizados pelo *World Bank Group* dificulta a análise e a descoberta de padrões demais. Essa dificuldade gera a necessidade em buscar ferramentas, tecnologias ou metologias que garantam a análise de forma correta e a descoberta de padrões.

1.2 Hipóteses

É possível descobrir padrões úteis nos dados do *World Bank Group*, utilizando corretamente algoritmos de mineração de dados para aperfeiçoar a análise.

1.3 Objetivo

O objetivo do presente trabalho de conclusão de curso é a obtenção de padrões e descoberta de conhecimento a partir da aplicação de técnicas de mineração de dados sobre os dados do *World Bank Group*.

1.3.1 Objetivos Específicos

Para a consecução do objetivo geral supracitado, foram definidos os seguintes objetivos específicos:

- Extrair os dados globais.
- Realizar o tratamento desses dados.
- Aplicar as técnicas e algoritmos de mineração de dados nos dados obtidos.
- Analisar os resultados obtidos após a aplicação das técnicas de mineração.

1.4 Visão geral do Trabalho

Este trabalho está dividido nos seguintes capítulos:

- Capítulo 1: Introdução
- Capítulo 2: Referencial Teórico
- Capítulo 3: Estudo de caso: Dados Globais do *World Bank Group*
- Capítulo 4: Elucidação da conclusão do trabalho e sugestões para os trabalhos futuros.

1.5 Metodologia

Obter os dados do *World Bank Group*, tratar e garantir a qualidade dos dados obtidos, utilizar algoritmos de mineração de dados para a descoberta de padrões existentes.

Capítulo 2

Referencial Teórico

Neste capítulo são apresentados os principais conceitos e ferramentas utilizados no desenvolvimento deste trabalho. Esses conceitos possuem grande importância para o entendimento do projeto e estão subdivididos nas seguintes seções: Seção 2.1 descreve teoricamente sobre Dados, Informação e Conhecimento; A Seção 2.2 apresenta os principais conceitos e propriedades de um *Data Warehouse*; A Seção 2.3 decorre sobre os conceitos de *Extraction, Transformation and Loading* (ETL) abordando como geralmente é utilizado no método de *Business Intelligence* (BI); A Seção 2.4 aborda a ferramenta *Pentaho Data Integration* (PDI) e suas principais formas de utilização; A Seção 2.5 explica os principais conceitos de Banco de Dados, Sistema Gerenciador de Banco de Dados (SGBD) e MySQL; A Seção 2.6 discorre sobre os conceitos de Mineração de Dados, suas formas de aprendizado de máquina, os principais algoritmos e técnicas; A Seção 2.7 apresenta a ferramenta Weka e suas formas de utilização.

2.1 Dados, Informação e Conhecimento

Peter Drucker [8] em 1999, classificou a época como a era da informação, ou sociedade da informação ou do conhecimento, atualmente essa era continua evoluindo a passos largos. Até então, a sociedade havia vivenciado duas revoluções industriais, ambas produzidas por transformações que se iniciaram nas relações de produção e, rapidamente, atingiram diversas esferas sociais, modificando a sociedade como um todo.

A primeira, ao final do século XVIII, conhecida como 1ª Revolução Industrial [8], trouxe novas tecnologias, como a máquina a vapor, a locomotiva, o tear mecânico e a fiandeira, entre outras inovações agrícolas. Essas tecnologias expandiram expressivamente a capacidade produtiva, e, neste contexto, a mão-de-obra deslocou-se do campo para os centros urbanos, causando assim um êxodo rural centrado na busca por novas oportunidades de trabalho. Nas cidades, a criação de novas máquinas associadas à ampla disponibilidade de mão-de-obra barata e matéria-prima, levou a uma nova explosão de produtividade.

Por outro lado, a 2ª Revolução Industrial, instaurada ao final do século XIX, se caracterizou pela difusão de novas tecnologias de comunicação - como o telegrafo - o desenvolvimento da eletricidade, de produtos químicos e a fundição do aço.

O contexto social atual insere-se na Terceira Revolução Industrial, em que a organização da sociedade se fundamenta na tecnologia da informação. Nesta configuração,as revoluções anteriores contribuíram para a possibilidade de armazenar grandes volumes de

dados, de processamento rápido com custos razoáveis de recuperação e, principalmente, transmissão de informação.

Assim como aquelas, a revolução industrial mais recente, alterou o modo como as pessoas vivem e trabalham, gerando uma reorganização social e cultural A terceira revolução criou um novo tipo de economia denominada, por Manuel Castells, de Economia Informacional Global [6], ela é informacional porque a competividade das empresas depende da sua capacidade de gerir informações, e é global porque as atividades produtivas e seus componentes, necessariamente, estão organizados em escala global. Castells afirma, ainda, que o paradigma tecnológico ajuda a organizar a essência da transformação tecnológica atual à medida que ela interage com a economia e com a sociedade [6].

Nesse contexto, é necessário compreender a diferença entre os dados armazenados pelas empresas, as informações obtidas através desses dados e o potencial conhecimento adquirido através de tais informações. Dados são imprescindíveis para a criação de informação, que, por sua vez, fazem parte do processo de construção do conhecimento, permitindo que este seja consolidado [7].

Apesar da distinção evidente entre esses elementos, nota-se que eles se inter-relacionam, construindo uma relação de dependência mútua, cada qual desempenhando um importante e específico papel para as organizações. Desta forma, analisar como se distinguem e de que forma se relacionam é essencial para o sucesso de trabalhos ligados ao conhecimento.

A Figura 2.1 e a Tabela 2.1 apresenta esses conceitos, de forma sintética, e suas respectivas correlações, que são detalhadas nas seções 2.1.1, 2.1.2, 2.1.3:

Figura 2.1: Relações entre os conceitos de dados, informação e conhecimento [7].

DADOS	INFORMAÇÃO	CONHECIMENTO
• Fácil estruturação • Fácil captura em máquinas • Frequentemente quantificado • Fácil transferência	• Requer unidade de análise • Exige consenso em relação ao significado • Exige necessariamente a medição humana	• Difícil estruturação • Difícil captura em máquinas • Frequentemente tácito • Difícil transferência

Tabela 2.1: Relação entre dado, informação e conhecimento [7]..

2.1.1 Dados

Para Rezende [27], *"o dado é entendido como um elemento da informação, um conjunto de letras, números ou dígitos, que, tomado isoladamente, não transmite nenhum conhecimento, ou seja, não contém um significado claro" (2006, p. 62).* De acordo com O'Brien [4], *"dados são fatos ou observações cruas, normalmente sobre fenômenos físicos ou transações de negócios" (2010, p. 12).* É um elemento que, quando tomado isoladamente, não produz qualquer compreensão sobre a realidade.

Abordando uma conceito mais próximo a linguagem matemática, Setzer define dado como uma sequência de símbolos quantificados ou quantificáveis [29]. Um simples texto é um dado ou uma sequência de dados formada por letras, letras pertencentes a um alfabeto, um conjunto finito de símbolos quantificados, sendo assim pode ser construída uma base numérica relacionada a cada símbolo do alfabeto.

Setzer define ainda que, um dado é necessariamente uma entidade matemática e, desta forma, é puramente sintático [29]. Isso consiste que representações formais ou estruturais descrevem integralmente os dados e eles podem ainda, claramente serem registrados e processados por um computador se forem quantificados e quantificáveis.

Interiormente em um computador, fragmentos de um texto podem ser unidos virtualmente a outros fragmentos, por intermédio de adjacência física na memória ou por ponteiros, ou seja, endereços da unidade de armazenamento sendo consumida, construindo assim estruturas de dados. Ponteiros podem fazer a interligação de um fragmento de um texto a uma representação quantificada de uma imagem, de um som, de um vídeo e outras coisas mais. Processar esses dados em um computador limita-se unicamente em realizar manipulações estruturais sobre eles. Essas manipulações são realizadas por programas, que são sempre funções matemáticas, sendo assim, também são considerados dados.

2.1.2 Informação

Segundo Setzer [29], informação é uma abstração informal, isto é, não pode ser formalizada através de uma teoria lógica ou matemática. A informação está presente na mente dos indivíduos e é representada por algo significativo para aquela determinada pessoa. Setzer afirma que isso não é uma definição [29], é uma caracterização, pois "algo", "significativo"e "indivíduos"não estão bem definidos;

> *"Um entendimento intuitivo desses termos. Por exemplo, a frase "Paris é uma cidade fascinante"é um exemplo de informação – desde que seja lida ou ouvida por alguém, desde que "Paris"signifique para essa pessoa a capital da França (supondo-se que o autor da frase queria referir-se a essa cidade) e "fascinante"tenha a qualidade usual e intuitiva associada com essa palavra."*

Assim, não é possível armazenar a informação em um computador, entretanto, uma representação em forma de dados pode ser convertida pela máquina, o que seria uma transformação sintática, que pode ser armazenada. Desta forma, fica claro que um computador não tem capacidade de processar diretamente a informação, novamente é indispensável reduzir a informação em dados. A informação pode ser vista de duas formas:

- Domínio interno de alguém, presente em sua esfera mental, e é gerado a partir de uma compreensão interna exemplificada por uma simples sensação de dor.
- Recebida por ela, a informação tem a capacidade de ser recebida em forma de texto, desenhos, imagens, áudios, etc. Ou seja, por intermédio de um entendimento simbólico formado unicamente por dados.

Desde que compreendida, uma informação pode ser completamente ou parcialmente absorvida com uma simples leitura de texto. Existe a possibilidade de se criar uma relação entre receber determinada informação por meio de dados e receber uma mensagem, entretanto, existem diversas formas de receber informação, como por exemplo, a sensação de frio ao entrar em uma piscina gelada. Veja, que está informação aparentemente não é formada por símbolos, portanto, não pode ser designada como mensagem. Em contrapartida, um latido de um animal, caracterizando um ruído vocal, não possui nenhum dado, mas pode conter inúmeras informações.

Distinguir dado e informação é uma tarefa fundamental em um processo de *Data Mining*, a principal e explícita característica que difere os dois termos é que o dado é especificamente sintático, enquanto uma informação indispensavelmente apresenta semântica. Paralelamente a isso, podemos inferir que um computador não possui capacidade de carregar e processar semântica, pois ele, assim como toda a teoria matemática é inteiramente sintática.

Searle [28] esclarece tais conceitos de forma precisa e simples, considere a Tabela 2.2, ela é composta por três colunas que possuem nomes de cidades, meses apresentados de 1 a 12, respectivamente com os meses de um ano e temperatura média de cada país em determinado mês, obviamente com títulos das colunas e o nome dos países escritos em Chinês, para um indivíduo brasileiro que não possui conhecimento sobre seus ideogramas, toda a tabela é constituída apenas por dados.

家	月	度
巴西	2	18°
德	1	12°
阿根廷	12	23°

Tabela 2.2: Tabela com colunas e registros escritos em Chinês.

Porém, a Tabela 2.3 possui a mesma informação escrita em Português, para este mesmo individuo expressaria diversos tipos de informação.

País	**Mês**	**Temperatura**
Brasil	2	18°
Alemanha	1	12°
Argentina	12	23°

Tabela 2.3: Tabela com colunas e registros escritos em Português.

Constata-se que ainda que a Tabela 2.2 seja ordenada por ordem alfabética ou em ordem decrescente de temperatura, esses processamentos seriam unicamente sintáticos, não trazendo nenhum significado para o indivíduo.

2.1.3 Conhecimento

Setzer caracteriza Conhecimento como uma abstração interior, pessoal, de algo que foi experimentado, vivenciado, por alguém [29]. Nessa perspectiva, o conhecimento pode ser relatado, exposto de alguma forma e o objeto que o descreve é a informação, diferentemente desta, o conhecimento não resulta simplesmente em uma interpretação pessoal, ele necessita de uma vivência, um aprendizado, alguma experiência [10]. O conhecimento está presente em uma esfera meramente abstrata do ser humano, onde este tem plenamente consciência do conhecimento que o pertence, além de conseguir correlacioná-las e criar, a partir dessas relações, novas informações, conclusões, críticas e novos significados [23].

Não existe a possibilidade de carregar o conhecimento em um computador, tendo em vista que este não é subordinado a representações, diferentemente de informações que são inseridos por meio de uma representação em forma de dados, como explicado na Seção 2.1.2. Setzer afirma que é absolutamente equivocado falar-se de uma "base de conhecimento"em um computador. O que se tem, de fato, é uma tradicional "base de dados"[29].

2.2 Banco de Dados

Segundo Korth [22], um banco de dados é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico, ou seja, um banco de dados é compreendido como um agrupamento de dados que se relacionam de alguma forma, mesmo que indiretamente.

Para Ferrari [11], um banco de dados é um local no qual é possível armazenar informações para consulta ou utilização quando necessário.

Assim, a partir da fusão das duas definições apresentadas pode-se conceituar um banco de dados como uma coleção lógica e coerente de dados que possui um significado implícito e cuja interpretação é dada por uma determinada aplicação[9].

2.2.1 Sistema Gerenciador de Banco de Dados - SGBD

Comumente, a aplicação permanece isolada do Banco de Dados por uma camada de aplicação, denominada Sistema Gerenciador de Banco de Dados (SGDB), o principal objetivo do SGDB é gerenciar o acesso, a manipulação e a organização dos dados, disponibilizando uma interface para que seus clientes possam incluir, alterar ou consultar dados previamente armazenados, como ilustrada na Figura 2.2:

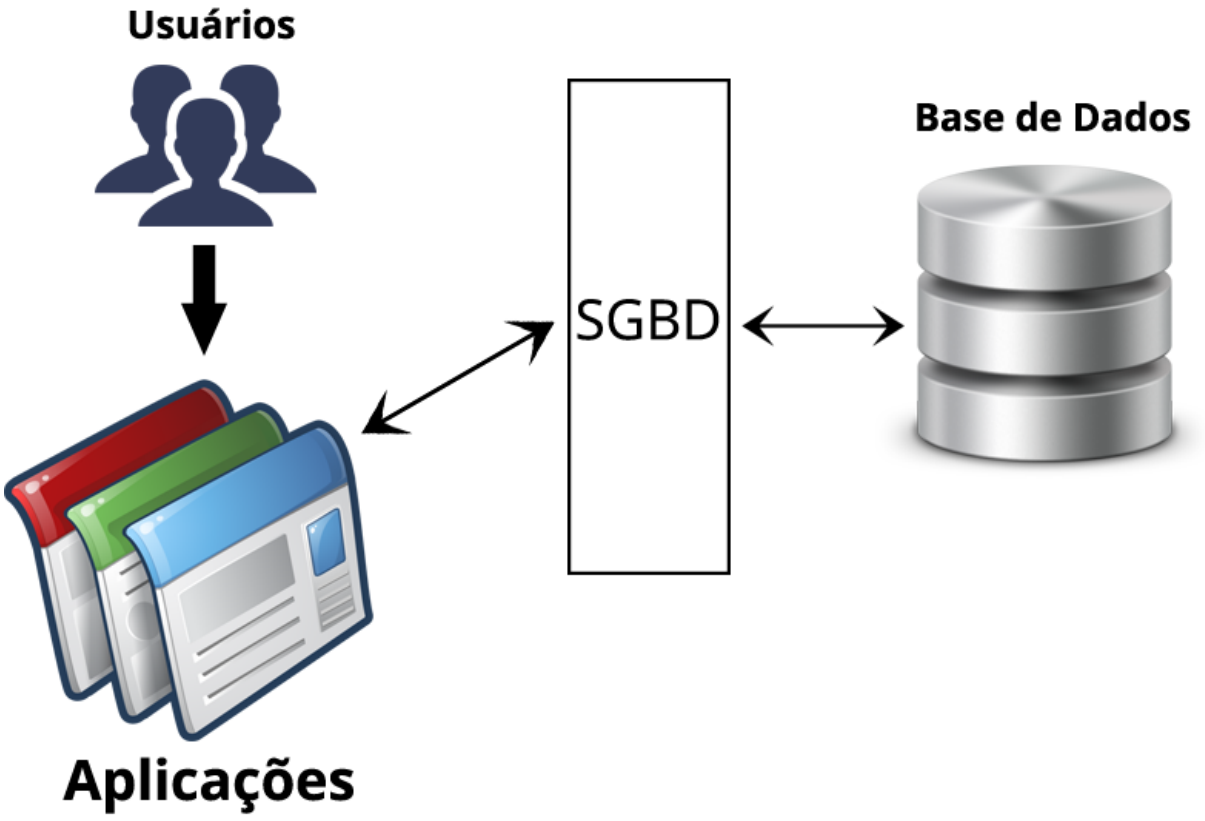


Figura 2.2: Função de um SGBD.

Segundo Ferrari [11], SGBD são bancos de dados que contêm mecanismos automatizados que se encarregam da gestão dos registros, em vista disto, as características auto descritivas de um banco de dados permitem que a aplicação que o acessa não necessite gerenciar a estrutura dos registros, e sim, limite-se a utilizá-los, tendo em vista que o próprio banco de dados se encarregará de criar espaço para novos registros, alterando seu conteúdo de acordo com as solicitações da aplicação que está acessando-o.

Os SGDB's foram amplamente utilizados para a automatização de processos críticos das organizações, como folha de pagamento e contabilidade, oferecendo suporte às funções do negócio organizacional. As transações realizadas nesse tipo de banco de dados são denominadas *Online transaction processing* (OLTP). Para esse tipo de transação é utilizado o conceito SGBD multiusuário, que permite o acesso simultâneo de vários usuários ao banco de dados.

Uma transação é um processo ou programa em execução que realiza vários acessos ao banco de dados, sendo necessário que cada transação seja realizada corretamente, ou seja, uma solicitação realizada pelo usuário deve trazer exatamente o que se pretende, sem interferência de outras transações. Sendo assim, o SGDB gerencia essas transações se tornando capaz de recuperar-se de erros e falhas, com as propriedades citadas abaixo [9]:

- Atomicidade: As transações são consideradas atômicas, desta forma, a transação deve ser realizada por completa, caso contrário, ela será desconsiderada;
- Consistência: A transação preservará a consistência do banco de dados se antes e após a execução total da transação o banco permanecer em um estado consistente. Tal situação indica que o banco satisfaz as restrições especificadas no esquema;
- Isolamento: Uma transação deve ser executada de modo que não interfira no resultado de outra transação, por conseguinte, o resultado de uma determinada transação será o mesmo em dois cenários distintos: quando executada separadamente, ou enquanto outras transações são executadas concomitantemente em determinado espaço de tempo;
- Durabilidade: Quaisquer alterações no banco de dados decorrentes de uma transação efetivada devem permanecer no banco de dados, mesmo em caso de falhas.

2.2.2 MySQL

O software MySQL[1] foi formulado há pouco mais de três décadas, em 1980, na Suécia por David Axmark (Suécia), Allan Larsson (Suécia) e Michael "Monty"Widenius (Finlândia). Recentemente, em 2008, a MySQL AB, desenvolvedora do MySQL foi adquirida pela *Sun Microsystems*, em uma transação que custou US 1 bilhão, um valor extremamente elevado para a categoria a qual pertence, *software open source*[2].

O MySQL é um sistema gerenciador de banco de dados (SGBD), de código aberto e multiplataforma, voltado para a utilização em aplicações de alto desempenho e redimensionáveis. A interface utilizada por esta ferramenta é a linguagem de consulta estruturada SQL.

[1] http://www.oracle.com/us/products/mysql/overview/index.html
[2] Termo utilizado para *softwares* de código aberto.

De acordo com o *DB-Engines Ranking*[3], responsável por medir mensalmente a popularidade dos SGBD's, o MySQL é a segunda solução mais utilizada pelo mercado de sistemas de gerenciamento de banco de dados, superando, inclusive, o Microsoft SQL Server, ficando atrás apenas do Oracle, sistema que pertence a mesma empresa[4]. Ao analisar o topo deste ranking, torna-se fundamental destacar que entre aqueles que ocupam as três primeiras colocações, apenas o MySQL é *open source*.

Na presente pesquisa utilizar-se-á o *MySQL Workbench* como ferramenta para gerenciar o banco de dados MySQL por ser uma ferramenta *open source* e possuir uma configuração simples, por meio dele podem ser criados, visualizados e gerenciados todos os *databases*, *schemas* e tabelas, tanto pela linguagem de *query*, quanto por sua interface amigável, ilustrada na Figura 2.3 abaixo:

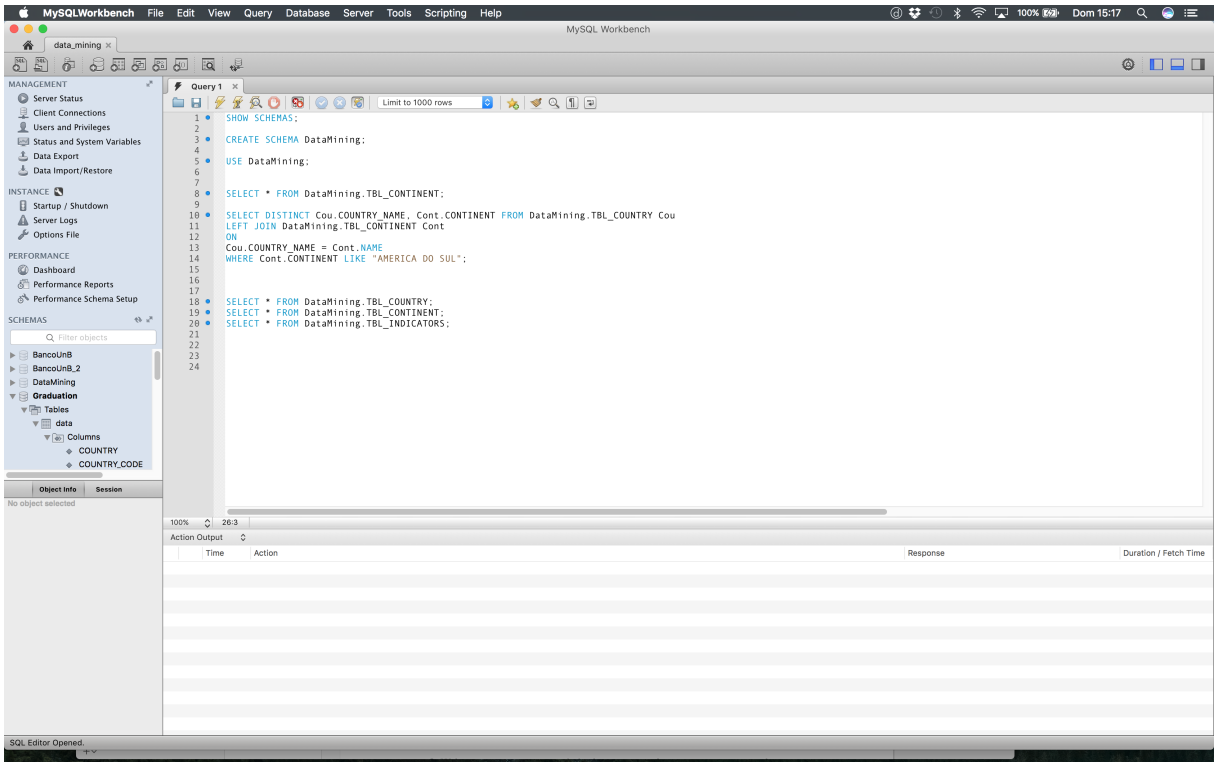


Figura 2.3: Interface do *MySQL Workbench*.

2.3 *Data Warehouse*

Segundo Kimball [21], *Data Warehouse* (DW) é uma fonte de dados para consulta da organização. A distinção entre o Banco de Dados e um DW encontra-se no fato que este último contém um repositório central de dados provenientes de diferentes fontes, armazenados após serem submetidos a tratamento e padronização. O Banco de Dados

[3] http://db-engines.com/en/ranking

[4] Em 2009, a *Oracle* anunciou a compra da *Sun Microsystems* e de todos os seus produtos, incluindo o MySQL.

é utilizado somente para o armazenamento dos dados, a análise dos dados contidos no banco exige a utilização de um SGBD para extrair os dados no formato adequado. O DW supera as fragilidades de um banco de dados, é uma solução voltada ao apoio à tomada de decisão, facilitando a elaboração de relatórios analíticos, visto que não requer um sistema de gerenciamento, conforme detalhado adiante.

A mineração de dados envolve técnicas multidisciplinares que auxiliam no processo, como tecnologias de banco de dados e de *data warehouse* [13],Conforme citado acima, um DW é conceituado por Inmon [16] como depósito integrado de dados orientados por assuntos, não volátil e variável de acordo com o tempo, para suporte ao gerenciamento dos processos de tomada de decisão. Buscando precisar melhor tal definição é importante detalhar alguns dos elementos que a compõem, conforme apresentado abaixo:

- Integrado: A partir de uma variedade de origens, os dados são reunidos no DW e fundidos em um todo coerente.
- Orientado a assunto: Os dados fornecem informações sobre assuntos específicos, possibilitando que se vá além de informações generalistas que abarcam somente informações sobre operações contínuas da companhia.
- Não volátil: Os dados são estáveis no DW. Assim, novos dados podem ser adicionados sem que os anteriores sejam removidos. Esta característica é essencial no gerenciamento, pois proporciona uma visão consistente dos negócios.
- Variável de acordo com o tempo: Todos os dados no DW são identificados em um período de tempo particular.

Percebe-se, então, que o DW não pode ser reduzido a um produto, ele constitui-se como uma estratégia em que os dados são armazenados separadamente em uma etapa posterior ao tratamento, visando, por fim, a sua utilização como uma ferramenta eficaz na tomada de decisão. Os dados contidos em um DW estão consolidados e centralizados, permitindo um fácil acesso às informações [21].

Um bom planejamento de modelagem, baseado nos conceitos descritos por Kimball [26] e representado pela Figura 2.4, é fundamental para o sucesso da mineração de dados.

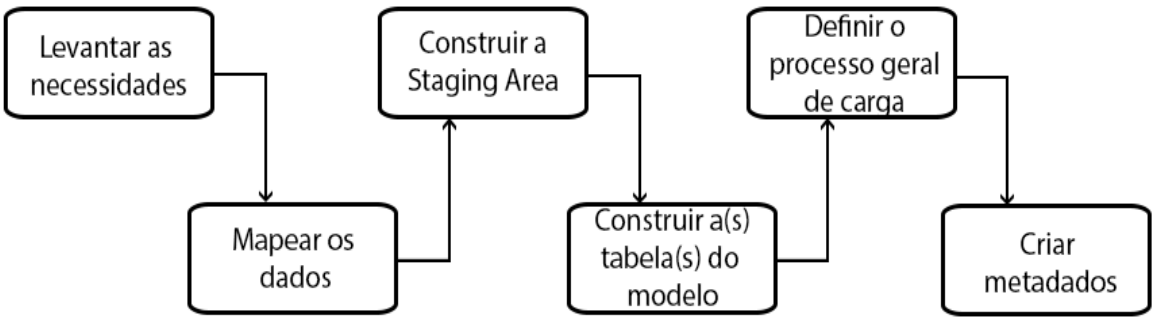


Figura 2.4: Etapas e planejamento de um DW.

O primeiro passo passa por levantar todos as necessidades do sistema; Criar os requisitos e mapear o local de origem dos dados e como chegar até eles; Construir um repositório

único de dados que possui dados sólidos e confiáveis onde possamos extrair informações, ou seja, fazer com que os dados deixem de ser somente dados para representar informações, como explicado na Seção 2.1, evitando também outros acessos ao sistema legado, esse repositório de transição é denominado *Staging Area*; Construir um modelo eficiente para extrair conhecimento; Definir a forma e ferramentas utilizadas para carregar as tabelas do DW; Desenvolver a documentação dos metadados, incluindo o processo de construção e o dicionário de dados (do inglês *data dictionary*, mantém um padrão entre abreviações de nomes e tipos de dados com a finalidade de preservar a consistência entre itens de dados através de diferentes tabelas), fornecendo apoio na gestão do conhecimento.

2.4 *Extraction, Transformation and Loading* - ETL

Atualmente, empresas de portes variados buscam formas inovadoras para se manterem e se destacarem no mercado competitivo. Na busca por métodos que as auxiliem na conquista de tal espaço, temos o *Business Intelligence* (BI), ou inteligência de negócios [17]. O BI refere-se ao processo de obtenção, organização, compartilhamento e monitoramento de informações presentes nos bancos de dados, oferecendo suporte a gestão de negócios.

O BI é um conjunto de ferramentas e técnicas que auxiliam a transformação de grandes quantidades de dados em informações significativas e úteis para analisar o negócio. Essas tecnologias são capazes de suportar uma enorme quantidade de dados desestruturados e auxiliam na identificação, desenvolvimento e, até mesmo, na criação de novas oportunidades estratégicas de negócios. Consequentemente, os principais objetivos do BI são permitir uma fácil interpretação destes dados, identificar novas oportunidades, encontrar estratégias efetivas baseadas nesses dados e promover negócios com vantagens competitivas no mercado, garantindo estabilidade a longo prazo.

Uma das principais etapas do BI é o ETL (*Extract Transform Load* ou Extração, Transformação e Carga, no português), esse processo incide sobre o mapeamento dos atributos dos dados de uma ou várias fontes para os atributos das tabelas do DW, pode ser dividido em três fases essenciais [26], ilustrado na Figura 2.5 seguinte:

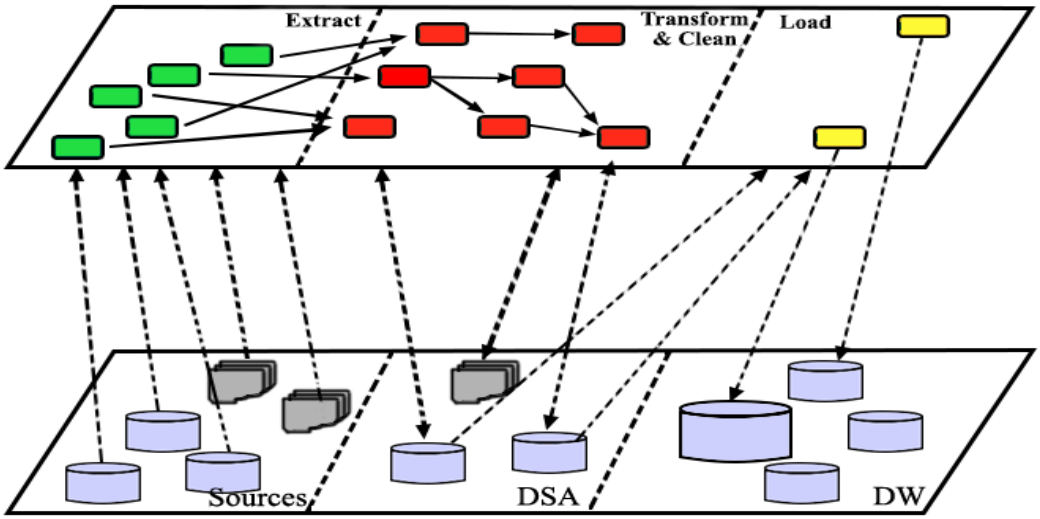


Figura 2.5: Processo de extração, transformação e carga dos dados [1].

A camada inferior da Figura 2.5 representa os dados que subsidiam todo o sistema, inclusive a etapa de extração, que consiste na obtenção de grandes volumes de dados de diversas fontes, desde as mais complexas, como um sistema transacional da empresa, até as mais simples, como planilhas e *flat files*[5], (descritas na figura como *"sources"*).

Além da capacidade de ler e extrair dados de diversos bancos de dados em formatos variados, as ferramentas de ETL são capazes de integrar esses dados, agregando informações provenientes dessas fontes, para posteriormente tratar, formatar e consolidar numa única estrutura de dados.

Os dados que compõem esse processo são obtidos por meio de rotinas de extração, representados na Figura 2.5 sob o termo *"extract"*. Tais dados não sofrem alterações em sua origem, tendo em vista que os ajustes são executados somente nas informações que serão operacionalizadas no DW, adequando-as às necessidades do modelo de DW a ser utilizado.

Após a extração, os dados são propagados para *Data Staging Area* (DSA), área destinada a arquivos intermediários, evitando vários acessos aos sistemas legados e arquivos de origem, e assim, tem papel fundamental em realizar a ligação entre os dados de origem e o DW, percorrendo a etapa de transformação (ilustrada na Figura 2.5 por *"transform and clean"*). Nesta etapa são desenvolvidos ajustes, nos quais adéquam-se os dados às necessidades do modelo de DW, atendendo, assim, às restrições necessárias ao modelo [20]. Este processo visa obter qualidade, limpeza e consistência dos dados, realizando ajustes indispensáveis para a validação dos conteúdos consoante com cada um dos seguintes atributos:

- Devido à codificação, o limite de caracteres entre cada esquema relacional, fonte e destino, não pode resultar em falhas no fluxo de dados, deve ser definido no dicionário de dados um padrão para que um dado proveniente de diversas fontes seja carregado no DW com consistência.
- Os dados devem ser transformados corretamente, seguindo fielmente as regras de negócio especificadas.
- A integridade referencial entre as tabelas precisa ser garantida.
- A rotina ETL deve rejeitar ou substituir os valores defeituosos, reportando todos os dados inválidos.
- Os valores necessitam de validação e, quando incorretos, devem ser corrigidos.
- As conversões de dados devem ser realizadas corretamente, garantindo que os valores não percam informações ou sentido em nenhuma circunstância.
- Caso esteja especificado nas regras de negócio, deve-se resolver a duplicidade dos dados.
- No caso de atributos nulos ou ausentes, deve-se inserir valores padronizados conforme as regras de negócio.
- A filtragem dos dados deve ser realizada, corrigindo erros de digitação e padronizando todos os tipos de atributos a serem carregados no DW.

[5] Arquivos textos.

Por fim, a carga dos dados padronizados, consistentes e limpos é realizada por rotinas de carga no DW (representadas na Figura 2.5 pelo termo *load*), respeitando as restrições de integridade e criando uma visão concreta e unificada das fontes. Devido a dependência da heterogeneidade dos bancos de dados, este processo torna-se extremamente complexo, apresentando obstáculos que dificultam a obtenção de êxito [19].

2.5 *Pentaho Data Integration* - PDI

O Pentaho[6] é uma solução de código aberto que possui funcionalidades para desenvolvimento de mineração de dados, criação de *workflow*, OLAP e capacidade de ETL. Recebeu por cinco anos consecutivos (entre 2008 e 2012) o título de melhor ferramenta de código aberto para BI pela InfoWorld[7]. Desenvolvido em Java, é estruturado por diversos componentes que permitem realizar extração de fontes variadas, transformações e cargas em diversos bancos e arquivos; mineração, análises de *clusters*, processamento de grandes bases de dados; geração de metadados e relatórios, como demonstrado na Figura 2.6:

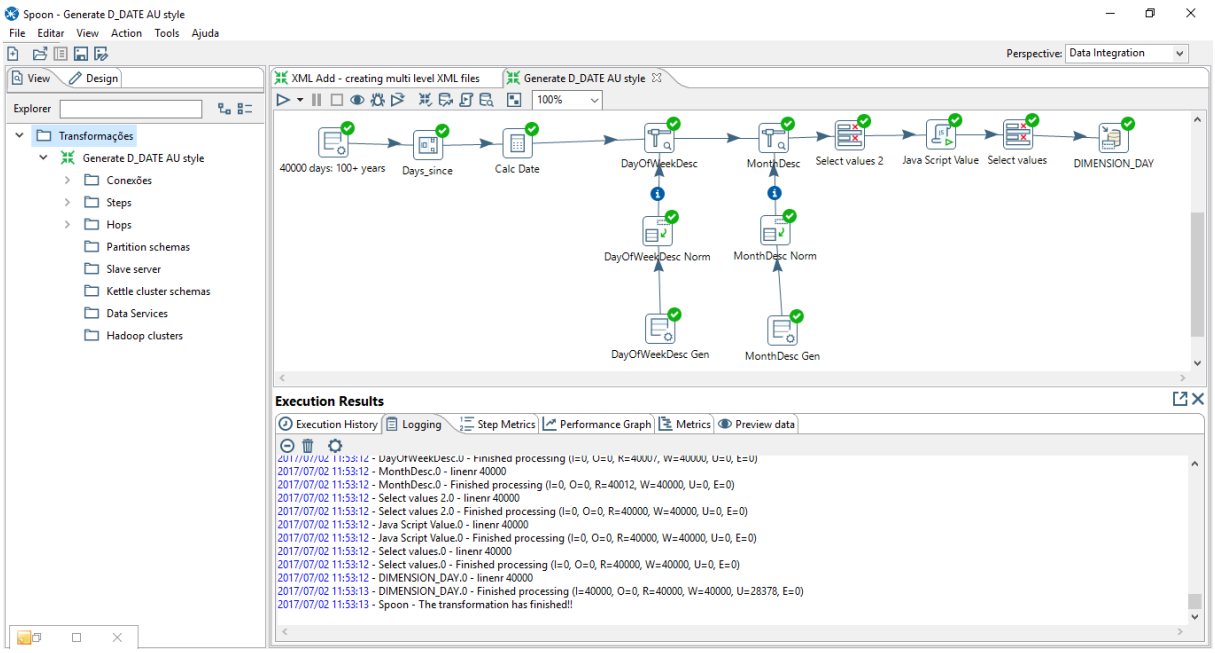


Figura 2.6: Interface *Pentaho Data Integration*

Pela sua vasta quantidade de componentes, subdivididos em pastas (especificadas na paleta presente na parte esquerda da Figura 2.6) e seu *desinger* gráfico, o *Pentaho* se torna uma ferramenta fácil e intuitiva, os componentes são arrastados para o ambiente de desenvolvimento (parte superior direita da Figura 2.6) criando um fluxo e um *pipeline* de dados. Ainda pode-se controlar o fluxo de dados, visualizando em tempo real durante

[6] http://www.pentaho.com

[7] InfoWorld é uma empresa de mídia online e uma organização de eventos e negócios com foco na tecnologia da informação, integrante do *InfoWorld Media Group*, uma divisão da IDG (*International Data Group*). Disponível em http://www.infoworld.com/, acessado em 25/11/2016

o processo de preparação dos dados (parte inferior direita da Figura 2.6), facilitando os testes e a correção de eventuais erros.

2.6 Mineração de Dados e o Processo de Extração do Conhecimento

No mundo atual constantemente surgem novas tecnologias, produzindo cada vez mais um imenso volume de dados em tempo real [24]. A cada venda, cada mensagem enviada ou recebida, transações bancárias, informações sobre cada habitante, buscas em sites de compras, convites em redes sociais, os dados estão presentes por toda a parte e grande parte deles guardados em sistemas digitais. Encontrar uma maneira de analisar esses dados é uma busca constante no meio tecnológico, encontrar formas de combinar e integrar diferente fontes de dados para explorar ao mesmo tempo e descobrir padrões entre eles, resultando em uma vasta quantidade de informação produzida e potencialmente um conhecimento que busque formas de melhorar uma empresa, um país ou até mesmo o mundo. Entretanto, esse enorme volume de dados torna a analise humana árdua e absolutamente trabalhosa, haja vista que a velocidade de produção de dados é muito maior que a velocidade de produção de conhecimento sobre eles [15].

O ser humano é capaz de levantar hipóteses, fazer deduções, descobrir padrões e compreender propriedades em conjuntos de dados menores e com quantidade reduzida de atributos, todavia, na medida em que esse conjunto aumenta juntamente com a quantidade de atributos presente no conjunto, a compreensão das propriedades e a descoberta de padrões transfigura-se em uma tarefa complicada e cansativa, padrões complexos entrelaçados entre vários atributos são dificilmente identificáveis e levariam um tempo elevado para serem encontrados.

Para isso, se faz necessário utilizar técnicas e ferramentas computacionais que facilitem a análise dos dados. Com o emprego da estatística e a possibilidade de visualizar os dados através de tabelas e gráficos, se torna viável alcançar algum conhecimento relacionado a uma quantidade relativamente grande de dados, no entanto, para uma análise mais criteriosa e profunda é absolutamente necessário algoritmos e métodos automatizados, ou seja, técnicas de mineração de dados.

Com o constante avanço da tecnologia, ferramentas gerenciais foram desenvolvidas para facilitar o processo de análise sobre grandes bases de dados, realizado de forma automatizada e confiável. Em vista que a mesma análise, quando realizada de forma manual, torna-se impraticável e susceptiva a erros, implicando em maiores custos de tempo, processamento e mão de obra.

Assim, a técnica de minerar dados (*Data Mining*) surge como uma metodologia de pesquisa e avaliação, de acordo com Han e Kamber [13], é um conjunto de técnicas multidisciplinares que engloba tecnologias de banco de dados e de *Data Warehouse*, computação de alta performance, *Machine Learning*, reconhecimento de padrões, redes neurais, estatística, recuperação de informações, visualização de dados, processamento de imagens e sinais e analisadores espaciais e temporais de dados.

É imprescindível que a mineração de dados possua como base técnicas eficientes e escaláveis. Um algoritmo é escalável quando o tempo de execução aumenta de forma linear proporcionalmente ao tamanho dos dados de acordo com os recursos disponíveis,

como memória e espaço em disco. Com algoritmos consideravelmente eficientes é possível obter conhecimento sólido que pode ser usado em diferentes situações, como: tomada de decisão, controle de processo, gerenciamento de informações e processamento de consultas.

Por conta disso, a mineração de dados é considerada uma das áreas de desenvolvimento interdisciplinar mais promissoras e importantes da tecnologia da informação e em sistemas de banco de dados. Muitas vezes, devido a grande complexidade da técnica de mineração de dados, o conceito costuma ser sinônimo de *Knowledge-Discovery in Databases* (KDD ou extração de conhecimento).

De acordo com Fayyad et al. [30], o processo de extração de conhecimento é composto por cinco etapas, representadas na Figura 2.7.

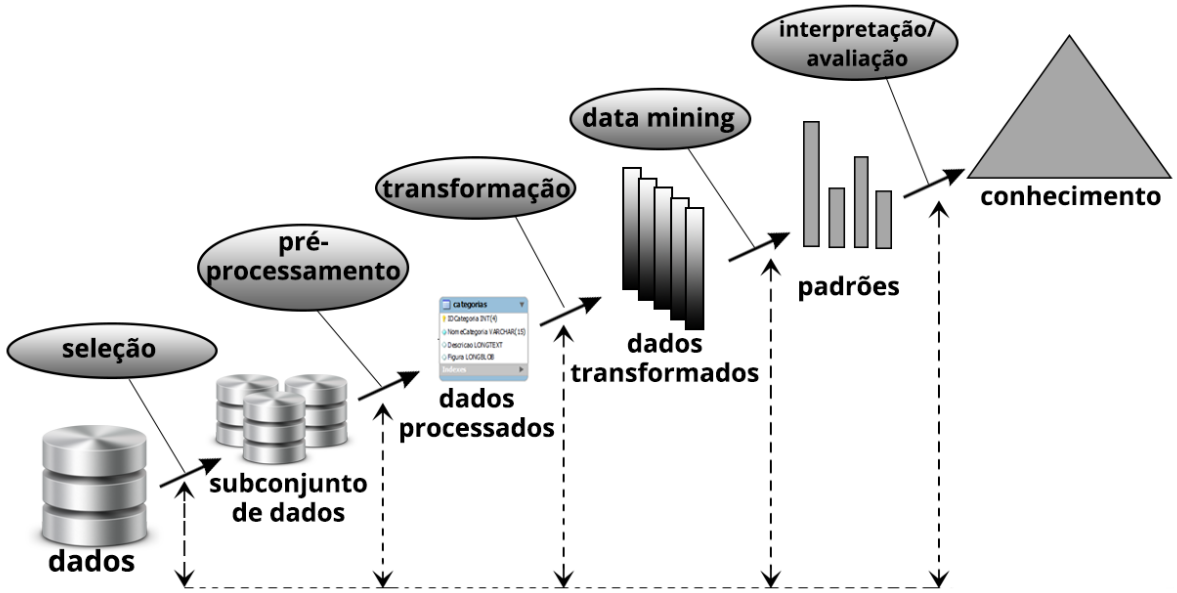


Figura 2.7: Etapas do processo de extração de conhecimento[30].

A primeira etapa do KDD é a Seleção dos dados, é necessário uma analise da relevância dos dados, selecionado um conjunto de dados, onde será executada o processo de extração de conhecimento.

Com o conjuntos de dados pronto, é importante executar a limpeza dos dados, visando criar uma consistência, remover dados errôneos, definir um padrão para valores faltantes, analisar redundâncias e qualquer outro tipo de inconsistências que podem interferir nos resultados da mineração, essa etapa é definida como pré-processamento.

A etapa seguinte é denominada transformação, é realizada uma edição adequada no conjunto de dados para que os algoritmos de mineração sejam aplicados corretamente.

A etapa de Mineração de Dados é realizada a exploração dos dados, análise e aplicação das técnicas de mineração, buscando encontrar padrões ou regras no determinado conjunto de dados.

Por fim, a etapa de interpretação consiste em analisar os resultados da mineração.

2.6.1 Tipos de aprendizado

A mineração de dados como apresentado anteriormente, engloba um conjunto de técnicas de *machine learning*, o que torna essencial entender a maneira que são divididos tais algoritmos de aprendizagem, que são organizados com base no resultado.

A Figura 2.8 caracteriza uma hierarquia de aprendizagem, de acordo com os tipos de aprendizagem. No topo encontramos a aprendizagem indutiva (processo pelo qual são realizadas as generalizações a partir dos dados). Em seguida, surgem os tipos de aprendizagem supervisionada (preditivo) e não supervisionada (descritivo).

Figura 2.8: Hierarquia da aprendizagem [12].

O aprendizado de máquina é classificado em três tipos, detalhados a seguir:

2.6.1.1 Aprendizado supervisionado

O principal objetivo do aprendizado supervisionado é aprender a fazer um mapeamento do *input* de dados para o *output* de dados, baseado em valores fornecidos corretamente por um supervisor ou indutor.

O processo de aprender possui uma fase de aprendizado com dados de testes, denominada treinamento, onde o indutor consegue extrair um bom classificador a partir de um conjunto de dados de entrada e resultados rotulados corretamente para todas as instancias dos dados. A saída do indutor, o classificador, pode então ser usada para classificar exemplos novos (ainda não rotulados) com a meta de predizer corretamente o rótulo de cada um. Também é reforçada a importância da generalização, que é a habilidade de produzir *outputs* razoáveis para *inputs* que não foram rotulados na fase de treinamento.

2.6.1.2 Aprendizado não supervisionado

Em tarefas de descrição, o principal objetivo consiste em explorar, ou descrever, um conjunto de dados. Essas tarefas ignoram o atributo de saída. Por esse motivo, diz-se que estes algoritmos seguem o paradigma de aprendizagem não supervisionada, diferentemente da aprendizagem supervisionada que possui características preditivas. Por exemplo, uma tarefa descritiva de agrupamento de dados tem por meta encontrar grupos de objetos

semelhantes no conjunto de dados. Outra tarefa descritiva consiste em encontrar regras de associação que relacionam um grupo de atributos com outro grupo de atributos.

2.6.1.3 Aprendizado por esforço

Uma tarefa que possui um paradigma diferente das anteriores, porém não menos importante, o aprendizado por reforço é necessário em aplicações cuja saída do sistema seja uma sequência de ações e, nesse caso, o que importa é a política definida pelo conjunto de ações onde o objetivo é reforçar, ou recompensar, uma ação considerada positiva, e punir uma ação considerada negativa para atingir um determinado objetivo. Uma única ação não é importante e não existe uma ação que seja melhor do que as outras em um estado intermediário, o que torna uma ação boa é se ela faz parte de uma política que levará ao alcance do objetivo. Assim, o programa deverá aprender com base em ações corretas ou incorretas realizadas anteriormente para criar uma boa política.

2.6.2 Técnicas e algoritmos de *DataMining*

Como já mencionado anteriormente, a mineração de dados é um conjunto de técnicas multidisciplinares que engloba uma vasta quantidade de tecnologias, na área da tecnologia da informação, nesta seção é apresentada as principais técnicas que serão focadas no desenvolvimento desse projeto. Para cada uma das demais técnicas, existem diversos algoritmos desenvolvidos.

2.6.2.1 Classificação

A classificação (também conhecida como árvores de classificação ou árvores de decisão) é uma técnica baseada no aprendizado supervisionado, possui algoritmos de mineração de dados que criam um guia passo a passo para determinar a saída de uma nova instância de dados. A árvore é criada da seguinte maneira: uma árvore em que cada nó na árvore representa um ponto onde uma decisão deve ser tomada com base na entrada, e se move para o próximo nó e o próximo até chegar a uma folha que lhe diz a saída prevista.

A classificação usa o conceito de usar um "conjunto de treinamento"para produzir um modelo. Isso leva um conjunto de dados com valores de saída conhecidos e usa este conjunto para produzir um modelo. Então, sempre que tiver um novo ponto de dados, com um valor de saída desconhecido, coloca-se o modelo e produz o resultado esperado.

2.6.2.2 Clusterização

A clusterização é uma técnica baseada no aprendizado não supervisionado, permite que um usuário faça grupos de dados para determinar os padrões dos dados. Clusterização tem suas vantagens quando o conjunto de dados é definido e um padrão geral precisa ser determinado a partir dos dados. Pode-se criar um número específico de grupos, dependendo das necessidades. Uma distinção entre classificação e clusterização é que cada atributo no conjunto de dados será usado para analisar os dados, enquanto a classificação usa apenas um subconjunto dos dados. Uma grande desvantagem é que o usuário precisa saber com antecedência quantos grupos deseja criar e para um usuário sem conhecimento real de seus dados, isso pode ser difícil.

2.7 *Weka*

Weka é uma coleção de algoritmos do estado da arte de *Machine Learning* para a realização de atividades de mineração de dados [31]. A sigla resulta de uma abreviação da expressão *Waikato Environment for Knowledge Analisys*[8], segundo Witten et al. [15], foi desenvolvido pela Universidade de Waikato, situada na Nova Zelândia, implementado pela primeira vez em sua forma moderna em 1997. Ele é composto por algoritmos de aprendizagem de máquina conjuntamente com uma coleção de recursos que realizam pré-processamento, regressão, classificação, clusterização, aplicação de regras de visualização dos dados e apresentação de resultados [31].

A ferramenta é desenvolvida utilizando a linguagem de programação Java e sua distribuição segue os termos da GNU (*GNU General Public License version 3.0 (GPLv3)*, ou Licença Pública Geral). Ela contém uma GUI voltada para a interação com arquivos de dados e produção de resultados visuais. Possui uma API geral, tornando possível incorpora-lo, como qualquer outra biblioteca, aos seus próprios aplicativos que realizam tarefas de mineração de dados automatizadas ao lado do servidor.

O software apresenta uma ampla variedade de recursos e ferramentas, como, por exemplo, o suporte a todas as etapas do processo experimental de mineração de dados, desde a preparação dos dados de entrada, análise estatística de esquemas de aprendizagem, até a visualização dos dados e apresentação dos resultados. Possui uma variedade de algoritmos de treinamento e ferramentas de pré-processamento que são apresentadas em uma interface amigável ao usuário. Possibilita, ainda, a integração direta com bancos de dados, o que permite ao usuário obter os dados diretamente da base e salvá-los em formato adequado para uso posterior no *Weka*.

O *Weka* utiliza o formato *ARFF* (*Attribute-Relation File Format*), é um arquivo de texto *ASCII* que descreve uma lista de instâncias que compartilham um conjunto de atributos [14]. Arquivos *.arff* possuem duas subdivisões:

- *Header*: Contém o nome da relação, uma lista dos atributos (as colunas nos dados) e seus tipos. A Figura 2.9 apresenta o *Header* do arquivo *Iris.arff*, um dos conhecidos exemplos que o *Weka* trás após o *download*.

```
1  % 1. Title: Iris Plants Database
2  %
3  % 2. Sources:
4  %      (a) Creator: Matheus Santana
5  %      (b) Date: July, 2017
6  %
7  @RELATION iris
8
9  @ATTRIBUTE sepallength  NUMERIC
10 @ATTRIBUTE sepalwidth   NUMERIC
11 @ATTRIBUTE petallength  NUMERIC
12 @ATTRIBUTE petalwidth   NUMERIC
13 @ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Figura 2.9: *Header* do arquivo *Iris.arff*.

[8]Traduzido como Ambiente Waikato para Análise do Conhecimento

- *Data*: Apresenta os dados de cada atributo para sua determinada classe, a Figura 2.10 apresenta a seção *Data* do arquivo *Iris.arff*.

```
1   @DATA
2   5.1,3.5,1.4,0.2,Iris-setosa
3   4.9,3.0,1.4,0.2,Iris-setosa
4   4.7,3.2,1.3,0.2,Iris-setosa
5   4.6,3.1,1.5,0.2,Iris-setosa
6   5.0,3.6,1.4,0.2,Iris-setosa
7   5.4,3.9,1.7,0.4,Iris-setosa
8   4.6,3.4,1.4,0.3,Iris-setosa
9   5.0,3.4,1.5,0.2,Iris-setosa
10  4.4,2.9,1.4,0.2,Iris-setosa
11  4.9,3.1,1.5,0.1,Iris-setosa
```

Figura 2.10: *Data* do arquivo *Iris.arff*.

A ferramenta ainda se destaca entre as demais disponíveis no mercado, tendo em vista que é de distribuição livre e multiplataforma, por ser criado em linguagem Java, como mencionado anteriormente, o que o torna adaptável a diferentes sistemas operacionais, como *Windows, GNU/Linux e Mac OS*.

A Figura 2.11 apresenta a interface inicial do *Weka* na plataforma *Mac OS*.

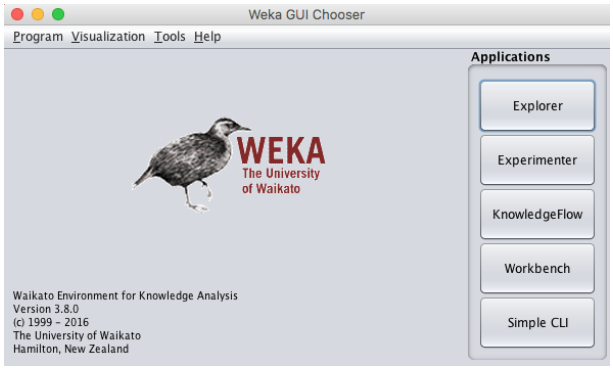


Figura 2.11: Interface inicial do *Weka* na plataforma *Mac OS*.

Segundo Witten et al. [15], a ferramenta apresenta ao usuário quatro distintos tipos de interfaces gráficas possíveis, além de uma interface mais simples, por linha de comando, como demonstrado abaixo:

- *Explorer*: Oferece ao usuário a possibilidade de acesso às opções existentes na barra de menu, bem como possibilita ao usuário carregar os dados a serem utilizados e verificar os resultados gerados pelos algoritmos de mineração. Entretanto, uma das desvantagens do modo *Explorer* (Figura 2.12) é que todo o conjunto de dados utilizado é mantido em memória, limitando-se a problemas de pequeno e médio porte.

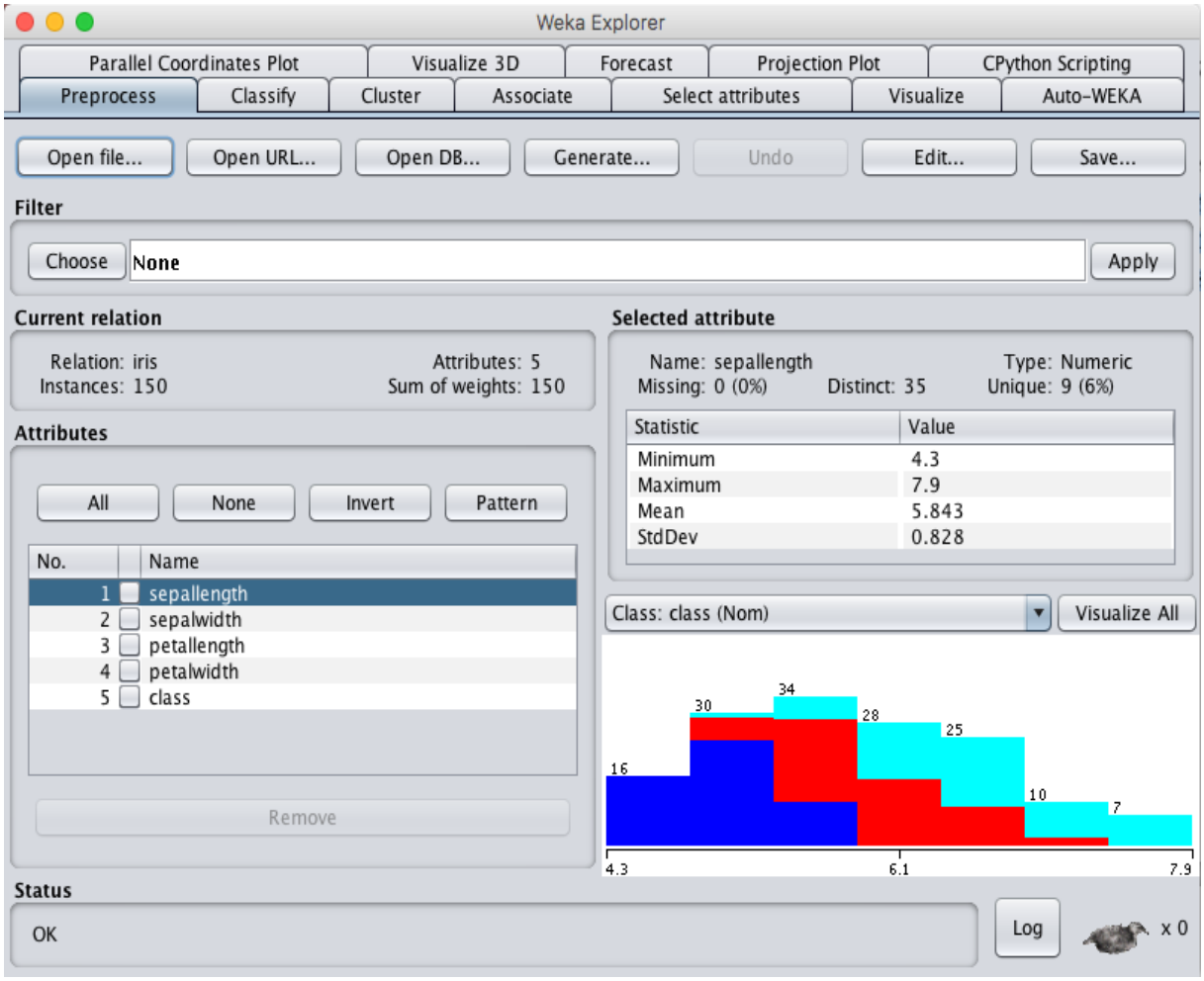


Figura 2.12: Interface *Explorer* na plataforma *Mac OS*.

- *Experimenter*: Tem por objetivo facilitar a identificação dos métodos e parâmetros nas técnicas de classificação e regressão mais adequados para determinado problema. A Figura 2.13 apresenta a interface que foi desenvolvida com o intuito de facilitar ao usuário a comparação de várias técnicas de aprendizagem, tornando mais fácil a execução de classificadores e filtros com diferentes definições de parâmetros sobre um conjunto de dados, a coleta de estatísticas de desempenho e a execução de testes significativos. Essa interface automatiza o processo experimental, as estatísticas podem ser armazenadas no formato *ARFF* e podem ser objeto de uma nova exploração de dados.

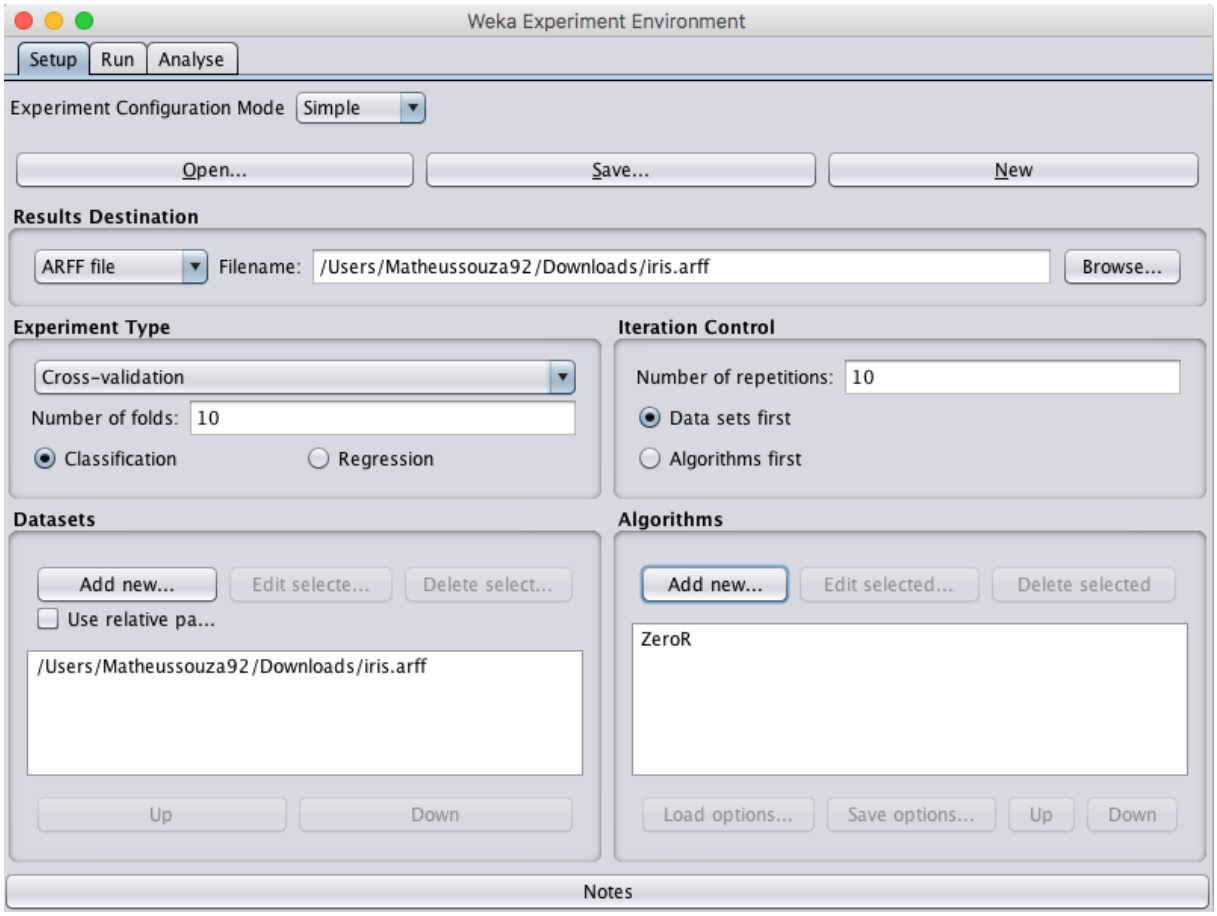


Figura 2.13: Interface *Experimenter* na plataforma *Mac OS*.

- *KnowledgeFlow*: Os usuários selecionam componentes *WEKA* a partir de uma barra de ferramentas, como ilustra a Figura 2.14, colocando-os em uma tela de layout que os conectam a um gráfico responsável pelo processamento e análise dos dados. Esta interface fornece uma alternativa ao *Explorer*, pois analisa como os dados fluem através do sistema, além de permitir o design e a execução de configurações para processamento de dados em fluxo por componentes conectados - que representam as fontes de dados - ferramentas de pré-processamento, algoritmos de mineração, métodos de avaliação e módulos de visualização.

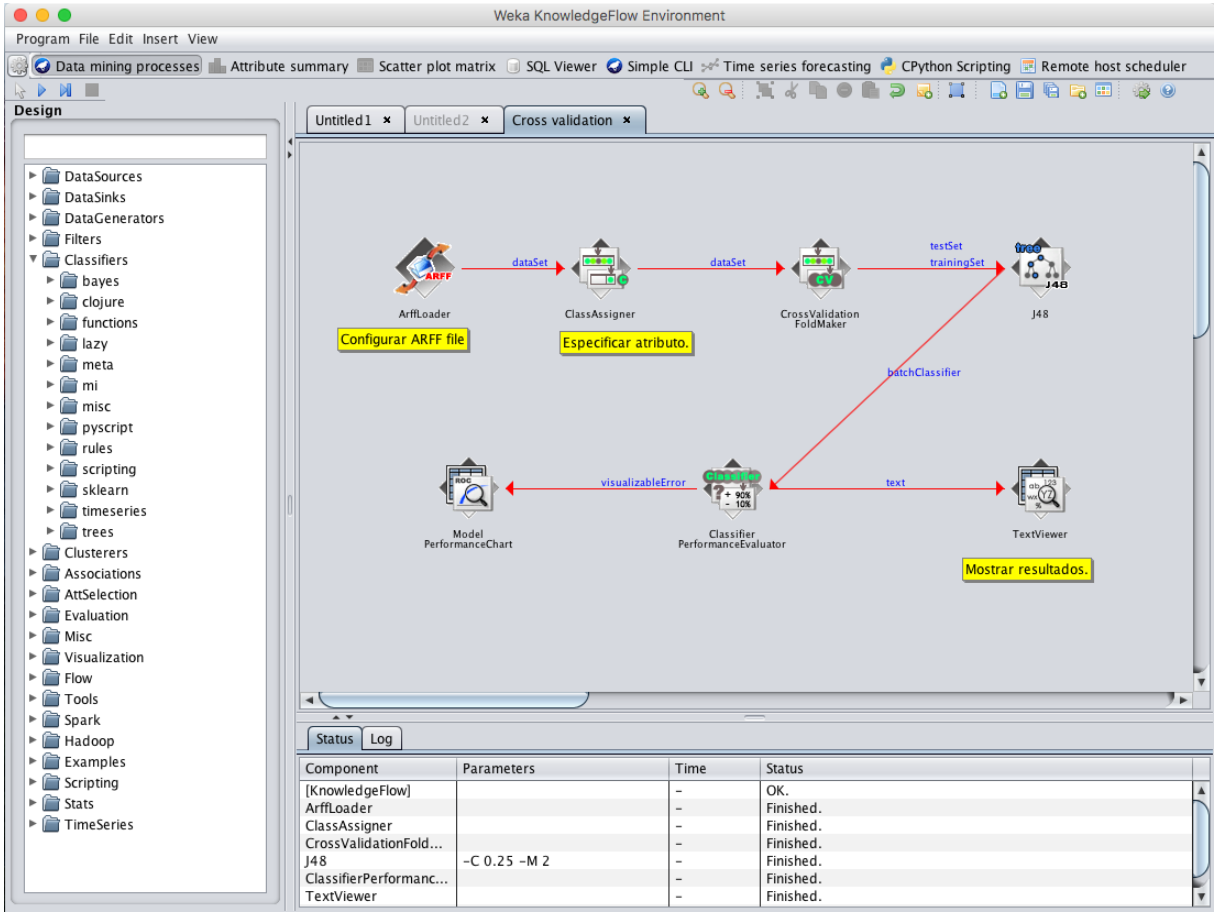


Figura 2.14: Interface *KnowledgeFlow* na plataforma *Mac OS*.

- *Workbench*: É um ambiente que combina todas as interfaces GUI em uma única interface. É útil se o usuário alterna com frequência entre duas ou mais interfaces distintas. A Figura 2.15 expõe o ambiente da interface *Workbench*.

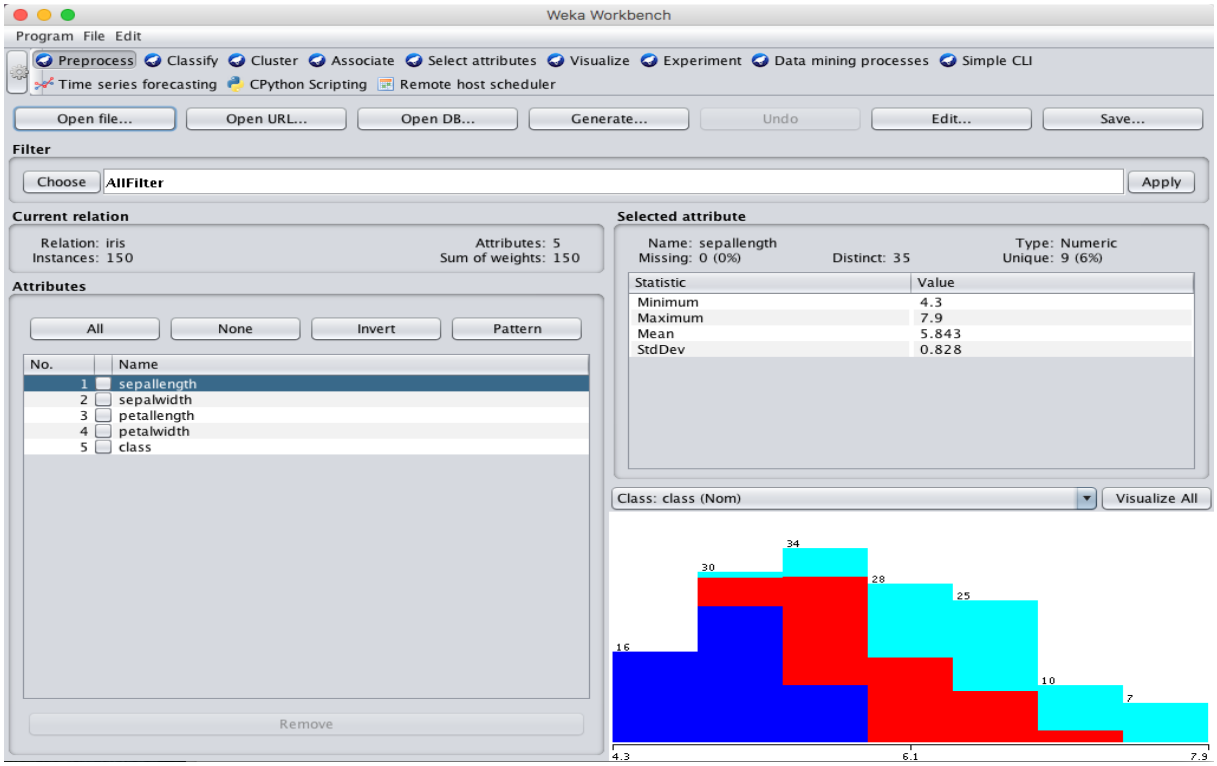


Figura 2.15: Interface *Workbench* na plataforma *Mac OS*.

- *Simple CLI*: A opção Simple CLI, demonstrada na Figura 2.16, apresenta dicas de como utilizar o *Weka* por linha de comando (via Terminal no *GNU/Linux/Mac OS* ou Prompt de Comando no *Windows*), e permite ao usuário informar os comandos a serem utilizados na mesma janela. Tal funcionalidade se diferencia das demais devido a possibilidade de escrever *shell scripts* usando a API completa de chamadas de linha de comando com parâmetros, permitindo ao usuário criar modelos, executar experimentos e realizar previsões sem uma interface gráfica de usuário.

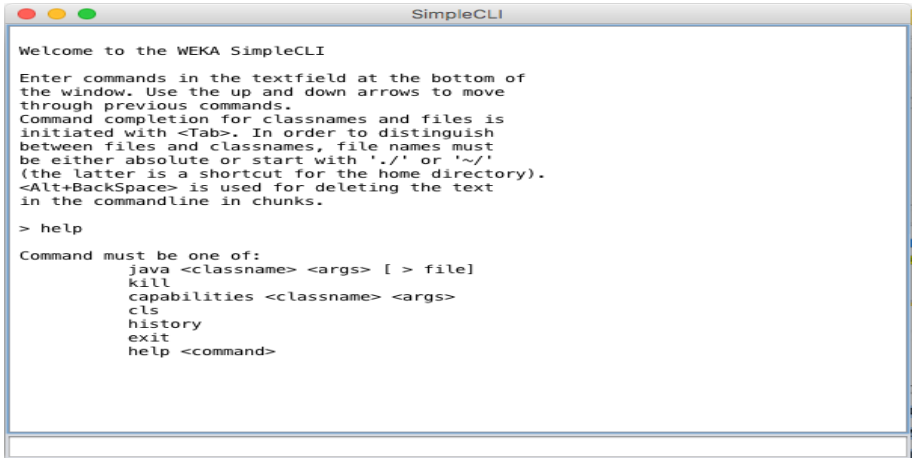


Figura 2.16: Interface *Simple CLI* na plataforma *Mac OS*.

Capítulo 3

Estudo de caso: Dados Globais do World Bank Group

Na mineração dos dados deste trabalho foi utilizada a arquitetura segundo Han e Kamber [13], detalhada no Seção 2.6. Conjuntamente foram utilizados o *Weka, Pentaho Data Integration e Workbench* como ferramentas para a conclusão de todo processo.

3.1 Coleta de Dados

O *World Bank*, criada em 1945, é uma instituição financeira internacional, formada por 189 países membros, assemelhando-se a uma associação[3]. Cada país membro é representado por um governador, que geralmente é selecionado entre ministros de finanças ou de desenvolvimento. Conjuntamente, os governadores de cada país membro formam o Conselho de Governadores, que se reúnem anualmente nas Assembleias de Governadores do Banco Mundial e do Fundo Monetário Internacional.

Em 2010 o World Bank começou a abrir seus dados ao público. Atualmente são disponibilizados diversos indicadores de dados e muitas ferramentas de visualização. Dentre essas ferramentas estão:

- *DataBank*: fornece dados de mais de 40 bases através de um acesso programático aos dados e metadados (APIs), em diversos idiomas: Inglês, Francês, Espanhol, Chinês e Árabe.
- *Open Data Readiness Assessment Tool*: permite que governos e agências avaliem, projetem e implementem iniciativas open data.
- *Maps.worldbank.org*: disponibiliza mapas de 143 países.
- *Climate Change Knowledge Portal*: é um centro de informações sobre o clima.
- *Microdata Library*: oferece acesso aos dados brutos ainda não tratados de mais de 700 questionários feitos a famílias e fontes.
- *Adepto*: ferramenta que automatiza a análise econômica dos dados pesquisado.
- *WITS*: ferramenta de dados que fornece acesso a dados comerciais e tarifárias internacionais.

Entre todas as ferramentas apresentadas acima, selecionou-se como objeto de pesquisa desse trabalho a *DataBank*. Ela fornece funções avançadas de seleção e exibição de dados, consultas personalizadas, *download* de dados, além da elaboração de gráficos e mapas. A ferramenta é subdividida em base de dados agrupadas por indicadores. Até o momento da produção deste trabalho existiam 61 bases de dados, constituídas de indicadores de desenvolvimento mundial, indicadores de capacidade estatística, estatística da educação, entre outros.

A Figura 3.1 apresenta a tela principal da ferramenta. Nesta primeira página é possível selecionar uma base de dados dentre as mais populares.

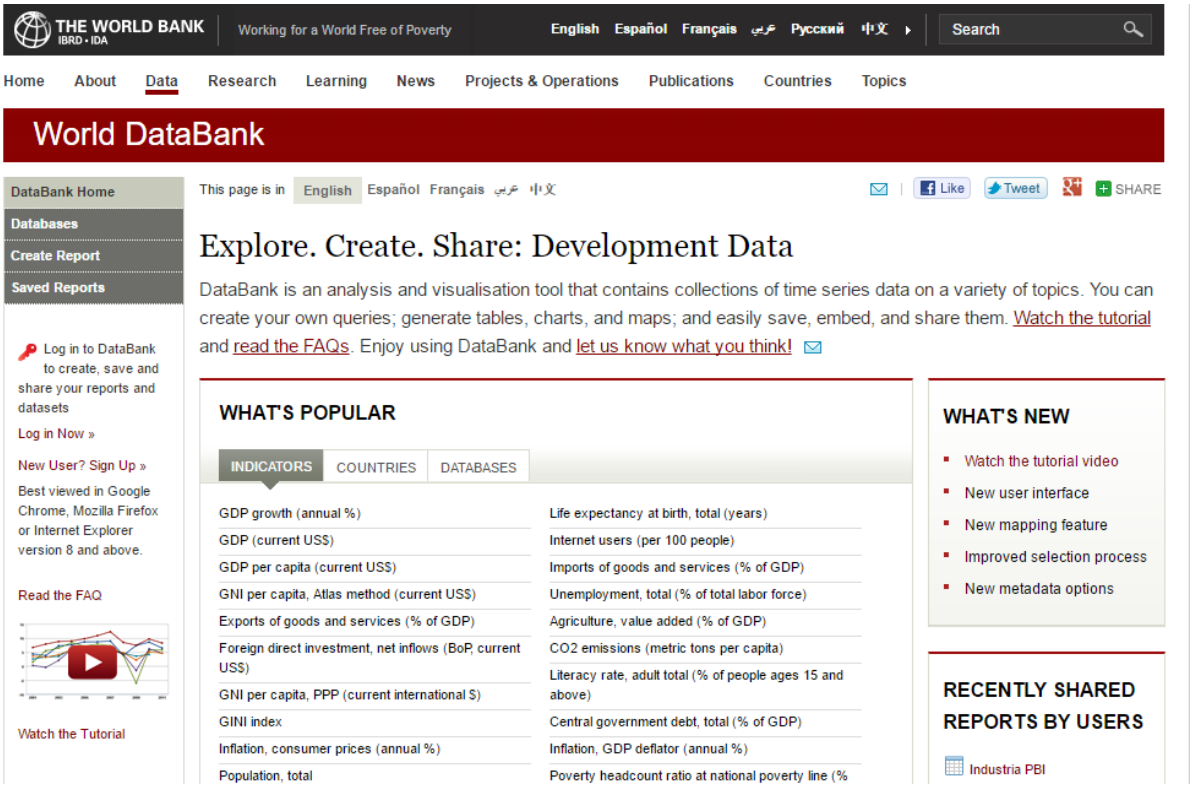


Figura 3.1: Tela inicial da ferramenta DataBank.

Ao clicar na base de dados escolhida, o usuário é encaminhado para uma nova tela, como ilustrado na Figura 3.2. Nesta tela é possível fazer seleções das variáveis que serão usadas a partir das seguintes opções: *Database*, *Country*, *Series*, *Time*. Ainda é disponibilizada, na mesma página, uma pré-visualização dos dados selecionados.

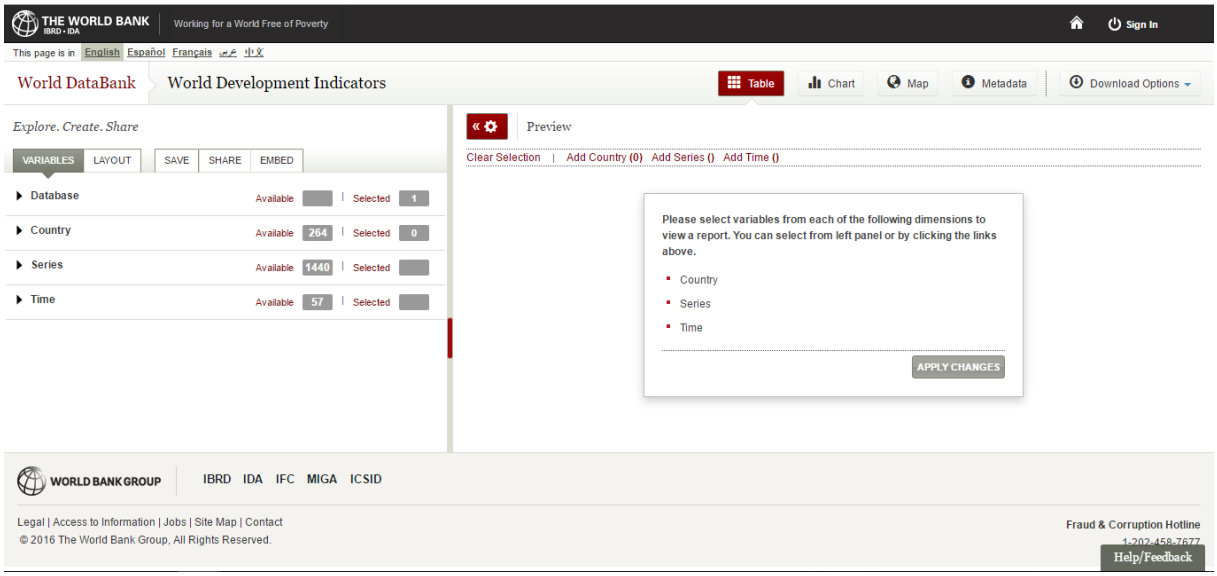


Figura 3.2: Tela seleção do dados.

As figuras seguintes, Figura 3.3 e Figura 3.4, apresentam gráficos e mapas que podem ser elaborados com o auxílio da própria ferramenta, a partir dos dados selecionados pelo usuário.

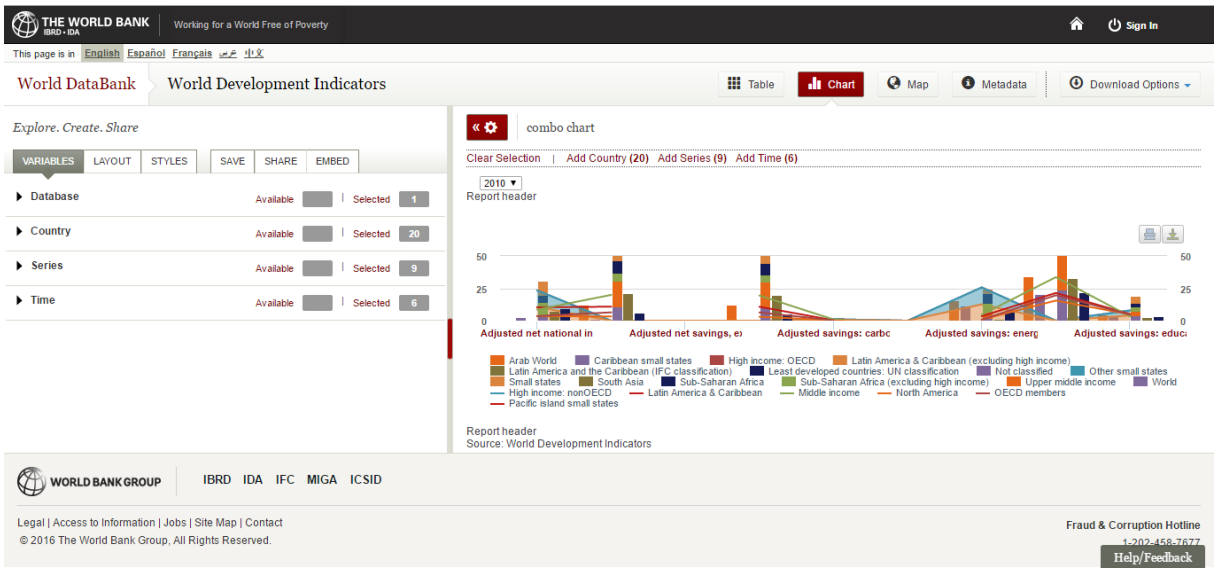


Figura 3.3: Imagem da visualização por gráfico.

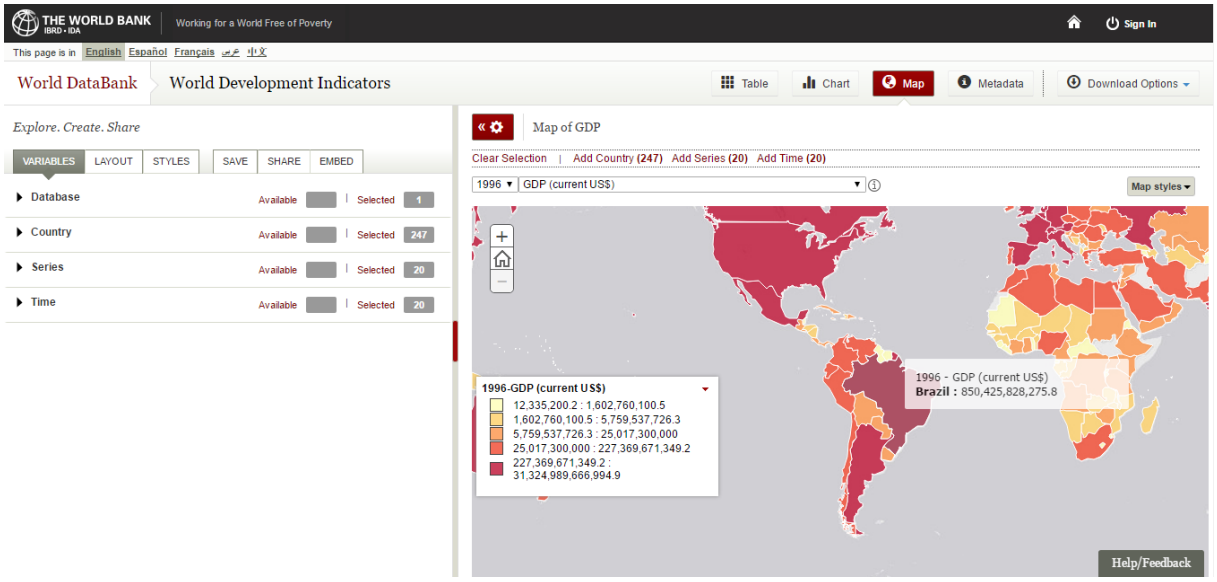


Figura 3.4: Imagem da visualização por mapa.

Foi utilizada a base de dados *World Development Indicators*(WDI) como fundamento deste trabalho, haja vista que ela é o maior conjunto de indicadores de desenvolvimento do *World Bank*. Foram selecionados todos os atributos da variável *Country* (países), totalizando 264 divididos entre países e agregados como União Europeia, conjuntamente com todos os atributos da variável *Series*, 1446 indicadores, e, por fim, os atributos situados no intervalo entre 2006 e 2015 da variável *Time*. A faixa de 10 anos foi escolhida pois em um intervalo muito grande os indicadores de um país pode mudar muito. O resultado dessa seleção é um retorno de 3817440 linhas que podem ser baixadas através do botão de *Download Options*, em formatos variados, como Excel, TXT, CSV e SDMX. Optou-se neste trabalho pelo formato CSV para *download*, formato que facilita a importação para o banco de dados. Assim foi realizada a coleta de dados desta pesquisa de forma direta no banco de dados do próprio *World Bank*.

3.2 Tratamento dos Dados

A extração do banco do *World Bank*, resultou em duas planilhas, *"Indicators"* e *"Definition"* respectivamente. A primeira é distribuída pelas colunas *"Contry Name"*(nome do país), *"Country Code"*(código do país), *"Series Code"*(código do indicador), *"Series Name"*(nome do indicador), e as colunas dos anos de 2006 a 2015. A segunda planilha é distribuída através das colunas *"Code"*(código do indicador), *"Indicator Name"*(nome do indicador), *"Long definition"*(definição do indicador) e *"Source"*(fonte da informação).

Na etapa destinada ao tratamento dos dados foi elaborado um modelo de DW, utilizando o SGBD MySQL, como explicado na Seção 2.2, que atendesse o propósito deste trabalho, subdividido em 2 tabelas, sendo que cada atributo dessas tabelas representam colunas das planilhas extraídas previamente, detalhada na Tabela 3.1. A Tabela 3.1 relaciona o arquivo e sua tabela no banco de dados e a Figura 3.5 mostra o modelo desse

banco. O modelo representa a relação das duas tabelas tendo o código da série como chave identificadora.

Planilha	Tabela
Indicators	tbl_contry
Definition	tbl_indicators

Tabela 3.1: Relação entre planilhas e tabelas.

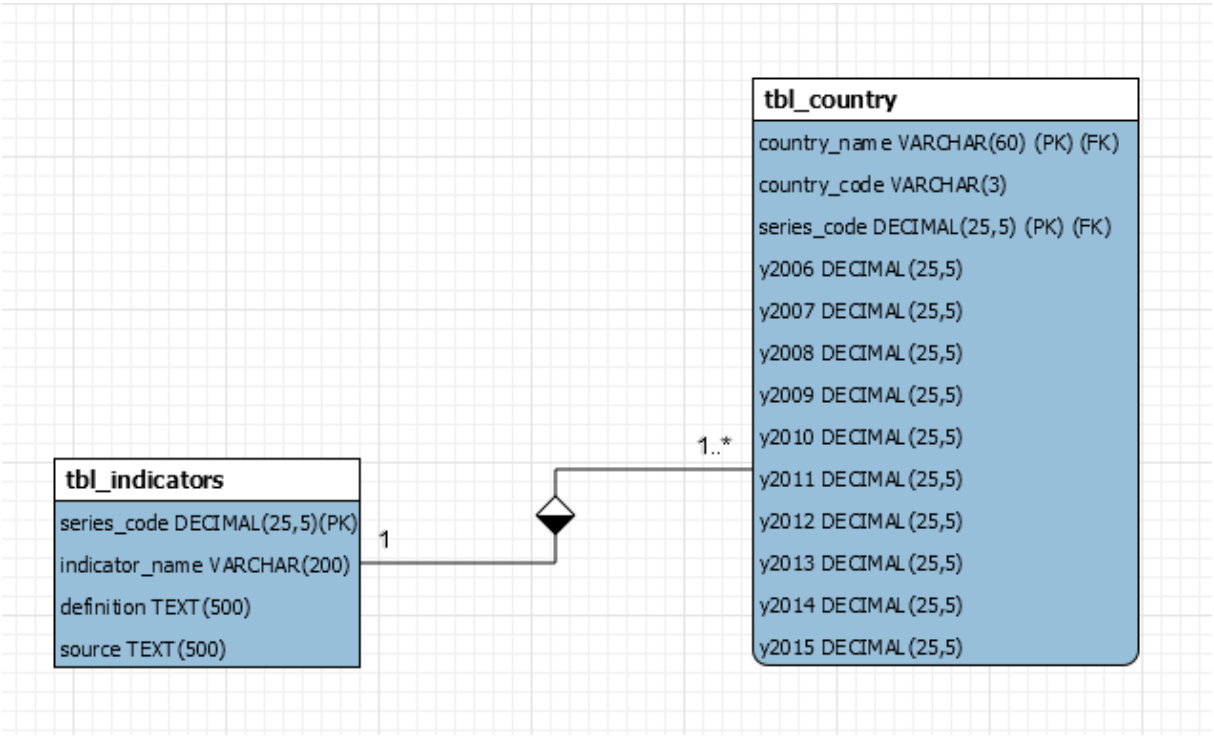


Figura 3.5: Modelo do banco.

Na etapa subsequente à elaboração do banco, utilizou-se a ferramenta *Pentaho Data Integration*, explicado na Seção 2.6, responsável pela limpeza e tratamento dos dados. Durante o tratamento foram desenvolvidos dois processos de tratamento de dados, que na ferramenta é chamado de *job*, um para cada tabela. As figuras seguintes apresentam a interface de cada *job*.

A Figura 3.6 expõe o *job* de tratamento e carga da tabela tbl_country. Este *job* é composto por cinco componentes. O primeiro é responsável pela leitura do arquivo, enquanto o segundo e o terceiro componentes realizam o tratamento dos dados, colocando todas as palavras em maiúsculo, além de retirar todos os acentos e caracteres especiais. Essas alterações tornaram-se necessárias para garantir que os atributos tenham nomes consistentes nas diferentes tabelas. No componente seguinte eliminam-se as colunas que

não serão utilizadas no trabalho, por não terem informações pertinentes a esse trabalho, e, por fim, o último componente é responsável pela carga no banco.

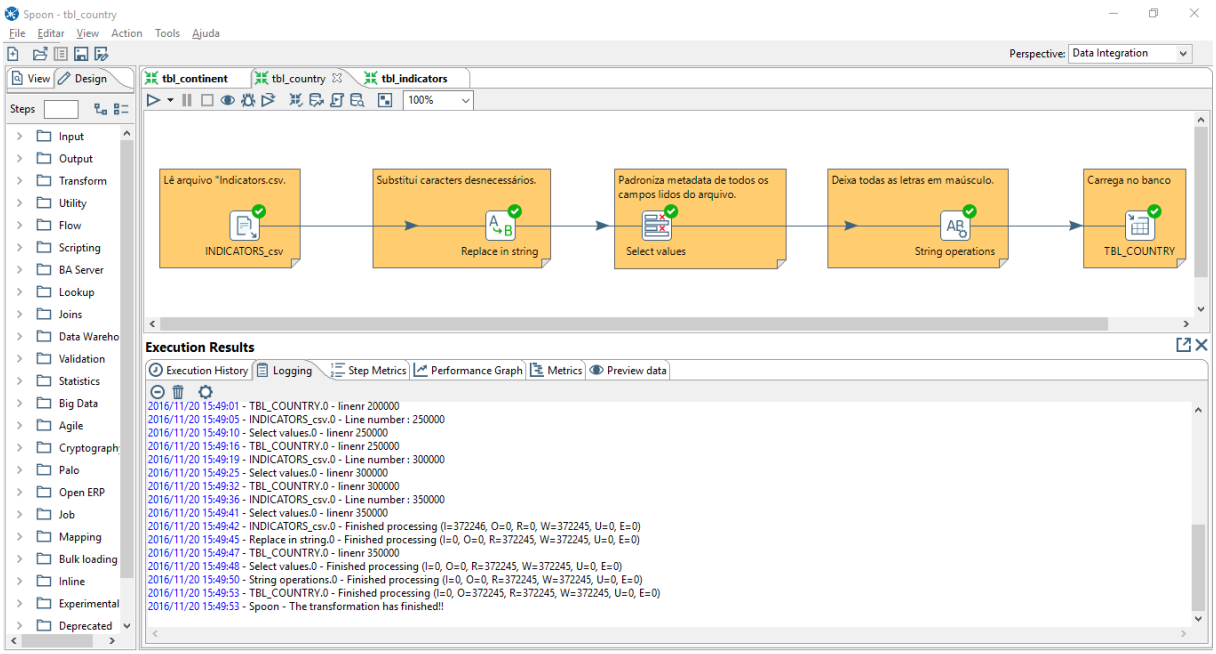


Figura 3.6: Job tbl_country.

A Figura 3.7 apresenta o *job* de tratamento e carga da tabela tbl_indicators. Este *job* é composto por quatro componentes. O primeiro componente é responsável pela leitura do arquivo de origem e os demais componentes referem-se ao tratamento dos dados, excetuando-se o último componente, que destina-se à carga no banco.

Ao fim desse processo os dados estão prontos para a etapa de mineração e análise.

3.3 Indicadores

Após uma análise detalhada de todos os indicadores baixados, foi decidido reduzi-los a apenas 125 indicadores, foram retirados dados que não estavam presentes em muitos países. Foi escolhido os indicadores mais populados para garantir uma maior confiabilidade nos resultados. Também foi reduzida a 33 países analisados, selecionando os mais conhecidos e os que mais tem dados em cada continente.

Os países usados foram: Argentina, Austrália, Áustria, Bolívia, Brasil, Canada, Chile, China, Colombia, Costa Rica, Cuba, Dinamarca, República Tcheca, Equador, Alemanha, Finlandia, França, Honduras, Índia, Israel, Itália, Japão, Irlanda, Korea, México, Portugal, Espanha, Suíça, Suécia, Turquia, Estados Unidos, Uruguai e Venezuela.

Abaixo será dada uma breve explicação sobre cada indicador usado no trabalho:

1. *Age Dependency Ratio(% of working-age population)*: Trata-se da relação da população dependentes(pessoas mais novas que 15 anos e mais velhas que 64) e a população com idade para trabalhar(entre 15 e 65 anos).

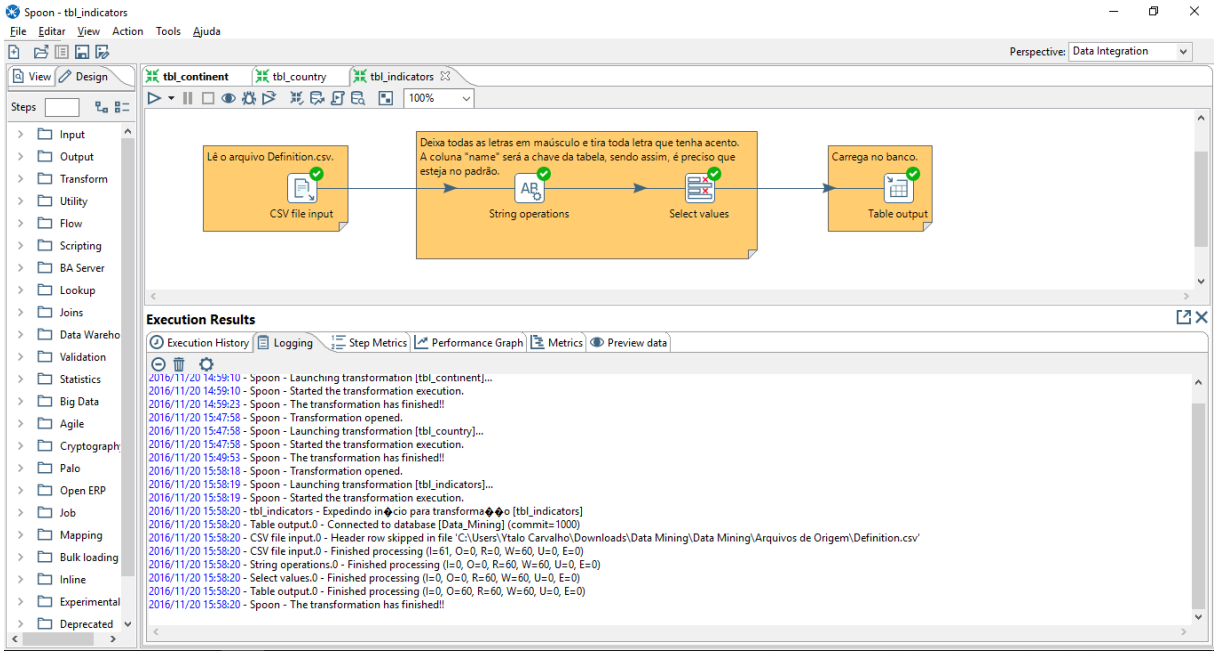


Figura 3.7: Job tbl_indicators.

2. *Age dependency ratio, old (% of working-age population)*: Trata-se da relação de pessoas idosas dependentes(mais de 64 anos) e a população com idade para trabalhar(entre 15 e 64 anos).

3. *Agricultural Raw Materials Exports (% of merchandise imports)*: Quantidade de materiais agrícolas exportados(excluindo exportação de petróleo e materiais combustíveis, pedras preciosas e metais), calculado pela porcentagem do total de exportações.

4. *Agricultural Raw Materials Imports*: Quantidade de materiais agrícolas importados(excluindo importação de petróleo e materiais combustíveis, pedras preciosas e metais), calculado pela porcentagem do total de importações.

5. *Agriculture Value Added Per Worker*: Medida de produtividade agrícola. O valor(em Dollar) acrescentado na agricultura por trabalhador.

6. *Agriculture Value Added(% of GDP)*: Valor da saída líquida do setor agrícola, porcentagem do PIB. A origem do valor é determinada pela ISIC(*International Standard Industrial classification).*

7. *Agriculture Value Added(Anual % Growth)*: Taxa anual de crescimento do valor adicionado a agricultura com base em moeda local constante. A origem do valor é determinada pela ISIC(*International Standard Industrial classification).*

8. *Bank Capital To Assets Ratio (%)*: É a proporção de capital bancário e reservas para o total de ativos(inclui todos os ativos não financeiros e financeiros).

9. *Bank Liquid Reserves to Bank Assets Ratio (%)*: Relação das participações dos depósitos em moeda nacional e os créditos de outros governos.

10. *Bank Nonperforming Loans to Total Gross Loans (%)*: São os valores totais dos empréstimos não performantes divididos pelo valor total da carteira de empréstimos.
11. *Broad Money(% of GDP)*: É a soma de moeda fora dos bancos frente ao PIB.
12. *Broad Money Growth(Anual %)*: É o crescimento anual da soma de moedas fora dos bancos.
13. *Broad Money To Total Reserves Ratio*: É a relacão do crescimento de moedas fora dos bancos com o total de reservas.
14. *Business Extent of Disclosure Index*: Índice(índice funciona de 0 a 10, com valores mais altos indicando a maior divulgação) que mede o quanto os investidores estão protegidos através da divulgação de propriedades e informações financeiras.
15. *Claims on Central Government(Annual Growth as % of Broad Money)*: Crescimento anual do créditos para o Governo Central(incluem empréstimos para instituições do governo).
16. *Claims on Central Government, ETC(% GDP)*: Créditos para o Governo Central(incluem empréstimos para instituições do governo) frente ao PIB..
17. *Claims on Other Sectors of the Domestic Economy (% of GDP)*: Créditos sobre outros setores da economia nacional, incluem crédito bruto do sistema financeiro às famílias, corporações não financeiras, governos estatutais e locais e fundos de segurança social.
18. *Claims on Other Sectors of the Domestic Economy (Annual Growth % of Broad Money)*: Crescimento anual dos créditos sobre outros setores da economia nacional, incluem crédito bruto do sistema financeiro às famílias, corporações não financeiras, governos estatutais e locais e fundos de segurança social.
19. *Claims on Private Sector (annual Growth as % of Broad Money)*: Créditos para setores privados, incluem crédito bruto do sistema financeiro a indivíduos, empresas e entidades públicas não financeiras não incluídas sob crédito interno líquido.
20. *Computer, Communications and Other Services(% of Commercial Service Exports)*: Exportações de serviços comerciais, incluem atividades como telecomunicações internacionais e serviçoes postais e correios; Dados de computadores; Transações de serviços relacionados com noticias entre residentes e não residentes; serviços de construção; *Royalties* e taxas de licença; serviços diversos de negócios, profissionais e técnicos; e serviços pessoais, culturais e recreativos.
21. *Computer, Communications and Other Services(% of Commercial Service Imports)*: Importações de serviços comerciais, incluem atividades como telecomunicações internacionais e serviçoes postais e correios; Dados de computadores; Transações de serviços relacionados com noticias entre residentes e não residentes; serviços de construção; *Royalties* e taxas de licença; serviços diversos de negócios, profissionais e técnicos; e serviços pessoais, culturais e recreativos.
22. *Consumer Price Index*:Reflete a alteração média no custo ao consumidor para adquirir uma cesta básica e serviços que podem ser fixados ou mudados em intevalos especificados, como anualmente.

23. *Cost of Business Start-UP Procedures(% of GNI Per Capita)*: Custo para registrar um negócio normalizado pelo percentual rendimento bruto nacional (GNI) *Per Capita*.

24. *Current Account Balance(% OF GDP)*: É a soma da balança comercial(exportações de bens de serviços menos as importações), o lucro líquido do exterior e as transferências correntes líquidas.

25. *Deposit Interest Rate (%)*: É a taxa paga por bancos comerciais ou similares por demanda, hora ou depósito de poupança.

26. *Domestic Credit To Private Sector (% Of GDP)*: Refere aos recursos financeiros fornecidos ao setor privado pelas sociedades financeiras, como através de empréstimos, compras de valores mobliários e créditos comerciais e outras contas a receber, que estabelecem uma reclamação de pagamento.

27. *Domestic Credit To Private Sector By Banks (% Of GDP)*: Refere os recursos financeiros fornecidos ao setor privado por outras corporações de depósitos, como através de emprestimos, compras de valores mobiliários e créditos comerciais e outras contas a receber, que estabelecem a reivindicação de reembolso.

28. *Exports Of Goods And Services (% Of Gdp)*: Representam o valor de todos os bens e outros serviços de merdado fornecidos ao resto do mundo. Incluem o valor da mercadoria, frete, seguros, transporte, viagens, *Royalties*, taxas de licença e outros serviços.

29. *Exports Of Goods And Services (Annual % Growth)*: Taxa anual de crescimento das exportações de bens e serviços com base na moeda local constante.

30. *External Balance On Goods And Services (% Of GDP)*: Balanço externo de bens de serviços igual a exportação de bens e serviços menos importações de bens e serviços.

31. *Final Consumption Expenditure, Etc. (% Of GDP)*: É a soma das despesas finais de consumo da família e despesas finais de consumo do governo geral.

32. *Final Consumption Expenditure, Etc. (Annual % Growth)*: Crescimento médio anual das despesas finais de consumo com base em moeda local constante.

33. *Food Exports (% Of Merchandise Exports)*: Exportação de alimento por porcentagem de mercadorias exportadas.

34. *Food Imports (% Of Merchandise Imports)*: Importação de alimento por porcentagem de mercadorias importadas.

35. *Foreign Direct Investment, Net Inflows (% Of Gdp)*: É a soma do capital social reivestimento dos resultados, outro capital de longo prazo e capital de curto prazo como mostrado no balanço dos pagamentos. Mostra o fluxo líquido na economia relatórios de investidores estrangeiros, e está dividido pelo PIB.

36. *Foreign Direct Investment, Net Outflows (% Of Gdp)*: Refere aos fluxos de investimentos direto na economia. É a soma do capital social, reinvestido dos resultados e outros capitais. Mostra os fluxos líquidos de investimento da economia de relatórios para o resto do mundo, e está dividido pelo PIB.

37. *Fuel Exports (% Of Merchandise Exports)*: Exportação de combustível por porcentagem de mercadorias exportadas.

38. *Fuel Imports (% Of Merchandise Imports)*: Importação de combustível por porcentagem de mercadorias importadas.

39. *GDP (Current US$)*: É a soma do valor bruto adicionado por todos os produtores residentes na economia mais qualquer imposto sobre os produtos e menos quaisquer subsídios não incluídos no valor do produto. O PIB é a soma do valor bruto adicionado por todos os produtores residentes na economia, mais quaisquer impostos sobre os produtos e menores, quaisquer subsídios não incluídos no valor dos produtos.

40. *GDP Growth (Annual %)*: Taxa anual de rrescimento da porcentagem do Produto Interno Bruto a preços de mercado baseados em moedas correntes constantes. O PIB é a soma do valor bruto adicionado por todos os produtores residentes na economia, mais quaisquer impostos sobre os produtos e menores, quaisquer subsídios não incluídos no valor dos produtos.

41. *GDP Per Capita (Current US$)*: É o Produto Interno Bruto dividido pela população. O PIB é a soma do valor bruto adicionado por todos os produtores residentes na economia, mais quaisquer impostos sobre os produtos e menores, quaisquer subsídios não incluídos no valor dos produtos.

42. *GDP Per Capita Growth (Annual %)*: Taxa anual de crescimento da percentagem do PIB por habitante com base em moedas correntes constantes. O PIB é a soma do valor bruto adicionado por todos os produtores residentes na economia, mais quaisquer impostos sobre os produtos e menores, quaisquer subsídios não incluídos no valor dos produtos.

43. *GDP Per Capita, PPP (Current International $)*: PIB Per capita com base na paridade de poder de compra(PPP) DE COMPRA (PPP). PPP PIB é produto interno bruto convertido a dólares internacionais usando taxas de participação do poder de compra. Um dólar internacional tem o mesmo poder de compra sobre o PIB como dólar americano tem nos Estados Unidos.

44. *General Government Final Consumption Expenditure (% Of GDP)*: Despesas finais de consumo do governo geral inclui todas as despesas correntes do governo para compras de bens e serviços.

45. *General Government Final Consumption Expenditure (Annual % Growth)*: Crescimento anual da receita de consumo final do governo geral com base em moeda local constante. Inclui todas as despesas correntes do governo para compras de bens de serviços.

46. *GNI (Current US$)*: É a soma do valor adicionado por todos os produtores residentes e qualquer imposto sobre os produtos não incluído a valorização da saída.

47. *GNI Growth (Annual %)*: O crescimento da soma do valor adicionado por todos os produtores residentes e qualquer imposto sobre os produtos não incluído a valorização da saída.

48. *GNI Per Capita Growth (Annual %)*: Taxa de crescimento anual da renda Per Capita com base em moeda local constante. GNI é o rendimento nacional bruto dividido pela população.

49. *Gni Per Capita, Ppp (Current International $)*: GNI PPP é o rendimento nacional bruto convertido em dólares internacionais utilizando taxas de paridade.

50. *Gross Capital Formation (% Of GDP)*: Consiste em descontos sobre adição a economia mais as alterações no nível inventário.

51. *Gross Capital Formation (Annual % Growth)*: Taxa de crescimento anual da formação bruta de capital com base na moeda local constante.

52. *Gross Capital Formation (Current US$)*:Consiste em descontos sobre adição a economia mais as alterações no nível inventário frente ao dólar.

53. *Gross Domestic Savings (% Of GDP)*: Crescimento das economias brutas domésticas. São calculadas com base no PIB, menos as despesas finais de consumo.

54. *Gross Fixed Capital Formation (% Of GDP)*:Formação bruta de capital fixo. Inclui melhorias de terras; compras de plantas, máquina e equipamentos; e a construção de estradas, estradas de ferro, escolas, escritórios, hospitais e moradias residenciais privadas e edifícios comerciais e industriais.

55. *Gross Fixed Capital Formation (Annual % Growth)*: Crescimento anual médio da formação bruta de capital fixo com base em moeda local constante.

56. *Gross Fixed Capital Formation (Current US$)*: Formação bruta de capital fixo. Inclui melhorias de terras; compras de plantas, máquina e equipamentos; e a construção de estradas, estradas de ferro, escolas, escritórios, hospitais e moradias residenciais privadas e edifícios comerciais e industriais.

57. *Gross Fixed Capital Formation, Private Sector (% Of GDP)*: Formação bruta do capital fixo de investimento privado. Inclui demonstrações brutas pelo setor privado sobre a adições a seus ativos domésticos simples.

58. *Gross National Expenditure (% Of GDP)*: É a soma das despesas final do consumo doméstico, da despesa final do consumo geral e da formação bruta do capital.

59. *Gross National Expenditure (Current US$)*: É a soma das despesas final do consumo doméstico, da despesa final do consumo geral e da formação bruta do capital.

60. *Gross Savings (% Of GDP)*: É a Receita nacional bruta menos o consumo total, mais transferências líquidas.

61. *Gross Savings (% Of Gni)*: É a Receita nacional bruta menos o consumo total, mais transferências líquidas.

62. *Gross Savings (Current US$)*: É a Receita nacional bruta menos o consumo total, mais transferências líquidas. O dado está em dólar corrente.

63. *Gross Value Added At Factor Cost (Current US$)*: Valor bruto acrescentado ao fator de custo.

64. *High-Technology Exports (% Of Manufactured Exports)*: Exportações de alta tecnologia. São produtos de alta complexidade tecnolótica como espaçonave, computadores, farmaceuticos, instrumentos científicos e máquinas elétricas.

65. *Household Final Consumption Expenditure (Annual % Growth)*: Crescimento anual de porcentagem de despesas final de bens de conumo com base em moeda local constante.

66. *Household Final Consumption Expenditure Per Capita (Constant 2010 US$)*: Despesa final do consumo familiar Per Capita.

67. *Household Final Consumption Expenditure Per Capita Growth (Annual %)*: Crescimento anual da despesa final do consumo familiar Per Capita.

68. *Household Final Consumption Expenditure, Etc. (% Of GDP)*: Despesa final do consumo familiar frente a porcentagem do PIB.

69. *Household Final Consumption Expenditure, Etc. (Annual % Growth)*: Crescimento anual da despesa final do consumo familiar frente a porcentagem do PIB.

70. *Ict Service Exports (% Of Service Exports, Bop)*: Exportações de serviços de tecnologia de informação e comunicação. Incluem serviços de computador e comunicação e serviços de informação.

71. *Ida Resource Allocation Index (1=Low To 6=High)*: Índice de atribuição de recurso da IDA. É obtido por cálculo da pontuação média para cada grupo. Os países são classificados de 1 a 6.

72. *Imports Of Goods And Services (% Of Gdp)*: Importação de bens de serviços. Representa o valor de todos os bens e outros serviços de mercado recebidos pelo resto do mundo.

73. *Imports Of Goods And Services (Annual % Growth)*: Crescimento anual da importação de bens de serviços. Representa o valor de todos os bens e outros serviços de mercado recebidos pelo resto do mundo.

74. *Imports Of Goods And Services (Current US$)*: Importação de bens de serviços em dólar corrente. Representa o valor de todos os bens e outros serviços de mercado recebidos pelo resto do mundo.

75. *Industry, Value Added (% Of GDP)*:Compreende ao valor adicionado em mineração, fabricação, construção, eletricidade, água e gás. Saída líquida de um setor após adicionar todas as saídas e subtrações de entrada intermediárias.

76. *Industry, Value Added (Annual % Growth)*: Crescimento anual do valor adicionado em mineração, fabricação, construção, eletricidade, água e gás. Saída líquida de um setor após adicionar todas as saídas e subtrações de entrada intermediárias.

77. *Inflation, Consumer Prices (Annual %)*: A inflação, tal como medida pelo índice de preços no consumidor, reflete a variação percentual do custo para o consumidor médio de adquirir uma cesta de produtos e serviços que podem ser fixados ou alterados a intervalos especificados, como anualmente.

78. *Insurance And Financial Services (% Of Commercial Service Exports)*: Seguros e serviços financeiros cobrem os seguros de frete de mercadorias exportadas e outros seguros diretos como seguros de vida; serviços de intermediação financeira, como comissões, operções de câmbio e serviços de corretagem; serviços auxiliares, como mercado financeiro, serviços operacionais e regulamentares.

79. *Insurance And Financial Services (% Of Commercial Service Imports)*: Seguros e serviços financeiros cobrem os seguros de frete de mercadorias importadas e outros seguros diretos como seguros de vida; serviços de intermediação financeira, como comissões, operções de câmbio e serviços de corretagem; serviços auxiliares, como mercado financeiro, serviços operacionais e regulamentares.

80. *Interest Rate Spread (Lending Rate Minus Deposit Rate, %)*: A taxa de juros carregada por bancos de empréstimos a clientes privados do setor mens a taxa de juros paga por bancos comerciais ou similares por demanda, hora ou depósito de poupança.

81. *Lending Interest Rate (%)*: É a taxa bancária que agrupa as necessidades de financiamento a curto e médio prazo do setor privado.

82. *Listed Domestic Companies, Total*: Quantidade de empresas domésticas, incluindo empresas estrangeiras que são exclusivamente listadas.

83. *Manufactures Exports (% Of Merchandise Exports)*: Percentual de exportação de manufaturas frente a quantidade de mercadorias exportadas.

84. *Manufactures Imports (% Of Merchandise Imports)*: Percentual de importações de manufaturas frente a quantidade de mercadorias importadas.

85. *Manufacturing, Value Added (% Of GDP)*: É a saída líquida do setor de manufatura após adicionar todas as saídas e entradas intermediárias.

86. *Manufacturing, Value Added (Annual % Growth)*: Crescimento anual das saídas líquidas do setor de manufatura após adicionar todas as saídas e entradas intermediárias.

87. *Market Capitalization Of Listed Domestic Companies (% Of Gdp)*: Porcentagem do PIB para a capitalização de mercado das empresas nacionais listadas.

88. *Merchandise Trade (% Of GDP)*: É a soma das exportações de mercadorias e das importações divididas pelo valor do PIB, todos em dólares correntes.

89. *Military Expenditure (% Of Gdp)*: Todas as despesas atuais e de capital nas forças armadas, incluindo forças de paz; ministérios da defesa e outras agências governamentais engajadas em projetos de defesa; forças paramilitares; e atividades do espaço militar.

90. *Ores And Metals Exports (% Of Merchandise Exports)*: Exportações de orés e metais frente a quantidade de mercadorias exportadas.

91. *Ores And Metals Imports (% Of Merchandise Imports)*: Importações de orés e metais frente a quantidade de mercadorias importadas.

92. *Overall Level Of Statistical Capacity (Scale 0 - 100)*: Indicador de capacidade estatística que avalia a capacidade do sistema estatístico do país. Usa a escala de 0 a 100.

93. *Periodicity And Timeliness Assessment Of Statistical Capacity (Scale 0 - 100)*: Indicador de periodicidade e oportunidade estatísticas. Avalia a disponibilidade e periodicidade dos principais indicadores socioeconômicos. Usa a escala de 0 a 100.

94. *Personal Remittances, Received (% Of GDP)*: Compreendem as tranferências pessoais e compensação de empregados.

95. *Population Growth (Annual %)*: Taxa anual de crescimento da população. É a taxa externa de crescimento populacional do ano anterior para o ano corrente.

96. *Population In The Largest City (% Of Urban Population)*: É a porcentagem da população urbana de um país que vive na maior área metropolitana do país.

97. *Private Credit Bureau Coverage (% Of Adults)*: Informa o número de indivíduos ou empresas listados por uma agência de crédito privada com informações atualizadas sobre o histórico de reembolso, dívidas não pagas ou crédito pendente. O número é expresso como uma porcentagem da população adulta.

98. *Proportion Of Seats Held By Women In National Parliaments (%)*: Porcentagem de mulheres ocupantes de cargos parlamentares.

99. *Public Credit Registry Coverage (% Of Adults)*: É o número de indivíduos e empresas em um registro de crédito público com informações atuais sobre histórico de reembolso de dívidas não pagas ou crédito excepcional. O número é expresso como porcentagem da população adulta.

100. *Real Interest Rate (%)*: Taxa de juros de crédito ajustada para a inflação medida pelo deflator do PIB.

101. *Risk Premium On Lending (Lending Rate Minus Treasury Bill Rate, %)*: Taxa de juros carregada por banco de empréstimos a clientes privados do setor menos a taxa de juros "livre de risco"em que os títulos de curto prazo do governo são emitidos ou negociados no mercado.

102. *Rural Population (% Of Total Population)*:Quantidade de pessoas que vivem nas zonas rurais. Calculada pela diferença entre a população total e a população urbana.

103. *Rural Population Growth (Annual %)*: Crescimento anual da quantidade de pessoas que vivem nas zonas rurais. Calculada pela diferença entre a população total e a população urbana.

104. *S&P Global Equity Indices (Annual % Change)*: Medem a mudança de preços no mercado de valores mobiliários.

105. *Services, Etc., Value Added (% Of GDP)*: Saída líquida do setor de serviços após adicionar todas as saídas e entradas intermediárias.

106. *Services, Etc., Value Added (Annual % Growth)*: Crescimento anual da saída líquida do setor de serviços após adicionar todas as saídas e entradas intermediárias.

107. *Source Data Assessment Of Statistical Capacity (Scale 0 - 100)*: Avaliação dos dados fonte da capacidade estatística do país. Os dados são mostrados na escala de 0 a 100.

108. *Stocks Traded, Total Value (% Of Gdp)*: Número total de ações negociadas, ambas domésticas e estrangeiras, multiplicadas por seus preços correspondentes de correspondência.

109. *Stocks Traded, Turnover Ratio Of Domestic Shares (%)*: Valor das ações domesticas negociadas e divididas por sua capitalização de mercado. O valor é anualizado multiplicando pela média mensal por 12.

110. *Tax Payments (Number)*: O número total de impostos pagos pelas empresas, incluindo arquivos eletrônicos.

111. *Time Required To Build A Warehouse (Days)*: Tempo necessário para construir um armazém. Contado em números de dias necessários para completar o procedimento necessário para a construção de um armazém.

112. *Time Required To Enforce A Contract (Days)*: Tempo necessário para concluir um contrato. Contado em número de dias a partir da apresentação do processo de tribunal até a determinação final, em casos apropriados, o pagamento.

113. *Time Required To Register Property (Days)*: Tempo necessário para a inscrição imobiliária. Contado em número de dias necessários para a empresa garantir o direitos à propriedade.

114. *Time Required To Start A Business (Days)*: Tempo necessário para começar um negocio. Contado em número de dias necessários para completar o procedimento para operar legalmente um negócio.

115. *Time To Prepare And Pay Taxes (Hours)*: Temppo para preparar e pagar impostos em tempo. Contado em horas por ano necessários para arquivar e pagar (ou retirar) trÊs tipos principais de impostos: o imposto de renda corporativo, o imposto de valor acrescentado ou de vendas e imposto de trabalho, incluindo taxas de pagamentos e contribuições de segurança social.

116. *Total Tax Rate (% Of Commercial Profits)*: Taxa total tributária. Mede a quantidade de impostos e contribuições obrigatórias pagáveis pelas empresas após a contabilidade por deduções e isenções permitidas como ação de lucros comerciais.

117. *Trade (% Of GDP)*: Soma das exportações e das importações de bens e serviços medidos como ação de Produto Interno Bruto(PIB).

118. *Trade In Services (% Of GDP)*: É a soma das exportações de serviços e das importações divididas pelo valor do PIB, todos em dólares atuais.

119. *Transport Services (% Of Commercial Service Exports)*: Cobre todos os servios de transporte realizados por residentes de uma economia para os de outa e envolvendo transporte de passageiros, aluguel de transportes com equipamento, e assistência relacionada e serviços auxiliares.

120. *Transport Services (% Of Commercial Service Imports)*: Cobre todos os servios de transporte realizados por residentes de uma economia para os de outa e envolvendo transporte de passageiros, aluguel de transportes com equipamento, e assistência relacionada e serviços auxiliares.

121. *Travel Services (% Of Commercial Service Exports)*: Cobre os serviços de mercadorias e serviços adquiridos de uma economia por viajantes para sua utilização própria durante viagens de menos de um ano para negócios ou propriedades pessoais.

122. *Travel Services (% Of Commercial Service Imports)*: Cobre os serviços de mercadorias e serviços adquiridos de uma economia por viajantes para sua utilização própria durante viagens de menos de um ano para negócios ou propriedades pessoais.

123. *Urban Population (% Of Total)*: Quantideade da população que vive em zonas urbanas definidas por escritórios nacionais de estatística.

124. *Urban Population Growth (Annual %)*: Crescimento anual da quantidade da população que vive em zonas urbanas definidas por escritórios nacionais de estatística.

125. *Wholesale Price Index (2010 = 100)*: Refere a mistura de produtos agrícolas e industriais em várias etapas de produção e distribuição, incluindo direitos de importação.

3.4 Usando os dados na plataforma Weka

A ferramenta Weka, como detalhado na Seção 2.7, utiliza preferencialmente o formato texto com extensão *.arff*. Neste trabalho, utiliza-se uma função disponível na ferramenta que possibilita a realização da mineração obtendo os dados diretamente das tabelas do DW.

3.4.1 Configurando conexão da Weka com o MySQL

Para que a conexão da Weka fosse bem sucedida, tornou-se necessário percorrer as seguintes etapas:

- Baixar o driver JDBC do MySQL e inserir na mesma pasta que se encontra o executável da weka.
- Em seguida deve-se configurar o arquivo "DatabaseUtils.props", que encontra-se dentro do executável do weka, e colocá-lo na mesma pasta que se encontra os arquivos citados acima.
- A Figura 3.8 apresenta as linhas que devem ser modificadas no arquivo "DatabaseUtils.props".
- Após a configuração, é criado um arquivo "exec.bat"contendo a seguinte linha de comando:

  java -cp mysql-connector-java-5.1.22-bin.jar;weka.jar weka.gui.GUIChooser

  Na linha de comando, o parâmetro -cp (ou –classpath) indica ao Java quais são as pastas onde ele deve procurar pelas bibliotecas necessárias para a execução do

programa. Após o –cp, especificou-se a lista de diretórios ou arquivos JAR separados por ";"(ponto-e-vírgula).

- Após esta configuração a Weka está pronto pra estabelecer conexão com o MySQL.

```
11  # JDBC driver (comma-separated list)
12  jdbcDriver=org.gjt.mm.mysql.Driver
13
14  # database URL
15  jdbcURL=jdbc:mysql://localhost:3306/datamining
16
```

Figura 3.8: Arquivo de configuração da Weka com o MySQL.

3.5 Carga dos dados e análise inicial

Ao configurar a Weka para leitura direta do banco, ganhou-se a liberdade de trazer as linhas e colunas que forem importantes para análise e relacionar uma tabela com outra, tudo isso através de *selects*. Todos os *selects* utilizados nesse trabalho estão disponíveis para consulta no apêndice. Com a *Query* A.1 foram selecionados os anos, nomes dos países e os indicadores necessários para a primeira análise.

No primeiro momento, na aba *Process*, é necessário clicar no botão *Open DB...*, abrindo a janela *SQL-Viewer* configura-se a URL=jdbc:mysql://localhost:3306/datamining, que é o local onde se encontra o banco desse trabalho. Em seguida é configurado o usuário e a senha do banco, clicando no primeiro botão após o espaço da URL. E clicando no segundo botão, se os dados estiverem corretos, estabelecemos conexão com o banco. No espaço em branco logo abaixo colocou-se a *Query* A.1. O resultado desse *select* foram 130 atributos e 330 instâncias.

No quadro inferior direito dessa tela, é possível selecionar um classificador, onde a partir deste, pode ser feita uma análise comparativa. Selecionando o atributo "*CountryName*"como classificador e clicando no botão para visualizar o comparativo de todos os atributos, conseguimos chegar a algumas conclusões apenas observando essa tela. Na Figura 3.9 temos a imagem dessa tela de visualização.

Na aba *Classify* foi feita a primeira classificação. Como primeiro passo é preciso clicar no botão *Choose* no canto superior esquerdo para escolher o classificador desejado. Após essa ação, foi aberta uma tela com várias pastas, onde existem vários tipos de algoritmos. Para essa primeira classificação foi utilizado o algoritmo de arvore de decisão J48, que é encontrado na pasta *trees*. A Figura 3.10 mostra a lista de classificadores.

Após isso, é necessário escolher o atributo classificador e então apertar o botão Start. Porém, dependendo do tamanho dos dados, do classificador escolhido e do algoritmo, essa execução pode demorar algum tempo.

As Figuras 3.11, 3.12, 3.13, 3.14, 3.15 e 3.16 mostram a saída da mineração com o algoritmo J48. Na primeira parte da saída, é mostrado um resumo dos dados de entrada e a opção de teste utilizada, que nesse caso foi *10-fold cross-validation*. Na *cross-validation*,

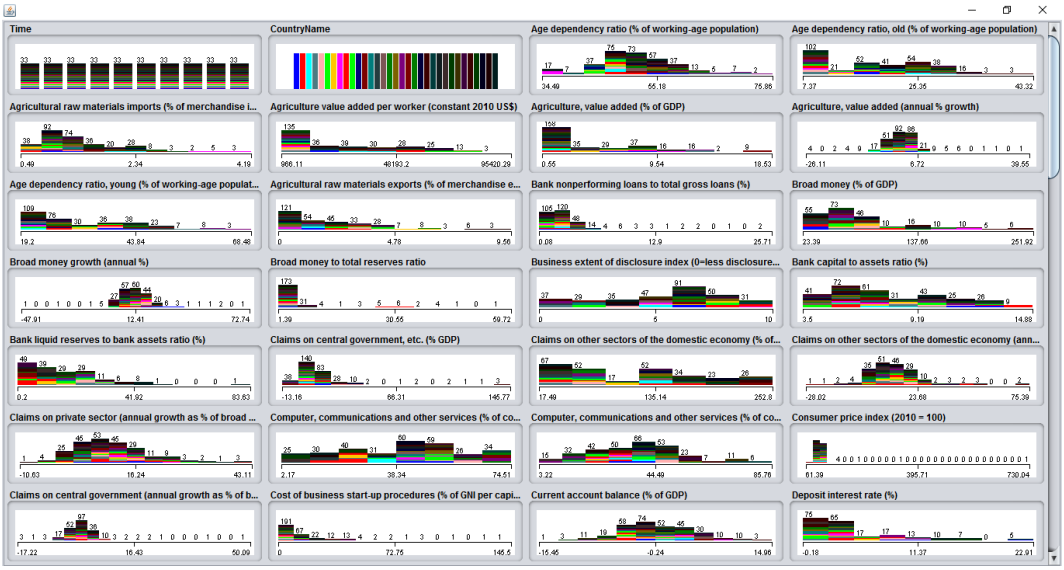


Figura 3.9: Tela de análise visual dos dados.

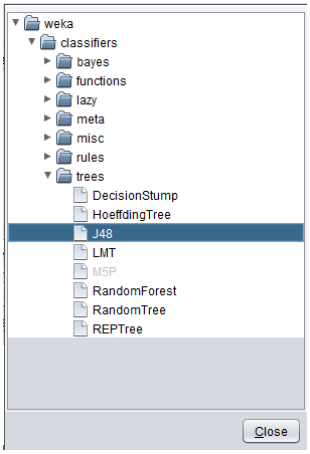


Figura 3.10: Lista de classificadores.

os dados são divididos em 10 partes, em cada um dos 10 passos, uma é separada para teste e as outras nove para treinamento e ao final de cada um deles é calculada a taxa de erro[15]. Sendo assim, a aprendizagem é executada dez vezes em diferentes conjuntos de dados obtendo dez estimativas de erros e gera um valor médio para esse resultado final.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult
Instances:    330
Attributes:   130
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

Figura 3.11: Parte 1 da saída da árvore de decisões.

Na segunda e terceira parte é mostrada a arvore de decisões gerada. É possível ver o indicador *Rural population (% of total population)* como o nó principal, ou raiz, da árvore seguido de outros 33 indicadores. Quando ao final da linha na qual encontra-se o atributo, existir o símbolo referente a dois pontos, significa que este representa uma folha da árvore, seguido do número de instâncias que utilizaram este mesmo caminho para chegar a esta folha, este número pode ser fracionário pois representa a média das 10 execuções. Logo abaixo da árvore encontramos seu tamanho, número de folhas e número total de folhas.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Rural population (% of total population) <= 19.901
|   Time required to build a warehouse (days) <= 197
|   |   Time required to build a warehouse (days) <= 96
|   |   |   Time required to build a warehouse (days) <= 28: KOREA (10.0)
|   |   |   Time required to build a warehouse (days) > 28
|   |   |   |   Time required to start a business (days) <= 10
|   |   |   |   |   Cost of business start-up procedures (% of GNI per capita) <= 0.3: DENMARK (10.0)
|   |   |   |   |   Cost of business start-up procedures (% of GNI per capita) > 0.3: UNITED STATES (10.0)
|   |   |   |   Time required to start a business (days) > 10: FINLAND (10.0)
|   |   Time required to build a warehouse (days) > 96
|   |   |   Time to prepare and pay taxes (hours) <= 132
|   |   |   |   Tax payments (number) <= 8.5: SWEDEN (10.0)
|   |   |   |   Tax payments (number) > 8.5: AUSTRALIA (10.0)
|   |   |   Time to prepare and pay taxes (hours) > 132
|   |   |   |   Time required to enforce a contract (days) <= 420: JAPAN (10.0)
|   |   |   |   Time required to enforce a contract (days) > 420: CHILE (10.0)
|   Time required to build a warehouse (days) > 197
|   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 6
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 3
|   |   |   |   Time required to build a warehouse (days) <= 308: URUGUAY (10.0)
|   |   |   |   Time required to build a warehouse (days) > 308: VENEZUELA (10.0)
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 3: BRAZIL (10.0)
|   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 6
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 7.4
|   |   |   |   Time required to build a warehouse (days) <= 274: ISRAEL (10.0)
|   |   |   |   Time required to build a warehouse (days) > 274: ARGENTINA (10.0)
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 7.4: CANADA (10.0)
Rural population (% of total population) > 19.901
```

Figura 3.12: Parte 2 da saída da árvore de decisões. Primeira metade da arvore.

```
Rural population (% of total population) > 19.901
|   Age dependency ratio, old (% of working-age population) <= 14.1307
|   |   Time required to enforce a contract (days) <= 588
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 3.1: ECUADOR (10.0)
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 3.1
|   |   |   |   Time required to register property (days) <= 9.5: TURKEY (10.0)
|   |   |   |   Time required to register property (days) > 9.5
|   |   |   |   |   Time required to register property (days) <= 39: CHINA (10.0)
|   |   |   |   |   Time required to register property (days) > 39: MEXICO (10.0)
|   |   Time required to enforce a contract (days) > 588
|   |   |   Total tax rate (% of commercial profits) <= 65.2
|   |   |   |   Time required to register property (days) <= 21: COSTA RICA (10.0)
|   |   |   |   Time required to register property (days) > 21
|   |   |   |   |   Tax payments (number) <= 34: INDIA (10.0)
|   |   |   |   |   Tax payments (number) > 34: HONDURAS (10.0)
|   |   |   Total tax rate (% of commercial profits) > 65.2
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 5: BOLIVIA (10.0)
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 5: COLOMBIA (10.0)
|   Age dependency ratio, old (% of working-age population) > 14.1307
|   |   Population in the largest city (% of urban population) <= 17.66794
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 3.1: CZECH REPUBLIC (10.0)
|   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 3.1
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 6
|   |   |   |   |   Time required to register property (days) <= 32: SPAIN (10.0)
|   |   |   |   |   Time required to register property (days) > 32: GERMANY (10.0)
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 6: ITALY (10.0)
|   |   Population in the largest city (% of urban population) > 17.66794
|   |   |   Population in the largest city (% of urban population) <= 27.53669
|   |   |   |   Proportion of seats held by women in national parliaments (%) <= 33.3
|   |   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 3.1: SWITZERLAND (10.0)
|   |   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 3.1: FRANCE (10.0)
|   |   |   |   Proportion of seats held by women in national parliaments (%) > 33.3: CUBA (10.0)
|   |   |   Population in the largest city (% of urban population) > 27.53669
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 5.5: AUSTRIA (10.0)
|   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 5.5
|   |   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) <= 7.4: PORTUGAL (10.0)
|   |   |   |   |   Business extent of disclosure index (0=less disclosure to 10=more disclosure) > 7.4: IRELAND (10.0)

Number of Leaves  :     33

Size of the tree :      65
```

Figura 3.13: Parte 3 da saída da árvore de decisões. Segunda metade da arvore.

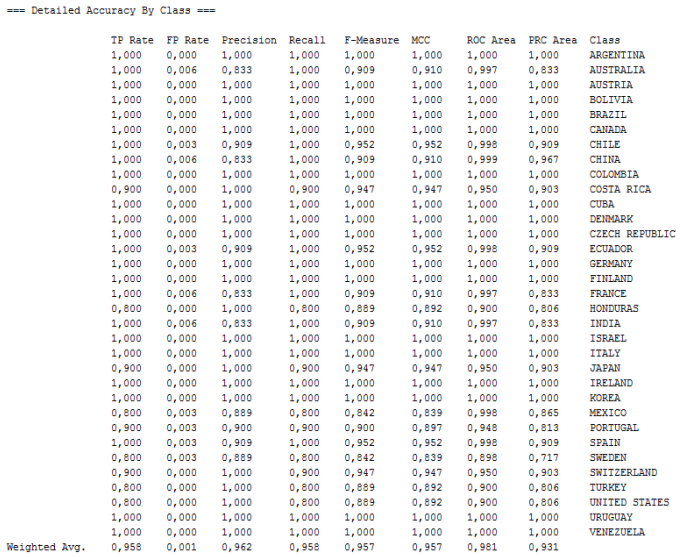
Na parte 4 encontramos a precisão da performance do algoritmo utilizando determinado tipo de teste. Neste caso aproximadamente 4% das instâncias foram classificadas incorretamente. E na Figura 3.16 temos a matriz de confusão, que mostra o total percentual de acerto do mapeamento e por classes, podendo-se identificar confusões entre as classes,como por exemplo, que uma instância da classe Costa Rica foi marcada como Portugal.

```
Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         316               95.7576 %
Incorrectly Classified Instances        14                4.2424 %
Kappa statistic                          0.9562
Mean absolute error                      0.0026
Root mean squared error                  0.0501
Relative absolute error                  4.375  %
Root relative squared error             29.1976 %
Total Number of Instances              330
```

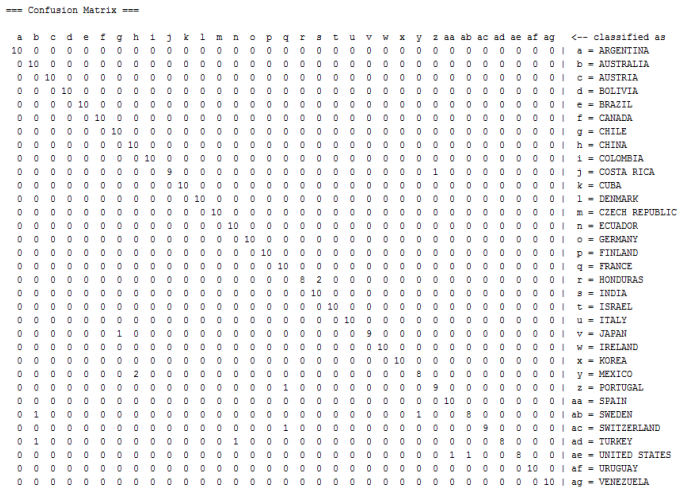
Figura 3.14: Parte 4 da saída da árvore de decisões.



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     ARGENTINA
               1,000    0,006    0,833      1,000    0,909      0,910    0,997     0,833     AUSTRALIA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     AUSTRIA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     BOLIVIA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     BRAZIL
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     CANADA
               1,000    0,003    0,909      1,000    0,952      0,952    0,998     0,909     CHILE
               1,000    0,006    0,833      1,000    0,909      0,910    0,999     0,967     CHINA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     COLOMBIA
               0,900    0,000    1,000      0,900    0,947      0,947    0,950     0,903     COSTA RICA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     CUBA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     DENMARK
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     CZECH REPUBLIC
               1,000    0,003    0,909      1,000    0,952      0,952    0,998     0,909     ECUADOR
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     GERMANY
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     FINLAND
               1,000    0,006    0,833      1,000    0,909      0,910    0,997     0,833     FRANCE
               0,800    0,000    1,000      0,800    0,889      0,892    0,900     0,806     HONDURAS
               1,000    0,006    0,833      1,000    0,909      0,910    0,997     0,833     INDIA
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     ISRAEL
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     ITALY
               0,900    0,000    1,000      0,900    0,947      0,947    0,950     0,903     JAPAN
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     IRELAND
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     KOREA
               0,800    0,003    0,889      0,800    0,842      0,839    0,998     0,865     MEXICO
               0,900    0,003    0,900      0,900    0,900      0,897    0,948     0,813     PORTUGAL
               1,000    0,003    0,909      1,000    0,952      0,952    0,998     0,909     SPAIN
               0,800    0,003    0,889      0,800    0,842      0,839    0,898     0,717     SWEDEN
               0,900    0,000    1,000      0,900    0,947      0,947    0,950     0,903     SWITZERLAND
               0,800    0,000    1,000      0,800    0,889      0,892    0,900     0,806     TURKEY
               0,800    0,000    1,000      0,800    0,889      0,892    0,900     0,806     UNITED STATES
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     URUGUAY
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     VENEZUELA
Weighted Avg.  0,958    0,001    0,962      0,958    0,957      0,957    0,981     0,931
```

Figura 3.15: Parte 5 da saída da árvore de decisões.



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y  z aa ab ac ad ae af ag   <-- classified as
 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | a = ARGENTINA
  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | b = AUSTRALIA
  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | c = AUSTRIA
  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | d = BOLIVIA
  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | e = BRAZIL
  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | f = CANADA
  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | g = CHILE
  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | h = CHINA
  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | i = COLOMBIA
  0  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0 | j = COSTA RICA
  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | k = CUBA
  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | l = DENMARK
  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | m = CZECH REPUBLIC
  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | n = ECUADOR
  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | o = GERMANY
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | p = FINLAND
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | q = FRANCE
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | r = HONDURAS
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | s = INDIA
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0 | t = ISRAEL
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0 | u = ITALY
  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0  0  0 | v = JAPAN
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0 | w = IRELAND
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0 | x = KOREA
  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  0  0  0  0  0  0  0  0 | y = MEXICO
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0 | z = PORTUGAL
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0 | aa = SPAIN
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  8  0  0  0  0  0 | ab = SWEDEN
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0 | ac = SWITZERLAND
  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  0  0  0 | ad = TURKEY
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  8  0  0 | ae = UNITED STATES
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0 | af = URUGUAY
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10 | ag = VENEZUELA
```

Figura 3.16: Parte 6 da saída da árvore de decisões.

Podemos também visualizar na Figura 3.14 a estatística Kappa, o erro médio absoluto e a raiz do erro médio quadrado das estimativas encontradas. A estatística Kappa, segundo Witten e Frank (2005), representa a porcentagem de sucesso de predição a partir das instâncias corretamente classificadas na matriz de confusão. O erro médio é a média sobre a amostra de verificação dos valores absolutos das diferenças entre a previsão e a observação correspondente. E a raiz do erro médio quadrado é calculado tirando a raiz quadrada do resultado do erro médio ao quadrado.

Na Figura 3.17 observa-se parte da arvore resultante de forma visual, que pode ser gerada a partir da lista de resultados no canto inferior esquerdo da tela, bastando clicar com o botão direito e selecionar *Visualize tree*. O que é visto no formato de elipse são os nós, que são as decisões, e os em formato retangular são as folhas da árvore, que são as classes atribuídas, como já detalhados nas Figuras 3.12 e 3.13.Podem ser geradas árvores

muito grandes o que dificulta a sua visualização por completo, sendo necessário consultar o *log* de saída para uma melhor análise.

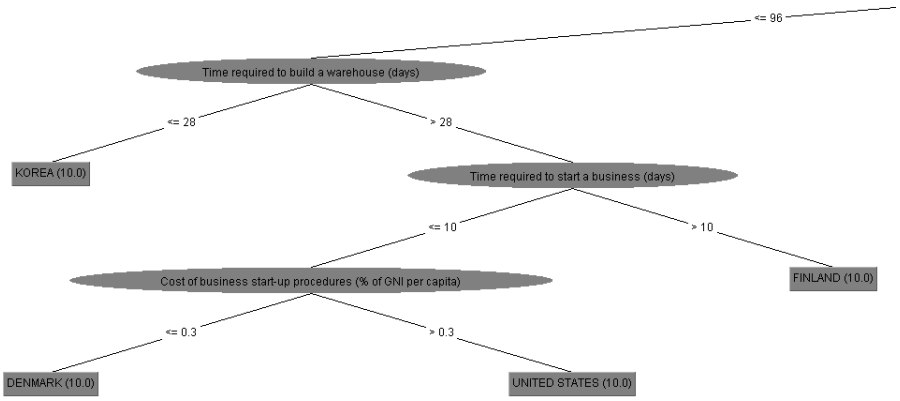


Figura 3.17: Arvore de decisões.

3.5.1 Análise dos Países como Classes

Nesta primeira análise buscou-se mostrar que não somente é possível identificar os países segundo os seus indicadores, como também obter essa informação com bastante confiança.

Notou-se que o algoritmo J48 destacou, durante sua execução, 13 indicadores dos 125 usados. Estes foram selecionados para compor os nós da arvore baseado no seu ganho de informação [5]. Para obter o ganho o algoritmo usa a entropia. A entropia é a medida de incerteza associada a um atributo e ela pode ser calculada da seguinte maneira, para todo p [25]:

A quantidade de informação que p tem a oferecer sobre a conclusão p_j:

$$\boxed{Entropia(p) = -\sum_{j=1}^{n} \frac{|p_j|}{|p|} \log \frac{|p_j|}{|p|}}$$

onde p é a classe e p_j é o nó.

A entropia condicional é:

$$Entropia(j|p) = \frac{|p_j|}{|p|} \log \frac{|p_j|}{|p|}$$

Com a entropia conseguimos o ganho desejado:

$$\boxed{Ganho(p, j) = Entropia(p) - Entropia(j|p)}$$

Após calcular o ganho de cada atributo, o algoritmo escolhe o com maior ganho e coloca como raiz, em ordem decrescente ele vai colocando os nós até chegar nas folhas da árvore. O algoritmo, segundo Hunt, Marin e Stone (1966), usa a abordagem de divisão e conquista o que diminui os tamanhos das árvores. Sendo assim, pode-se afirmar que o algoritmo J48 escolheu os 13 indicadores com os maiores ganhos.

3.6 Clusterizando os Dados

Para uma análise mais profunda, primeiramente, foi utilizada a técnica de clusterização, como explicado na Seção 2.6, separando em 3 *clusters*, essa quantidade foi escolhida pois verificou-se que com uma quantidade maior, alguns *clusters* apareciam com poucos elementos. Na aba "*Cluster*"clicou-se no botão *Choose* e escolheu o algoritmo *SimpleK-Means*. Após ser selecionado mudou-se o número de *cluster* para 3, ao clicar no nome do algoritmo onde se abrirá a janela de configurações.

O algoritmo *SimpleKMeans* realisa basicamente 3 passos:

1. Escolhe o centroide de cada *cluster*.
2. Determina a distância de cada objeto até o centroide.
3. Agrupa os objetos baseado na menor distância até o centroide.

Escolhido a quantidade de *cluster* foi executado o algoritmo com configuração padrão. A Figura 3.18 mostra o resultado dessa clusterização. Baseado em todos os indicadores, o algoritmo escolheu 178 instâncias para o primeiro *cluster*, 131 para o segundo e 21 para o terceiro. Buscando características de cada *cluster* separou-se os 13 indicadores usados para a classificação mostrada na seção anterior. Acrescentou-se também os indicadores sobre PIB(Produto Interno Bruto).

```
Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      178 ( 54%)
1      131 ( 40%)
2       21 (  6%)
```

Figura 3.18: Resultado da clusterização.

Attribute	***Full Data***	***Cluster 0***	***Cluster 1***	***Cluster 2***
Age dependency ratio, old (% of working-age population)	18.953	25.0275	12.1465	9.9238
Business extent of disclosure index(0=less disclosure to 10=more disclosure)	5.4025	6.0731	4.4788	5.481
Cost of business start-up procedures(% of GNI per capita)	13.4194	4.5711	25.1835	15.0333
Population in the largest city (% of urban population)	24.0973	23.4866	27.9251	5.3961
Proportion of seats held by women in national parliaments	24.5598	26.7994	22.891	15.9857
Rural population (% of total population)	24.5798	21.9147	22.9135	57.5645
Tax payments (number)	17.7328	11.7674	25.814	17.8857
Time required to build a warehouse(days)	171.4425	151.8361	202.2748	145.2952
Time required to enforce a contract(days)	564.3628	492.8376	659.7812	575.3952
Time required to register property(days)	33.7734	29.9035	40.7276	23.1952
Time required to start a business(days)	25.1309	12.9168	42.2955	21.5857
Time to prepare and pay taxes(hours)	332.9019	181.6365	555.7322	225.019
Total tax rate (% of commercial profits)	48.1528	42.8099	55.999	44.4952
GDP (current US$)	1565054059779.8079	2103297189070.3472	430463984957.9072	4080483716348.99
GDP Growth (annual %)	2.7514	1.349	3.7633	8.3261
GDP per capita (current US$)	27034.2424	43196.0022	8761.2484	4032.7643
GDP per capita growth (annual %)	1.8512	0.6863	2.553	7.3478

Tabela 3.2: Clusterização com 3 *cluster*

Pelo resultado mostrado na Tabela 3.2, observou-se algumas características:

- ***Cluster* 0**: Pelo primeiro indicador, são países com uma grande quantidade dos pessoas acima de 60 anos. Pelo segundo indicador, países com uma maior sigilo de informações financeiras. Já pelo sexto indicador, trata-se de países com pouca população rural. Os indicadores de tempo, como o *Time required to build a warehouse (days)*, mostra que são países que precisam de pouco tempo para construir uma casa ou começar uma empresa. O média do PIB é a maior dos 3 *clusters* e o crescimento, o menor. Visto isso, esperou-se este *cluster* contasse predominantemente com países desenvolvidos, como o Estados Unidos da América e países Europeus.
- ***Cluster* 1**: Pelo primeiro indicador, são países com pouca população acima dos 60 anos. Pelo segundo indicador, países com uma baixi sigilo de informações financeiras. São também países com baixa população rural, porém com um pouco a mais que os países do *cluster* 0. E tem os maiores valores para os indicadores de tempo. Notou-se que possui o PIB maior que o *cluster* 2, porém o crescimento menor. Esperou-se que este *cluster* contasse com países subdesenvolvidos.
- ***Cluster* 2**: Já o último *cluster*, apresentou a menor média de quantidade de pessoas acima de 60 anos. Uma taxa média de sigilo de informação financeiras. Possui a maior população rural entre os *clusters*. De imediato não foi encontrado um grupo de países que se encaixe nas características do *cluster*.

A Weka permite salvar como foi dividido os *clusters*, informando a que *cluster* cada instância pertence. Para isso, é preciso clicar com o botão direito no *Result list* e ao aparecer as opções, clique em *Visualize cluster assignments*. Será aberta uma janela onde deverá ser clicado em *Save*.

Abrindo, na aba *Preprocess*, o arquivo criado anteriormente, pode-se analisar o resultado clicando no botão *Edit*. E ao deslocar a barra de rolagem até o final observa-se que a última coluna estão os *clusters*, como é mostrado na Figura 3.19.

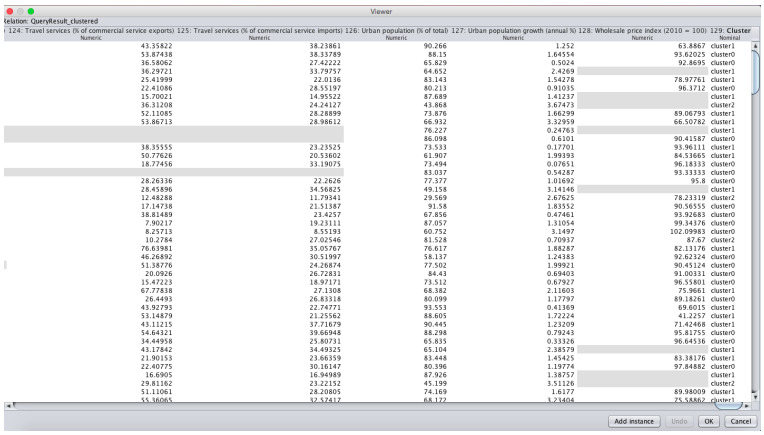


Figura 3.19: Tela de visualização do arquivo após clusterização.

Para verificar em qual *cluster* cada país ficou, pode-se ordenar os dados clicando na coluna em que se quer ordenar, no caso foi clicado primeiramente na coluna *CountryName* e depois na coluna *Cluster*. Com isso verificou-se a seguinte distribuição por *cluster*:

- ***Cluster* 0**: Austrália, Áustria, Canada, Coreia, Cuba, Dinamarca, República Tcheca, Alemanha, Finlândia, França, Israel, Itália, Japão, Irlanda, Portugal, Espanha, Suíça, Suécia e Estados Unidos da América.
- ***Cluster* 1**: Argentina, Bolívia, Brasil, Chile, Colômbia, Costa Rica, República Tcheca, Equador, Honduras, Israel, México, Turquia, Uruguai e Venezuela.
- ***Cluster* 2**: China, Índia e Coreia.

Analisando os resultados, observou-se que os países Cuba e Israel estão no *cluster* 0 e o país Israel no ano de 2010 e o país Republica Tcheca nos ano de 2006, 2007 e 2010 se encontra no *cluster* 1 também. Verificado todos os indicadores do país e a forma que o processo de clusterização selecionou essa instância para os dois *clusters*, descobriu-se que essa mudança se deu pelo variação expressiva na maioria dos indicadores que se referem a agricultura e exportações, sendo suficiente para o algoritmo seleciona-lo para o segundo *cluster*. Cuba foi selecionado para o *cluster* 0 por não existirem dados na maioria dos indicadores.

Buscando algumas características de cada *cluster*, rodou-se o classificador de arvore J48. Por existirem pouquíssimos dados de indicadores válidos para o país Cuba, decidiu-se removê-lo, visando maior confiança no resultado. Todas as instâncias do país foram removidas clicando em *Edit*, ordenando pelos nomes dos países, selecionando todas as instâncias para exclusão, clicando com o botão direito e clicando em *Delete ALL selected instance*.

Para essa análise colocou-se os *cluster* como classe visando obter padrões sobre cada cluster. A Figura 3.20 mostra o resultado do algoritmo e na Figura 3.21 mostra a árvore gerada pelo algoritmo J48. Do resultado foram tirados algumas análises:

- 43,4% das instâncias analisadas apresentam RNB per capta abaixo de $ 22.920,00. E dentre estas 43,4% nenhuma faz parte do *cluster* 0.
- Apenas 2,3% das instâncias do *cluster* 1 possuem RNB per capta acima dos $ 22.920,00.
- Olhando a matriz de confusão, observa-se que 2 instâncias pertencentes ao *cluster* 0 foram classificadas como *cluster* 1, 5 instâncias pertencentes ao *cluster* 1 foram classificadas como *cluster* 0 e uma instância pertencente ao *cluster* 2.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         312               97.5    %
Incorrectly Classified Instances         8                2.5    %
Kappa statistic                          0.954
Mean absolute error                      0.0196
Root mean squared error                  0.1231
Relative absolute error                  5.3669 %
Root relative squared error             28.842  %
Total Number of Instances              320

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,989    0,041    0,966      0,989    0,977      0,950    0,968     0,946     cluster0
               0,960    0,010    0,983      0,960    0,971      0,954    0,976     0,973     cluster1
               0,952    0,000    1,000      0,952    0,976      0,974    0,965     0,956     cluster2
Weighted Avg.  0,975    0,027    0,975      0,975    0,975      0,953    0,971     0,957

=== Confusion Matrix ===

   a   b   c   <-- classified as
 173   2   0 |   a = cluster0
   5 119   0 |   b = cluster1
   1   0  20 |   c = cluster2
```

Figura 3.20: Resultado algoritmo J48.

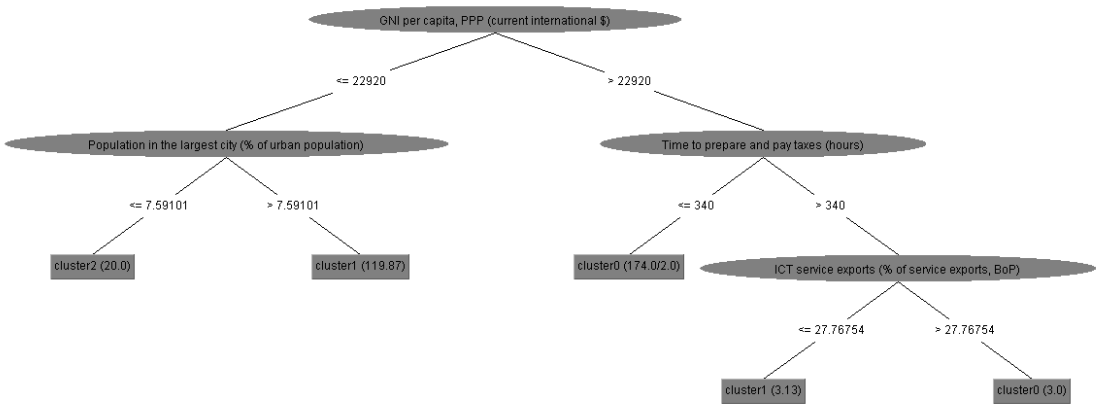


Figura 3.21: Arvore de decisões.

A Figura 3.22 mostra a representação gráfica do resultado. Podemos ver que a taxa de acerto é alta e consequentemente o erro médio é muito baixo. Sabendo que o algoritmo *SimpleKMeans* usou todos os indicadores para dividir os dados em 3 *cluster*, observou-se que com apenas 4 indicadores podemos ter 97,5% de precisão na classificação.

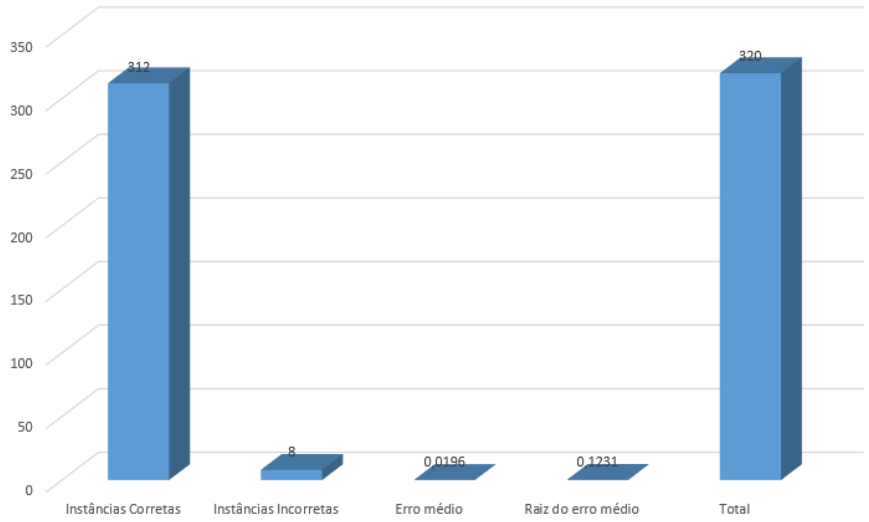


Figura 3.22: Avaliação do algoritmo J48.

3.7 Análise de Indicadores

Para a análise dos Indicadores foi utilizado a *Query* A.2, onde são buscados os mesmos indicadores da seção anterior menos os atributos com os nomes dos países e o atributo de ano. Foi realizado o mesmo processo da Seção 3.5 para fazer a carga dos dados. Buscando encontrar padrões não triviais, decidiu-se escolher alguns dos indicadores que são relevantes e não repetitivos e analisar os resultados obtidos, visto que muitos indicadores podem dar resultados ruins. Os escolhidos foram:

- GDP Per Capita (Current US$)
- GDP Per Capita Growth (Annual %)
- Current Account Balance(% OF GDP)
- Industry, Value Added (% Of GDP)
- Industry, Value Added (Annual % Growth)
- Inflation, Consumer Prices (Annual %)
- Military Expenditure (% Of Gdp)

Sabendo que todos os indicadores são numéricos e que o algoritmo de arvore J48 não classifica atributos numéricos, se fez necessário usar um filtro para transformar esses dados. O filtro que faz essa transformação é o filtro *Discretize*. Discretização de dados é um técnica que consiste em transformar valores numéricos em valores nominais ou discretos que possam representar melhor os dados em determinados conjuntos[31].

Para escolher um filtro na ferramenta *Weka*, clica-se no botão *Choose* na sessão *Filter* e seleciona *filters* > *unsupervised* > *attribute* > *Discretize*. As configurações que serão modificadas durante o processo serão basicamente a *attributeIndices*(onde coloca-se o índice do atributo que será discretizado), *bins*(cada *bin* é um intervalo, por exemplo, se os

valores reais entre 0 e 1 forem divididos em dois *bins* um *bin* pode representar o intervalo [0-0.5) e o outro o intervalo (0.5-1]) e *useEqualFrequency*(seleciona se todos os *bins* terão o mesmo número de instâncias ou não).

A *weka* gera os *bins* através de um processo chamado *binning*. Usando o conceito de vizinhança entre os dados, este processo ordena os valores dos atributos. Após a ordenação, os valores são distribuídos por grupos(*bin*). Esses grupos são divididos segundo um critério aplicado que pode ser a média aritmética, mediana ou um valor de limite. Após a divisão, os valores são substituídos pelas medidas calculadas em cada grupo.

Para garantir que novas informações fossem encontradas, foram removido da lista de atributos os outros indicadores que se assemelhavam muito com o indicador em questão. Foram disponibilizados no Apendice B todos os *logs* resultantes das classificações a seguir.

3.7.1 PIB per Capita e Crescimento do PIB per Capita

Analisou-se o indicador do PIB per capta, que refere-se a soma de todos os bens de consumo divididos pela quantidade de habitantes do país, conjuntamente com a análise do seu crescimento. Um PIB elevado e, consequentemente, o PIB per capta também elevado, são características de países desenvolvidos. Nesta análise buscou-se a padrões dos países em relação ao seu PIB per capta e o seu crescimento.

Para esses indicadores foram retirados todos os atributos que referiam ao PIB(Produto Interno Bruto), RNB(Rendimento Nacional Bruto) e também as instâncias que referiam à despesas per capta. Para a escolha dos números de *bins* e para escolher se os números de instâncias em cada *bin* será com a mesma frequência (com a mesma quantidade de instâncias em cada *bin*) ou não, foram feitos testes com o algoritmo J48 para 2, 3 e 4 *bins* e para cada um deles o uso de mesma frequência e de frequências diferentes. As Figuras 3.23 e 3.24 mostram a taxa de acerto dos resultados.

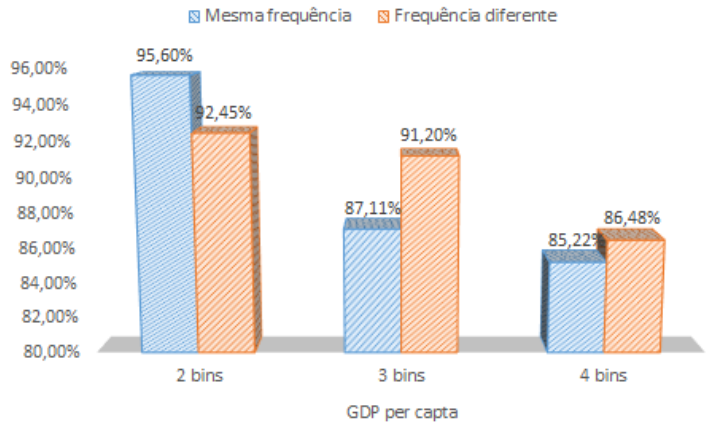


Figura 3.23: Taxa de acerto da classificação após a discretização do atributo GDP Per Capita (Current US$).

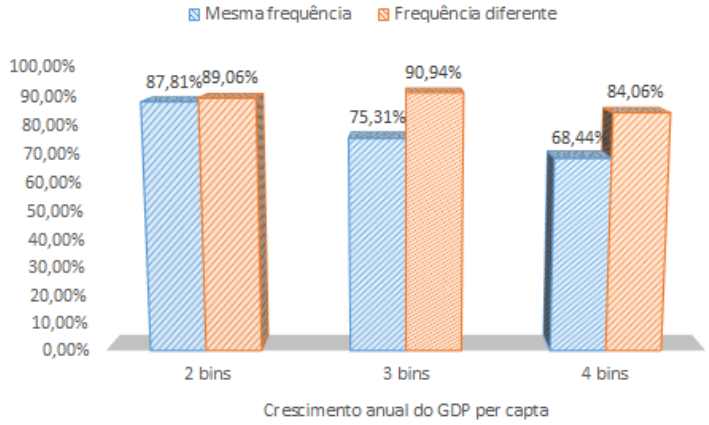


Figura 3.24: Taxa de acerto da classificação após a discretização do atributo GDP Per Capita Growth (Annual %).

Para o indicador de PIB per capta, a discretização com 2 *bins* e com a mesma frequência que obteve a melhor classificação. As instâncias no intervalo (-infinito, 22714,718895] pertencem ao primeiro *bin* e as no intervalo (22714,718895, +infinito) pertencem ao segundo *bin* e em cada um deles possuem 159 instâncias. A Figura 3.25 mostra a árvore de decisão resultante.

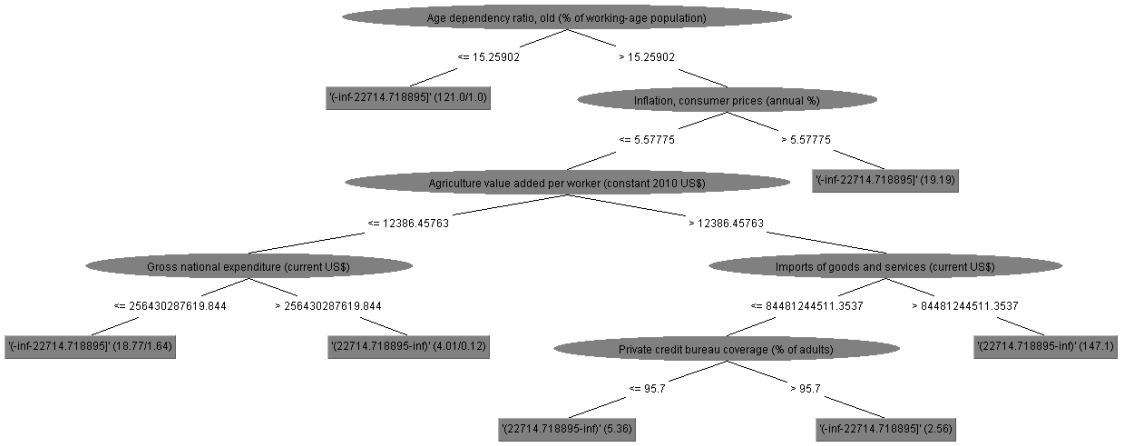


Figura 3.25: Arvore de decisões do indicador *GDP Per Capita (Current US$)*.

Para o indicador de crescimento de PIB per capta, a discretização com 3 *bins* e com frequência diferente obteve a melhor classificação. As 41 instâncias no intervalo (-infinito, -1,27109] pertencem ao primeiro *bin*, as 249 do intervalo (-1,27109, 6,16451] pertencem ao segundo *bin* e no intervalo (6,16451 , +infinito) estão as últimas 30 instâncias. A figura 3.26 mostra a árvore de decisão resultante.

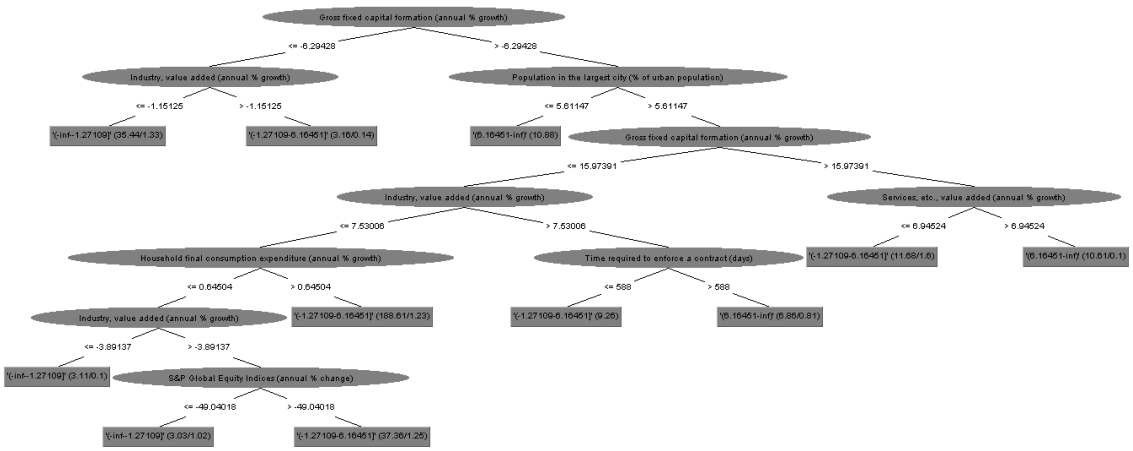


Figura 3.26: Arvore de decisões do indicador *GDP Per Capita Growth (Annual %)*.

Após a análise das árvores e dos *logs* de saída, foram destacados alguns pontos:

- Aproximadamente 76% das instâncias com o PIB per capta abaixo de US$ 22.714,72, possuem menos de 15,25% de aposentados.
- Por volta de 85% das instâncias com crescimento do PIB per capta menor que -1,27% também demonstram redução na formação de capital fixo e redução também no investimento na indústria.
- Das instâncias com o PIB per capta acima de US$ 22.714,72, aproximadamente 92% possuem mais de 15,25% de aposentado, inflação menor que 5,57% ao ano, produção agrícola maior que US$12.386,45 por trabalhador e gastam mais de US$84,4 bilhões em importações de bens de serviço.

3.7.2 Balanço da Conta Corrente Nacional

Analisou-se o balanço da conta corrente nacional, que representa a soma da balança comercial (exportações de bens e serviços menos as importações), o lucro líquido do exterior e as transferências correntes líquidas. Um saldo de conta corrente positivo indica que a nação é um bom credor para o resto do mundo, enquanto um saldo da conta corrente negativa indica que é um mutuário(quem pega epréstimos) para o resto do mundo. O *superávit* da conta corrente aumenta os ativos externos líquidos de uma nação pelo montante do excedente, e o *déficit* da conta corrente diminui esse montante.

Para esse indicador foi retirado o atributo *External Balance On Goods And Services (% Of GDP)* e *Gross savings (% of GDP)*. E para escolha do número de *bins* foi realizado o mesmo processo explicado anteriormente e o resultado mostrado na Figura 3.27.

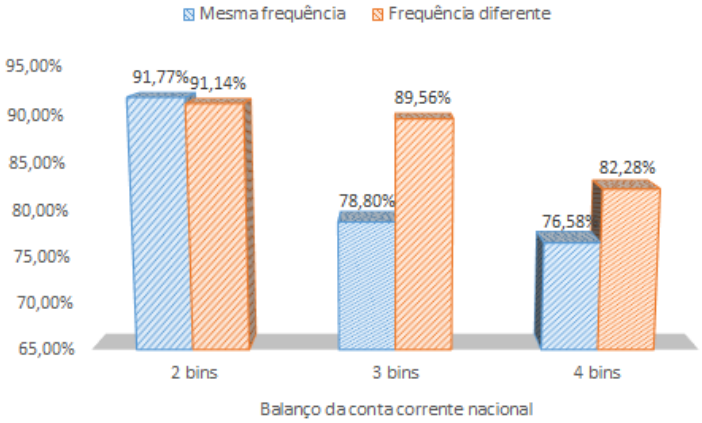


Figura 3.27: Taxa de acerto da classificação após a discretização para o atributo *Current Account Balance(% OF GDP)*.

Na Figura 3.28 pode-se ver parte da árvore resultante da classificação J48, ela pode ser melhor observada no *log* de saída que encontra-se no Apendice B.3.

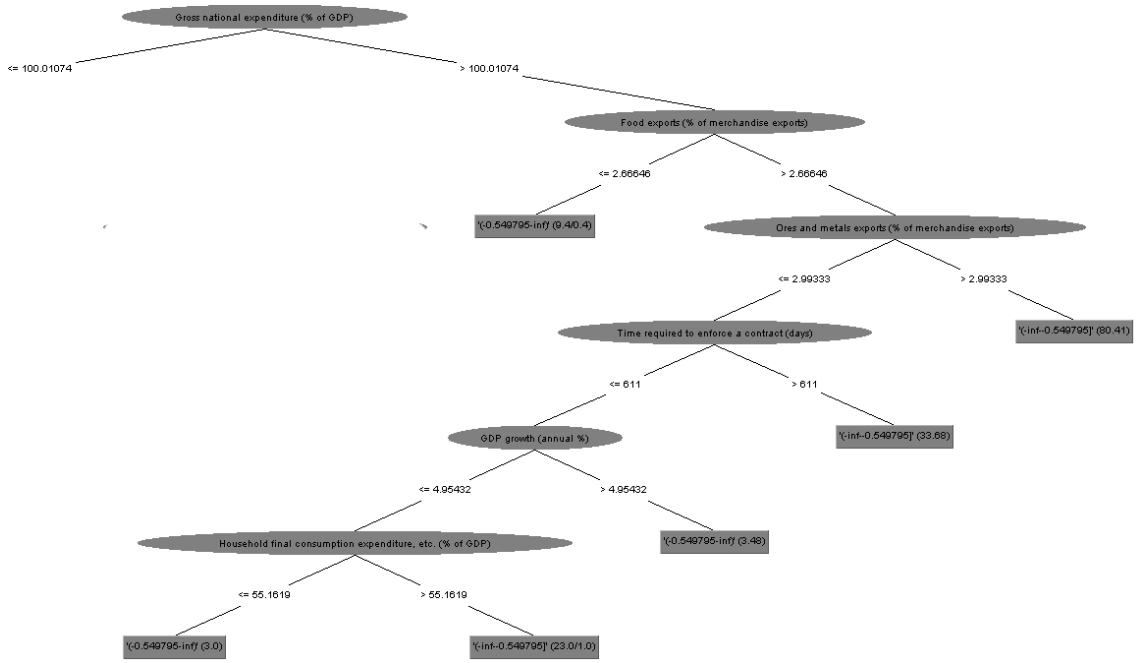


Figura 3.28: Arvore de decisões do indicador *Current Account Balance(% OF GDP)*.

Os dois *bins* foram divididos da forma que o primeiro estivesse contido no intervalo (- infinito, -0,549795] e o segundo estivesse contido no intervalo (-0,549795, +infinito) e cada um com 158 instâncias.

Após análise do *log* e da árvore, alguns pontos foram destacados:

- Aproximadamente 50% da instâncias com o balanço negativo abaixo de -0,55% do PIB a exportação de metais são superiores a 2,9% do total de exportações de mercadorias.

- Aproximadamente 20% das instâncias com abalanço negativo abaixo de -0,55% do PIB demoram mais de 611 dias para receber o resultado de um processo judiciário.
- Já as instâncias que tem melhores balanços, acima de -0,55% do PIB existem mais burocracias para a abertura de empresas o que leva a aproximadamente 68% das instâncias analisadas precisarem de mais 6,5 dias para registrar uma empresa.

3.7.3 Investimento no setor Industrial e Crescimento do Investimento no setor Industrial

Foram analisados também o investimento na industria e o crescimento desse investimento. Este indicador representa a diferença entre a produção bruta da indústria e o custo de seus insumos intermediários.

Como o investimento adicionado à determinada área, está relacionado com o PIB do país, pra o indicador investimento adicionado à indústria foram retirados todos os atributos relacionados a investimentos adicionados, como os de manufaturas e serviços, para influenciar nos resultados. A análise de classificação pode ser vista na Figura 3.29. Seguindo os passos já vistos, o melhor resultado de classificação foi para a discretização com dois *bins* de frequências diferentes. O primeiro *bin* cobrindo o intervalo de (- infinito, 37,99396] contendo 256 instâncias e o segundo *bin* cobrindo o intervalo de (37,99396, +infinito) contendo 40 instâncias. Na Figura 3.30 pode-se ver a árvore resultante da classificação J48.

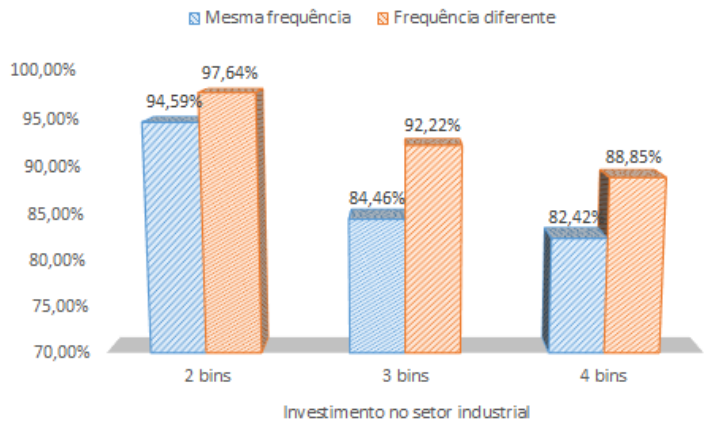


Figura 3.29: Taxa de acerto da classificação após a discretização para o atributo *Industry, Value Added (% Of GDP)*.

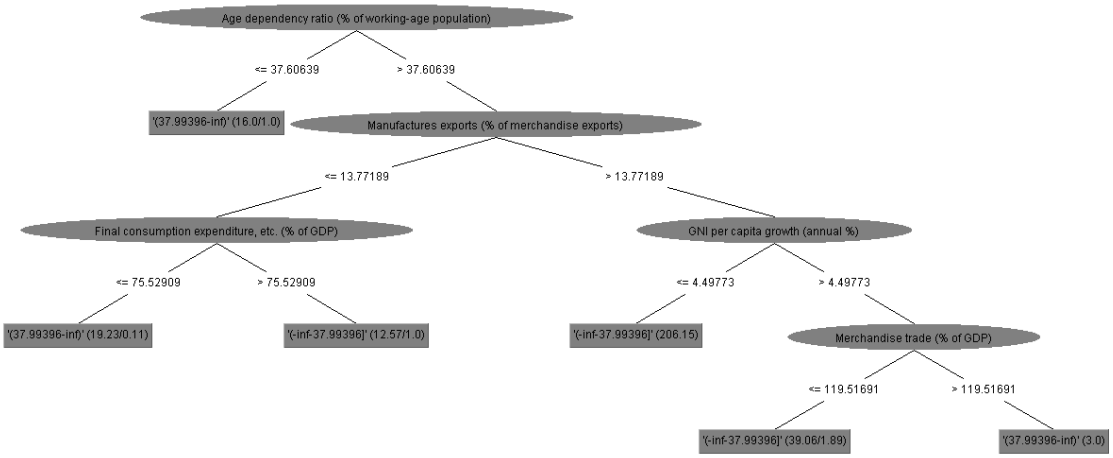


Figura 3.30: Arvore de decisões do indicador *Industry, Value Added (% Of GDP)*.

Seguindo a mesma linha de pensamento do indicador de investimento na indústria, para o indicador de crescimento de investimento adicionado à indústria foram retirados os atributos de crescimento de valores adicionados e também removemos o atributo correspondente ao crescimento do PIB. Como pode ser visto na Figura 3.31, o melhor resultado foi visto usando 2 *bins* com frequências diferentes. O primeiro *bin* cobre o intervalo de (-infinito, -1,21294] contendo 60 instâncias e o segundo *bin* cobre o intervalo de (-1,21294 , +infinito) contendo 242 instâncias. Na figura 3.32 pode-se ver a árvore resultante da classificação J48.

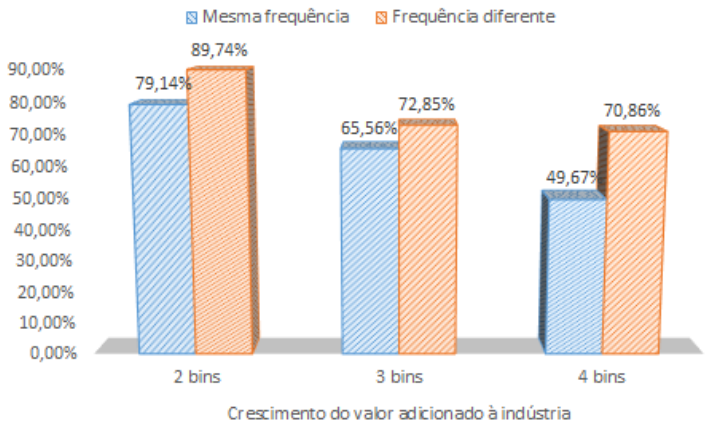


Figura 3.31: Taxa de acerto da classificação após a discretização para o atributo *Industry, Value Added (Annual % Growth)*.

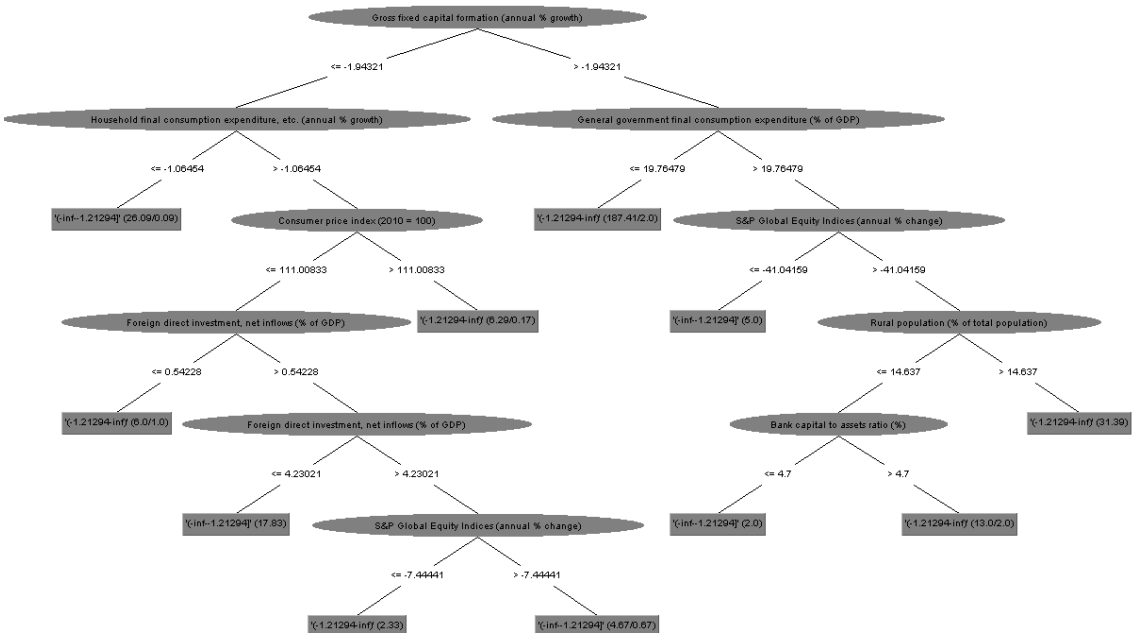


Figura 3.32: Arvore de decisões do indicador *Industry, Value Added (Annual % Growth)*.

Analisando as árvores resultantes das classificações e dos seus *logs*, destacaram-se alguns pontos:

- Das instâncias analisadas com o valor investido na industria inferior a 37,99% do PIB, aproximadamente 80% delas possui mais de 37,60% de dependentes(faixa etária de menos de 15 anos e mais de 64) por trabalhador(faixa etária entre 15 e 64 anos). Esses mesmos 80% a exportação de manufaturas representam mais de 13% da suas exportações.
- Para as instâncias com o valor de investimento na indústria superior a 37,99% do PIB, aproximadamente 40% delas possui menos de 37,60% de dependentes(faixa etária de menos de 15 anos e mais de 64) por trabalhador(faixa etária entre 15 e 64 anos).
- Aproximadamente 77% das instâncias que apresentaram crescimento de investimento na indústria superior a -1,21%, gastam menos de 19,76% do PIB com gastos governamentais.
- Aproximadamente 43% das instâncias que apresentaram deficit de crescimento de investimento na indústria superior a -1,21% também apresentaram deficit superior a -1,06% no crescimento de despesas de consumo doméstico.

3.7.4 Inflação

Analisou-se também a inflação medida pelo índice de preços ao consumidor, que reflete a variação percentual anual no custo para o consumidor médio de adquirir uma cesta de produtos e serviços.

O indicador sobre a inflação foi tratado de uma forma diferente. Para que a análise não fosse prejudicada retiramos todas as instâncias do país Venezuela por apresentar números

exagerados de inflação que acredita-se está relacionada com a crise política enfrentada no país durantes os anos analisados. Também removemos uma instância do país Irlanda que apresentou uma taxa muito baixa de inflação no ano de 2009. Decidiu-se também não remover nenhum indicador para essa análise. O melhor resultado na classificação foi com 2 *bins* e com frequências diferentes como pode ser notado na Figura 3.33. O primeiro *bin* compreende o intervalo de (-infinito, 6,326845] e o segundo compreende o intervalo (6,326845, +infinito).

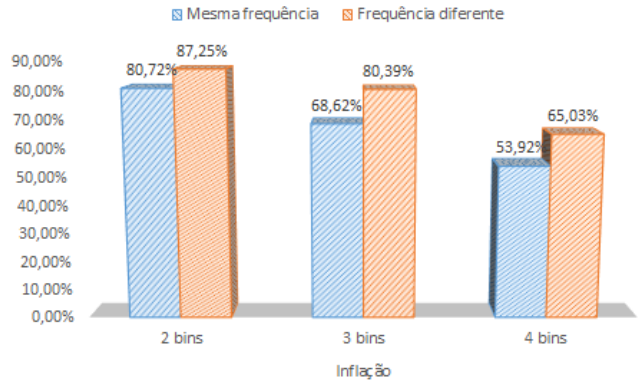


Figura 3.33: Taxa de acerto da classificação após a discretização para o atributo *Inflation, Consumer Prices (Annual %)*.

Na Figura 3.34 mostra a arvore resultante da classificação J48. No *log* que se encontra no Apêndice B.6 pode-se ter uma melhor compreensão do resultado.

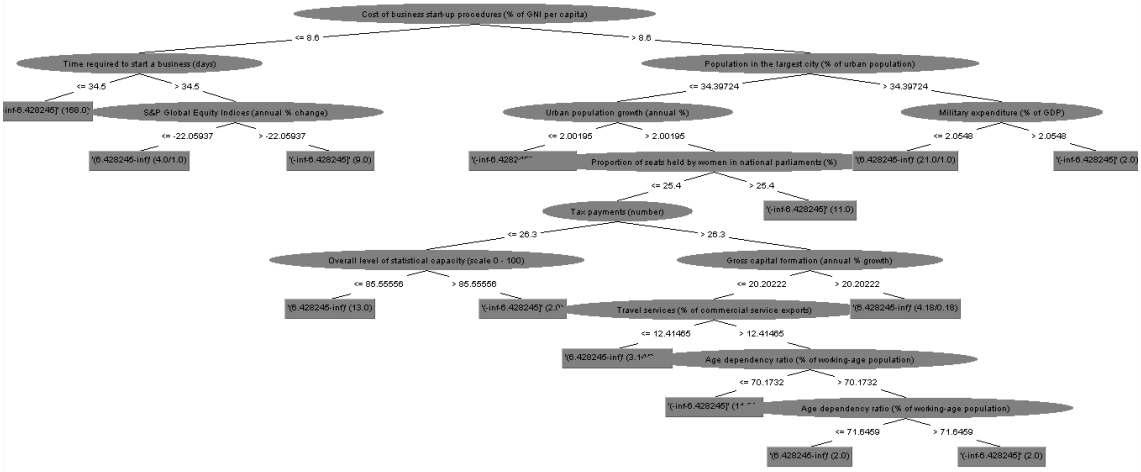


Figura 3.34: Arvore de decisões do indicador *Inflation, Consumer Prices (Annual %)*.

Através de uma analise mais aprofundada destacou-se alguns pontos:

- Das instâncias analisadas com a inflação abaixo de 6,32% aproximadamente 65% levam menos de 34,5 dias para abrir uma empresa com custo abaixo de 8,6% da RNB.

- Para as instâncias com inflação acima de 6,32%, cerca de 46% possuem mais de 34% da população nas maiores cidades e gastam menos de 2% do PIB com investimentos militares.
- Por volta de 47% das instâncias com inflação acima de 6,32% apresentam um crescimento urbano acima 2% ao ano. Dessas instâncias, cerca de 60% delas, as empresas pagam, em média, menos de 26,3 impostos diferentes.

3.7.5 Despesas militares

As despesas militares representam todos os gastos nas forças armadas, incluindo as forças de manutenção da paz; Ministérios de defesa e outras agências governamentais envolvidas em projetos de defesa; Forças paramilitares, se estes forem julgados como treinados e equipados para operações militares; E atividades espaciais militares. Essas despesas incluem pessoal militar e civil, incluindo pensões de aposentadoria de pessoal militar e serviços sociais para pessoal; operação e manutenção; Aquisição; Pesquisa e desenvolvimento militar; E ajuda militar (nas despesas militares do país doador).

Para o indicador de investimentos militares removemos todos indicadores de PIB e RNB. A Figura 3.35 mostra a comparação dos resultados das classificações. Mesmo a discretização com 2 *bins* e frequência diferente ter apresentado a melhor classificação, observou que a discretização com 3 *bins* e frequência diferente apresentaria melhores análises sem perder tanta precisão na classificação. O intervalo para o primeiro *bin* é (-infinito, 2,468583], para o segundo (2,468583, 4,937167] e o terceiro (4,937167, +infinito).

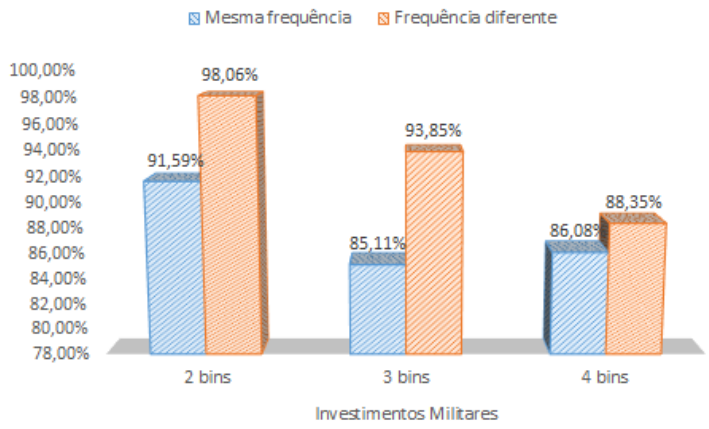


Figura 3.35: Taxa de acerto da classificação após a discretização para o atributo *Military Expenditure (% Of Gdp)*.

A Figura 3.36 mostra a arvore resultante da classificação J48. No *log* que se encontra no Apêndice B.7 pode-se ter uma melhor compreensão do resultado.

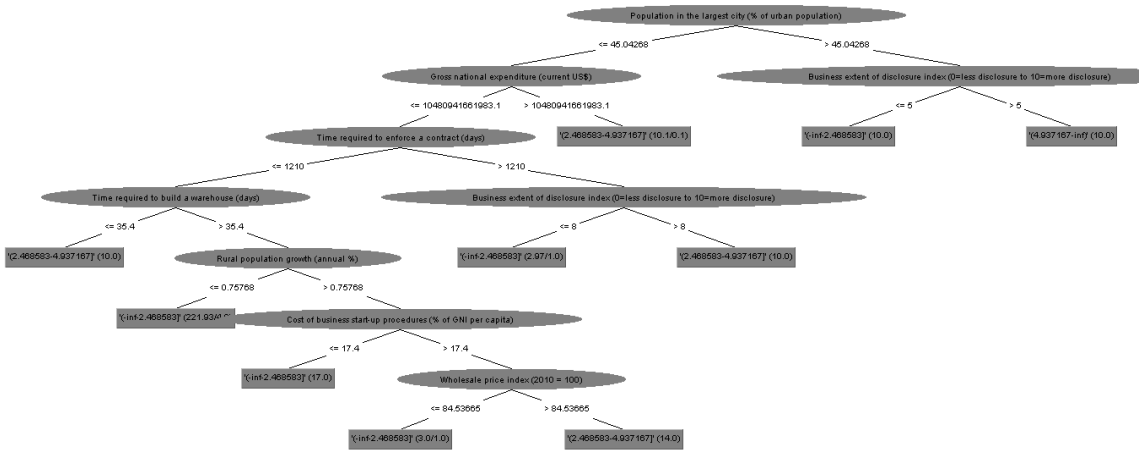


Figura 3.36: Arvore de decisões do indicador *Military Expenditure (% Of Gdp)*.

Após análise do *log* e da árvore, destacou-se alguns pontos:

- 100% das instâncias analisadas que investem mais de 4,9% do PIB na área militar apresentam o índice de proteção de informações empresariais acima de 5(o índice vai de 0 a 10). Também concentram mais de 45% da população nas maiores cidades do país.
- Aproximadamente 88% das instâncias analisadas que investem menos de 2,47% do PIB na área militar apresentam crescimento da população rural inferior a 0,75% ao ano.
- Cerca de 48% das instâncias analisadas que investem entre 2,47% e 4,9% do PIB em área militar, gastam menos de 1210 dias para finalizar um processo judicial.

Capítulo 4

Conclusão

Neste trabalho de graduação, encontrou-se padrões de forma não supervisionada nos dados extraídos do *site* do Banco Mundial, banco que é a maior fonte global de assistência para o desenvolvimento, proporcionando empréstimos e doações de cerca de US$ 60 bilhões aos 187 países-membros[2].

No primeiro momento foi estudado qual o conjunto de dados mais consistente para análise, optando-se pelo conjunto de indicadores *World Development Indicators(WDI)* por englobar a maior quantidade de informações. Segundo o Banco Mundial, esse conjunto de indicadores apresenta os dados de desenvolvimento global mais atualizados e precisos disponíveis no *site*, incluindo estimativas nacionais, regionais e globais.

Escolhido os dados, passou-se para o processo de extração, tratamento e carga destes dados, explicado na Seção 2.4. O sistema do Banco Mundial oferece a ferramenta *DataBank* que facilitou o processo de extração. No Capítulo 3 foi explicado que a ferramenta fornece a flexibilidade na extração dos dados, podendo escolher quais dados serão baixados e se eles serão linhas ou colunas no arquivo. Foram baixados todos os indicadores de vários países em um período de dez anos. A escolha de dez anos veio com intuito de cobrir possíveis mudanças e que não fosse um grande intervalo de tempo. A ferramenta também permite escolher em qual formato os dados serão baixados, foi escolhido o formato CSV por ser aceito nas ferramentas de tratamento e de mineração.

Para a transformação foi utilizada a ferramenta *Pentaho Data Integration*, explicado na Seção 2.5. Com auxílio dela, tratamos os dados excluindo o que não era necessário para a mineração e formatando os dados para a garantia da consistência. A carga foi realizada em um banco de dados *MySQL*, explicado na Seção 2.2.

Após esse intenso trabalho de tratamento e carga, os dados estavam prontos para a mineração. A ferramenta utilizada para essa mineração foi a *Weka*, detalhada na Seção 2.7. Analisando os 1446 indicadores observou-se que muitos deles estavam sem dados para a maioria dos países, por isso foi escolhido 125 com mais dados possíveis. Para a escolha dos 125 foi levado em conta também a exclusão de indicadores que falavam da mesma coisa de maneira diferente. Após a conexão do *Weka* com o banco de dados, explicado na Seção 3.5, com auxílio de *queries*, foram selecionado os 125 indicadores para a mineração.

A primeira análise teve como objetivo confirmar o quanto o valor dos indicadores estavam atrelados aos seus países. E isso foi constatado após selecionar o atributo com os nomes dos países como classe e executar o algoritmo de classificação, ficando claro

após uma classificação com precisão de 95% que pode-se identificar o país pelos seus indicadores.

Confirmada a consistência dos indicadores, abordou-se a técnica de clusterização onde foi dividido em 3 *clusters* esperando que fossem agrupados países com mesmo nível de desenvolvimento. Isso foi visto na maioria dos casos. Alguns países mudaram de *cluster* em alguns anos outro foi colocado em um *cluster* por não apresentar muitos dados na maioria dos indicadores. O resultado da classificação com os *cluster* como classe mostrou que aparentemente os países foram divididos por níveis de desenvolvimento.

Por fim, foi escolhido 7 indicadores para análise. Para que essa análise fosse possível foi necessário do uso da discretização para transformar os valores numéricos em nominais ou discretos, possibilitando a execução do algoritmo J48. Também foi preciso um estudo ainda mais profundo dos indicadores para retirar da análise outros indicadores que deixariam o resultado tendencioso evitando respostas óbvias.

Ao final de cada classificação alguns padrões não triviais foram encontrados. Como exemplo os indicadores de PIB foi descoberto a relação entre a diminuição de investimentos na indústria e o baixo crescimento do PIB. Também foi descoberto que existe uma forte relação entre a importação de bens de serviços com países com o PIB per capta elevado. Ficou visível que é possível encontrar padrões em alguns indicadores escolhidos.

4.1 Trabalhos Futuros

Como sugestão de trabalhos futuros, seria uma mineração de indicadores que refletissem não somente o PIB per capita mas também analisar o indicador GINI que mostra melhor essa divisão de renda. Poderia também agregar uma análise sociológica aprofundada nos resultados dos padrões.

Como o uso da tabela que mostrava o país por continente fugiu do nosso escopo, outra sugestão de trabalhos futuros seria o utilização da tabela para análise não somente dos países mas também dos continentes, descobrindo diferenças dentro dos próprios continentes e entre eles.

Apêndice A

Querys utilizadas

Linsting A.1: Query contendo os dados principais

```
SELECT
        `Time`
    , `CountryName`
        , `SP.POP.DPND`                  AS 'Age dependency ratio (% of
            working-age population)'
        , `SP.POP.DPND.OL`               AS 'Age dependency ratio, old (% of
            working-age population)'
        , `TM.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials imports (%
            of merchandise imports)'
        , `TX.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials export (% of
            mechandise exports)'
        , `EA.PRD.AGRI.KD`               AS 'Agriculture value added per
            worker (constant 2010 US$)'
        , `NV.AGR.TOTL.ZS`               AS 'Agriculture, value added (% of
            GDP)'
        , `NV.AGR.TOTL.KD.ZG`    AS 'Agriculture, value added (annual %
            growth)'
        , `FB.BNK.CAPA.ZS`               AS 'Bank capital to assets ratio
            (%)'
        , `FD.RES.LIQU.AS.ZS`    AS 'Bank liquid reserves to bank assets
            ratio (%)'
        , `FB.AST.NPER.ZS`               AS 'Bank nonperforming loans to
            total gross loans (%)'
        , `FM.LBL.BMNY.GD.ZS`    AS 'Broad money (% of GDP)'
        , `FM.LBL.BMNY.ZG`               AS 'Broad money growth (annual %)'
        , `FM.LBL.BMNY.IR.ZS`    AS 'Broad money to total reserves ratio'
        , `IC.BUS.DISC.XQ`               AS 'Business extent of disclosure
            index (0=less disclosure to 10=more disclosure)'
        , `FM.AST.CGOV.ZG.M3`    AS 'Claims on central government (annual
            growth as % of broad money)'
        , `FS.AST.CGOV.GD.ZS`    AS 'Claims on central government, etc. (%
            GDP)'
        , `FS.AST.DOMO.GD.ZS`    AS 'Claims on other sectors of the domestic
             economy (% of GDP)'
        , `FM.AST.DOMO.ZG.M3`    AS 'Claims on other sectors of the domestic
             economy (annual growth as % of broad money)'
        , `FM.AST.PRVT.ZG.M3`    AS 'Claims on private sector (annual growth
             as % of broad money)'
```

```
23          ,`TX.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
               services (% of commercial service exports)'
24          ,`TM.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
               services (% of commercial service imports)'
25          ,`FP.CPI.TOTL`                  AS 'Consumer price index (2010 =
               100)'
26          ,`IC.REG.COST.PC.ZS`    AS 'Cost of business start-up procedures (%
                of GNI per capita)'
27          ,`BN.CAB.XOKA.GD.ZS`    AS 'Current account balance (% of GDP)'
28          ,`FR.INR.DPST`                  AS 'Deposit interest rate (%)'
29          ,`FD.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector by
               banks (% of GDP)'
30          ,`FS.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector (% of
                GDP)'
31          ,`NE.EXP.GNFS.ZS`               AS 'Exports of goods and services
               (% of GDP)'
32          ,`NE.EXP.GNFS.KD.ZG`    AS 'Exports of goods and services (annual %
                growth)'
33          ,`NE.RSB.GNFS.ZS`               AS 'External balance on goods and
               services (% of GDP)'
34          ,`NE.CON.TETC.ZS`               AS 'Final consumption expenditure,
               etc. (% of GDP)'
35          ,`NE.CON.TETC.KD.ZG`    AS 'Final consumption expenditure, etc. (
               annual % growth)'
36          ,`TX.VAL.FOOD.ZS.UN`    AS 'Food exports (% of merchandise exports)
               '
37          ,`TM.VAL.FOOD.ZS.UN`    AS 'Food imports (% of merchandise imports)
               '
38          ,`BX.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net inflows
               (% of GDP)'
39          ,`BM.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net outflows
                (% of GDP)'
40          ,`TX.VAL.FUEL.ZS.UN`    AS 'Fuel exports (% of merchandise exports)
               '
41          ,`TM.VAL.FUEL.ZS.UN`    AS 'Fuel imports (% of merchandise imports)
               '
42          ,`NY.GDP.MKTP.CD`               AS 'GDP (current US$)'
43          ,`NY.GDP.MKTP.KD.ZG`    AS 'GDP growth (annual %)'
44          ,`NY.GDP.PCAP.CD`               AS 'GDP per capita (current US$)'
45          ,`NY.GDP.PCAP.KD.ZG`    AS 'GDP per capita growth (annual %)'
46          ,`NY.GDP.PCAP.PP.CD`    AS 'GDP per capita, PPP (current
               international $)'
47          ,`NE.CON.GOVT.ZS`               AS 'General government final
               consumption expenditure (% of GDP)'
48          ,`NE.CON.GOVT.KD.ZG`    AS 'General government final consumption
               expenditure (annual % growth)'
49          ,`NY.GNP.MKTP.CD`               AS 'GNI (current US$)'
50          ,`NY.GNP.MKTP.KD.ZG`    AS 'GNI growth (annual %)'
51          ,`NY.GNP.PCAP.KD.ZG`    AS 'GNI per capita growth (annual %)'
52          ,`NY.GNP.PCAP.PP.CD`    AS 'GNI per capita, PPP (current
               international $)'
53          ,`NE.GDI.TOTL.ZS`               AS 'Gross capital formation (% of
               GDP)'
54          ,`NE.GDI.TOTL.KD.ZG`    AS 'Gross capital formation (annual %
               growth)'
```

```
55          ,`NE.GDI.TOTL.CD`               AS 'Gross capital formation (
              current US$)'
56          ,`NY.GDS.TOTL.ZS`               AS 'Gross domestic savings (% of
              GDP)'
57          ,`NE.GDI.FTOT.ZS`               AS 'Gross fixed capital formation
              (% of GDP)'
58          ,`NE.GDI.FTOT.KD.ZG`    AS 'Gross fixed capital formation (annual %
               growth)'
59          ,`NE.GDI.FTOT.CD`               AS 'Gross fixed capital formation (
              current US$)'
60          ,`NE.GDI.FPRV.ZS`               AS 'Gross fixed capital formation,
              private sector (% of GDP)'
61          ,`NE.DAB.TOTL.ZS`               AS 'Gross national expenditure (%
              of GDP)'
62          ,`NE.DAB.TOTL.CD`               AS 'Gross national expenditure (
              current US$)'
63          ,`NY.GNS.ICTR.ZS`               AS 'Gross savings (% of GDP)'
64          ,`NY.GNS.ICTR.GN.ZS`    AS 'Gross savings (% of GNI)'
65          ,`NY.GNS.ICTR.CD`               AS 'Gross savings (current US$)'
66          ,`NY.GDP.FCST.CD`               AS 'Gross value added at factor
              cost (current US$)'
67          ,`TX.VAL.TECH.MF.ZS`    AS 'High-technology exports (% of
              manufactured exports)'
68          ,`NE.CON.PRVT.PC.KD`    AS 'Household final consumption expenditure
               per capita (constant 2010 US$)'
69          ,`NE.CON.PRVT.PC.KD.ZG` AS 'Household final consumption expenditure
               per capita growth (annual %)'
70          ,`NE.CON.PETC.ZS`               AS 'Household final consumption
              expenditure, etc. (% of GDP)'
71          ,`NE.CON.PETC.KD.ZG`    AS 'Household final consumption expenditure
              , etc. (annual % growth)'
72          ,`NE.CON.PRVT.KD.ZG`    AS 'Household final consumption expenditure
               (annual % growth)'
73          ,`BX.GSR.CCIS.ZS`               AS 'ICT service exports (% of
              service exports, BoP)'
74          ,`IQ.CPA.IRAI.XQ`               AS 'IDA resource allocation index
              (1=low to 6=high)'
75          ,`NE.IMP.GNFS.ZS`               AS 'Imports of goods and services
              (% of GDP)'
76          ,`NE.IMP.GNFS.KD.ZG`    AS 'Imports of goods and services (annual %
               growth)'
77          ,`NE.IMP.GNFS.CD`               AS 'Imports of goods and services (
              current US$)'
78          ,`NV.IND.TOTL.ZS`               AS 'Industry, value added (% of GDP
              )'
79          ,`NV.IND.TOTL.KD.ZG`    AS 'Industry, value added (annual % growth)
              '
80          ,`FP.CPI.TOTL.ZG`               AS 'Inflation, consumer prices (
              annual %)'
81          ,`TX.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
              commercial service exports)'
82          ,`TM.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
              commercial service imports)'
83          ,`FR.INR.LNDP`                  AS 'Interest rate spread (lending
              rate minus deposit rate, %)'
84          ,`FR.INR.LEND`                  AS 'Lending interest rate (%)'
```

```
85          ,`CM.MKT.LDOM.NO`                AS 'Listed domestic companies,
               total [CM.MKT.LDOM.NO]'
86          ,`TM.VAL.MANF.ZS.UN`     AS 'Manufactures imports (% of merchandise
               imports)'
87          ,`TX.VAL.MANF.ZS.UN`     AS 'Manufactures exports (% of merchandise
               exports)'
88          ,`NV.IND.MANF.ZS`                AS 'Manufacturing, value added (%
               of GDP)'
89          ,`NV.IND.MANF.KD.ZG`     AS 'Manufacturing, value added (annual %
               growth)'
90          ,`CM.MKT.LCAP.GD.ZS`     AS 'Market capitalization of listed
               domestic companies (% of GDP)'
91          ,`TG.VAL.TOTL.GD.ZS`     AS 'Merchandise trade (% of GDP)'
92          ,`MS.MIL.XPND.GD.ZS`     AS 'Military expenditure (% of GDP)'
93          ,`TX.VAL.MMTL.ZS.UN`     AS 'Ores and metals exports (% of
               merchandise exports)'
94          ,`TM.VAL.MMTL.ZS.UN`     AS 'Ores and metals imports (% of
               merchandise imports)'
95          ,`IQ.SCI.OVRL`                   AS 'Overall level of statistical
               capacity (scale 0 - 100)'
96          ,`IQ.SCI.PRDC`                   AS 'Periodicity and timeliness
               assessment of statistical capacity (scale 0 - 100)'
97          ,`BX.TRF.PWKR.DT.GD.ZS` AS 'Personal remittances, received (% of
               GDP)'
98          ,`SP.POP.GROW`                   AS 'Population growth (annual %)'
99          ,`EN.URB.LCTY.UR.ZS`     AS 'Population in the largest city (% of
               urban population)'
100         ,`IC.CRD.PRVT.ZS`                AS 'Private credit bureau coverage
               (% of adults)'
101         ,`SG.GEN.PARL.ZS`                AS 'Proportion of seats held by
               women in national parliaments (%)'
102         ,`IC.CRD.PUBL.ZS`                AS 'Public credit registry coverage
                (% of adults)'
103         ,`FR.INR.RINR`                   AS 'Real interest rate (%)'
104         ,`FR.INR.RISK`                   AS 'Risk premium on lending (
               lending rate minus treasury bill rate, %)'
105         ,`SP.RUR.TOTL.ZS`                AS 'Rural population (% of total
               population)'
106         ,`SP.RUR.TOTL.ZG`                AS 'Rural population growth (annual
                %)'
107         ,`CM.MKT.INDX.ZG`                AS 'S&P Global Equity Indices (
               annual % change)'
108         ,`NV.SRV.TETC.ZS`                AS 'Services, etc., value added (%
               of GDP)'
109         ,`NV.SRV.TETC.KD.ZG`     AS 'Services, etc., value added (annual %
               growth)'
110         ,`IQ.SCI.SRCE`                   AS 'Source data assessment of
               statistical capacity (scale 0 - 100)'
111         ,`CM.MKT.TRAD.GD.ZS`     AS 'Stocks traded, total value (% of GDP)'
112         ,`CM.MKT.TRNR`                   AS 'Stocks traded, turnover ratio
               of domestic shares (%)'
113         ,`IC.TAX.PAYM`                   AS 'Tax payments (number)'
114         ,`IC.WRH.DURS`                   AS 'Time required to build a
               warehouse (days)'
115         ,`IC.LGL.DURS`                   AS 'Time required to enforce a
               contract (days)'
```

```
116         ,`IC.PRP.DURS`                  AS 'Time required to register
                property (days)'
117         ,`IC.REG.DURS`                  AS 'Time required to start a
                business (days)'
118         ,`IC.TAX.DURS`                  AS 'Time to prepare and pay taxes (
                hours)'
119         ,`IC.TAX.TOTL.CP.ZS`    AS 'Total tax rate (% of commercial profits
                )'
120         ,`NE.TRD.GNFS.ZS`               AS 'Trade (% of GDP)'
121         ,`BG.GSR.NFSV.GD.ZS`    AS 'Trade in services (% of GDP)'
122         ,`TX.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
                service exports)'
123         ,`TM.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
                service imports)'
124         ,`TX.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
                service exports)'
125         ,`TM.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
                service imports)'
126         ,`SP.URB.TOTL.IN.ZS`    AS 'Urban population (% of total)'
127         ,`SP.URB.GROW`                  AS 'Urban population growth (annual
                %)'
128         ,`FP.WPI.TOTL`                  AS 'Wholesale price index (2010 =
                100)'
129 FROM TBL_COUNTRY
```

Linsting A.2: Query para mineração de indicadores

```
1  SELECT
2           ,`SP.POP.DPND`                  AS 'Age dependency ratio (% of
                working-age population)'
3           ,`SP.POP.DPND.OL`               AS 'Age dependency ratio, old (% of
                 working-age population)'
4           ,`TM.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials imports (%
                of merchandise imports)'
5           ,`TX.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials export (% of
                 mechandise exports)'
6           ,`EA.PRD.AGRI.KD`               AS 'Agriculture value added per
                worker (constant 2010 US$)'
7           ,`NV.AGR.TOTL.ZS`               AS 'Agriculture, value added (% of
                GDP)'
8           ,`NV.AGR.TOTL.KD.ZG`    AS 'Agriculture, value added (annual %
                growth)'
9           ,`FB.BNK.CAPA.ZS`               AS 'Bank capital to assets ratio
                (%)'
10          ,`FD.RES.LIQU.AS.ZS`    AS 'Bank liquid reserves to bank assets
                ratio (%)'
11          ,`FB.AST.NPER.ZS`               AS 'Bank nonperforming loans to
                total gross loans (%)'
12          ,`FM.LBL.BMNY.GD.ZS`    AS 'Broad money (% of GDP)'
13          ,`FM.LBL.BMNY.ZG`               AS 'Broad money growth (annual %)'
14          ,`FM.LBL.BMNY.IR.ZS`    AS 'Broad money to total reserves ratio'
15          ,`IC.BUS.DISC.XQ`               AS 'Business extent of disclosure
                index (0=less disclosure to 10=more disclosure)'
16          ,`FM.AST.CGOV.ZG.M3`    AS 'Claims on central government (annual
                growth as % of broad money)'
```

```
17          ,`FS.AST.CGOV.GD.ZS`    AS 'Claims on central government, etc. (%
              GDP)'
18          ,`FS.AST.DOMO.GD.ZS`    AS 'Claims on other sectors of the domestic
              economy (% of GDP)'
19          ,`FM.AST.DOMO.ZG.M3`    AS 'Claims on other sectors of the domestic
              economy (annual growth as % of broad money)'
20          ,`FM.AST.PRVT.ZG.M3`    AS 'Claims on private sector (annual growth
              as % of broad money)'
21          ,`TX.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
              services (% of commercial service exports)'
22          ,`TM.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
              services (% of commercial service imports)'
23          ,`FP.CPI.TOTL`                  AS 'Consumer price index (2010 =
              100)'
24          ,`IC.REG.COST.PC.ZS`    AS 'Cost of business start-up procedures (%
              of GNI per capita)'
25          ,`BN.CAB.XOKA.GD.ZS`    AS 'Current account balance (% of GDP)'
26          ,`FR.INR.DPST`                  AS 'Deposit interest rate (%)'
27          ,`FD.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector by
              banks (% of GDP)'
28          ,`FS.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector (% of
              GDP)'
29          ,`NE.EXP.GNFS.ZS`               AS 'Exports of goods and services
              (% of GDP)'
30          ,`NE.EXP.GNFS.KD.ZG`    AS 'Exports of goods and services (annual %
              growth)'
31          ,`NE.RSB.GNFS.ZS`               AS 'External balance on goods and
              services (% of GDP)'
32          ,`NE.CON.TETC.ZS`               AS 'Final consumption expenditure,
              etc. (% of GDP)'
33          ,`NE.CON.TETC.KD.ZG`    AS 'Final consumption expenditure, etc. (
              annual % growth)'
34          ,`TX.VAL.FOOD.ZS.UN`    AS 'Food exports (% of merchandise exports)
              '
35          ,`TM.VAL.FOOD.ZS.UN`    AS 'Food imports (% of merchandise imports)
              '
36          ,`BX.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net inflows
              (% of GDP)'
37          ,`BM.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net outflows
              (% of GDP)'
38          ,`TX.VAL.FUEL.ZS.UN`    AS 'Fuel exports (% of merchandise exports)
              '
39          ,`TM.VAL.FUEL.ZS.UN`    AS 'Fuel imports (% of merchandise imports)
              '
40          ,`NY.GDP.MKTP.CD`               AS 'GDP (current US$)'
41          ,`NY.GDP.MKTP.KD.ZG`    AS 'GDP growth (annual %)'
42          ,`NY.GDP.PCAP.CD`               AS 'GDP per capita (current US$)'
43          ,`NY.GDP.PCAP.KD.ZG`    AS 'GDP per capita growth (annual %)'
44          ,`NY.GDP.PCAP.PP.CD`    AS 'GDP per capita, PPP (current
              international $)'
45          ,`NE.CON.GOVT.ZS`               AS 'General government final
              consumption expenditure (% of GDP)'
46          ,`NE.CON.GOVT.KD.ZG`    AS 'General government final consumption
              expenditure (annual % growth)'
47          ,`NY.GNP.MKTP.CD`               AS 'GNI (current US$)'
48          ,`NY.GNP.MKTP.KD.ZG`    AS 'GNI growth (annual %)'
```

```
49          ,`NY.GNP.PCAP.KD.ZG`    AS 'GNI per capita growth (annual %)'
50          ,`NY.GNP.PCAP.PP.CD`    AS 'GNI per capita, PPP (current
              international $)'
51          ,`NE.GDI.TOTL.ZS`               AS 'Gross capital formation (% of
              GDP)'
52          ,`NE.GDI.TOTL.KD.ZG`    AS 'Gross capital formation (annual %
              growth)'
53          ,`NE.GDI.TOTL.CD`               AS 'Gross capital formation (
              current US$)'
54          ,`NY.GDS.TOTL.ZS`               AS 'Gross domestic savings (% of
              GDP)'
55          ,`NE.GDI.FTOT.ZS`               AS 'Gross fixed capital formation
              (% of GDP)'
56          ,`NE.GDI.FTOT.KD.ZG`    AS 'Gross fixed capital formation (annual %
               growth)'
57          ,`NE.GDI.FTOT.CD`               AS 'Gross fixed capital formation (
              current US$)'
58          ,`NE.GDI.FPRV.ZS`               AS 'Gross fixed capital formation,
              private sector (% of GDP)'
59          ,`NE.DAB.TOTL.ZS`               AS 'Gross national expenditure (%
              of GDP)'
60          ,`NE.DAB.TOTL.CD`               AS 'Gross national expenditure (
              current US$)'
61          ,`NY.GNS.ICTR.ZS`               AS 'Gross savings (% of GDP)'
62          ,`NY.GNS.ICTR.GN.ZS`    AS 'Gross savings (% of GNI)'
63          ,`NY.GNS.ICTR.CD`               AS 'Gross savings (current US$)'
64          ,`NY.GDP.FCST.CD`               AS 'Gross value added at factor
              cost (current US$)'
65          ,`TX.VAL.TECH.MF.ZS`    AS 'High-technology exports (% of
              manufactured exports)'
66          ,`NE.CON.PRVT.PC.KD`    AS 'Household final consumption expenditure
               per capita (constant 2010 US$)'
67          ,`NE.CON.PRVT.PC.KD.ZG` AS 'Household final consumption expenditure
               per capita growth (annual %)'
68          ,`NE.CON.PETC.ZS`               AS 'Household final consumption
              expenditure, etc. (% of GDP)'
69          ,`NE.CON.PETC.KD.ZG`    AS 'Household final consumption expenditure
              , etc. (annual % growth)'
70          ,`NE.CON.PRVT.KD.ZG`    AS 'Household final consumption expenditure
               (annual % growth)'
71          ,`BX.GSR.CCIS.ZS`               AS 'ICT service exports (% of
              service exports, BoP)'
72          ,`IQ.CPA.IRAI.XQ`               AS 'IDA resource allocation index
              (1=low to 6=high)'
73          ,`NE.IMP.GNFS.ZS`               AS 'Imports of goods and services
              (% of GDP)'
74          ,`NE.IMP.GNFS.KD.ZG`    AS 'Imports of goods and services (annual %
               growth)'
75          ,`NE.IMP.GNFS.CD`               AS 'Imports of goods and services (
              current US$)'
76          ,`NV.IND.TOTL.ZS`               AS 'Industry, value added (% of GDP
              )'
77          ,`NV.IND.TOTL.KD.ZG`    AS 'Industry, value added (annual % growth)
              '
78          ,`FP.CPI.TOTL.ZG`               AS 'Inflation, consumer prices (
              annual %)'
```

```
79          ,`TX.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
               commercial service exports)'
80          ,`TM.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
               commercial service imports)'
81          ,`FR.INR.LNDP`                  AS 'Interest rate spread (lending
               rate minus deposit rate, %)'
82          ,`FR.INR.LEND`                  AS 'Lending interest rate (%)'
83          ,`CM.MKT.LDOM.NO`               AS 'Listed domestic companies,
               total [CM.MKT.LDOM.NO]'
84          ,`TM.VAL.MANF.ZS.UN`    AS 'Manufactures imports (% of merchandise
               imports)'
85          ,`TX.VAL.MANF.ZS.UN`    AS 'Manufactures exports (% of merchandise
               exports)'
86          ,`NV.IND.MANF.ZS`               AS 'Manufacturing, value added (%
               of GDP)'
87          ,`NV.IND.MANF.KD.ZG`    AS 'Manufacturing, value added (annual %
               growth)'
88          ,`CM.MKT.LCAP.GD.ZS`    AS 'Market capitalization of listed
               domestic companies (% of GDP)'
89          ,`TG.VAL.TOTL.GD.ZS`    AS 'Merchandise trade (% of GDP)'
90          ,`MS.MIL.XPND.GD.ZS`    AS 'Military expenditure (% of GDP)'
91          ,`TX.VAL.MMTL.ZS.UN`    AS 'Ores and metals exports (% of
               merchandise exports)'
92          ,`TM.VAL.MMTL.ZS.UN`    AS 'Ores and metals imports (% of
               merchandise imports)'
93          ,`IQ.SCI.OVRL`                  AS 'Overall level of statistical
               capacity (scale 0 - 100)'
94          ,`IQ.SCI.PRDC`                  AS 'Periodicity and timeliness
               assessment of statistical capacity (scale 0 - 100)'
95          ,`BX.TRF.PWKR.DT.GD.ZS` AS 'Personal remittances, received (% of
               GDP)'
96          ,`SP.POP.GROW`                  AS 'Population growth (annual %)'
97          ,`EN.URB.LCTY.UR.ZS`    AS 'Population in the largest city (% of
               urban population)'
98          ,`IC.CRD.PRVT.ZS`               AS 'Private credit bureau coverage
               (% of adults)'
99          ,`SG.GEN.PARL.ZS`               AS 'Proportion of seats held by
               women in national parliaments (%)'
100         ,`IC.CRD.PUBL.ZS`               AS 'Public credit registry coverage
                (% of adults)'
101         ,`FR.INR.RINR`                  AS 'Real interest rate (%)'
102         ,`FR.INR.RISK`                  AS 'Risk premium on lending (
               lending rate minus treasury bill rate, %)'
103         ,`SP.RUR.TOTL.ZS`               AS 'Rural population (% of total
               population)'
104         ,`SP.RUR.TOTL.ZG`               AS 'Rural population growth (annual
                %)'
105         ,`CM.MKT.INDX.ZG`               AS 'S&P Global Equity Indices (
               annual % change)'
106         ,`NV.SRV.TETC.ZS`               AS 'Services, etc., value added (%
               of GDP)'
107         ,`NV.SRV.TETC.KD.ZG`    AS 'Services, etc., value added (annual %
               growth)'
108         ,`IQ.SCI.SRCE`                  AS 'Source data assessment of
               statistical capacity (scale 0 - 100)'
109         ,`CM.MKT.TRAD.GD.ZS`    AS 'Stocks traded, total value (% of GDP)'
```

```
110         ,`CM.MKT.TRNR`                  AS 'Stocks traded, turnover ratio
               of domestic shares (%)'
111         ,`IC.TAX.PAYM`                  AS 'Tax payments (number)'
112         ,`IC.WRH.DURS`                  AS 'Time required to build a
               warehouse (days)'
113         ,`IC.LGL.DURS`                  AS 'Time required to enforce a
               contract (days)'
114         ,`IC.PRP.DURS`                  AS 'Time required to register
               property (days)'
115         ,`IC.REG.DURS`                  AS 'Time required to start a
               business (days)'
116         ,`IC.TAX.DURS`                  AS 'Time to prepare and pay taxes (
               hours)'
117         ,`IC.TAX.TOTL.CP.ZS`    AS 'Total tax rate (% of commercial profits
               )'
118         ,`NE.TRD.GNFS.ZS`               AS 'Trade (% of GDP)'
119         ,`BG.GSR.NFSV.GD.ZS`    AS 'Trade in services (% of GDP)'
120         ,`TX.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
               service exports)'
121         ,`TM.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
               service imports)'
122         ,`TX.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
               service exports)'
123         ,`TM.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
               service imports)'
124         ,`SP.URB.TOTL.IN.ZS`    AS 'Urban population (% of total)'
125         ,`SP.URB.GROW`                  AS 'Urban population growth (annual
               %)'
126         ,`FP.WPI.TOTL`                  AS 'Wholesale price index (2010 =
               100)'
127 FROM TBL_COUNTRY
```

Linsting A.3: Query para mineração de um País no decorrer dos anos

```
1  SELECT
2          `Time`
3      ,`CountryName`
4          ,`SP.POP.DPND`                  AS 'Age dependency ratio (% of
               working-age population)'
5          ,`SP.POP.DPND.OL`               AS 'Age dependency ratio, old (% of
               working-age population)'
6          ,`TM.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials imports (%
               of merchandise imports)'
7          ,`TX.VAL.AGRI.ZS.UN`    AS 'Agricultural raw materials export (% of
               mechandise exports)'
8          ,`EA.PRD.AGRI.KD`               AS 'Agriculture value added per
               worker (constant 2010 US$)'
9          ,`NV.AGR.TOTL.ZS`               AS 'Agriculture, value added (% of
               GDP)'
10         ,`NV.AGR.TOTL.KD.ZG`    AS 'Agriculture, value added (annual %
               growth)'
11         ,`FB.BNK.CAPA.ZS`               AS 'Bank capital to assets ratio
               (%)'
12         ,`FD.RES.LIQU.AS.ZS`    AS 'Bank liquid reserves to bank assets
               ratio (%)'
```

```
13          ,`FB.AST.NPER.ZS`                AS 'Bank nonperforming loans to
              total gross loans (%)'
14          ,`FM.LBL.BMNY.GD.ZS`    AS 'Broad money (% of GDP)'
15          ,`FM.LBL.BMNY.ZG`                AS 'Broad money growth (annual %)'
16          ,`FM.LBL.BMNY.IR.ZS`    AS 'Broad money to total reserves ratio'
17          ,`IC.BUS.DISC.XQ`                AS 'Business extent of disclosure
              index (0=less disclosure to 10=more disclosure)'
18          ,`FM.AST.CGOV.ZG.M3`    AS 'Claims on central government (annual
              growth as % of broad money)'
19          ,`FS.AST.CGOV.GD.ZS`    AS 'Claims on central government, etc. (%
              GDP)'
20          ,`FS.AST.DOMO.GD.ZS`    AS 'Claims on other sectors of the domestic
               economy (% of GDP)'
21          ,`FM.AST.DOMO.ZG.M3`    AS 'Claims on other sectors of the domestic
               economy (annual growth as % of broad money)'
22          ,`FM.AST.PRVT.ZG.M3`    AS 'Claims on private sector (annual growth
               as % of broad money)'
23          ,`TX.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
              services (% of commercial service exports)'
24          ,`TM.VAL.OTHR.ZS.WT`    AS 'Computer, communications and other
              services (% of commercial service imports)'
25          ,`FP.CPI.TOTL`                   AS 'Consumer price index (2010 =
              100)'
26          ,`IC.REG.COST.PC.ZS`    AS 'Cost of business start-up procedures (%
               of GNI per capita)'
27          ,`BN.CAB.XOKA.GD.ZS`    AS 'Current account balance (% of GDP)'
28          ,`FR.INR.DPST`                   AS 'Deposit interest rate (%)'
29          ,`FD.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector by
              banks (% of GDP)'
30          ,`FS.AST.PRVT.GD.ZS`    AS 'Domestic credit to private sector (% of
               GDP)'
31          ,`NE.EXP.GNFS.ZS`                AS 'Exports of goods and services
              (% of GDP)'
32          ,`NE.EXP.GNFS.KD.ZG`    AS 'Exports of goods and services (annual %
               growth)'
33          ,`NE.RSB.GNFS.ZS`                AS 'External balance on goods and
              services (% of GDP)'
34          ,`NE.CON.TETC.ZS`                AS 'Final consumption expenditure,
              etc. (% of GDP)'
35          ,`NE.CON.TETC.KD.ZG`    AS 'Final consumption expenditure, etc. (
              annual % growth)'
36          ,`TX.VAL.FOOD.ZS.UN`    AS 'Food exports (% of merchandise exports)
              '
37          ,`TM.VAL.FOOD.ZS.UN`    AS 'Food imports (% of merchandise imports)
              '
38          ,`BX.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net inflows
              (% of GDP)'
39          ,`BM.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net outflows
               (% of GDP)'
40          ,`TX.VAL.FUEL.ZS.UN`    AS 'Fuel exports (% of merchandise exports)
              '
41          ,`TM.VAL.FUEL.ZS.UN`    AS 'Fuel imports (% of merchandise imports)
              '
42          ,`NY.GDP.MKTP.CD`                AS 'GDP (current US$)'
43          ,`NY.GDP.MKTP.KD.ZG`    AS 'GDP growth (annual %)'
44          ,`NY.GDP.PCAP.CD`                AS 'GDP per capita (current US$)'
```

```
45          ,`NY.GDP.PCAP.KD.ZG`    AS 'GDP per capita growth (annual %)'
46          ,`NY.GDP.PCAP.PP.CD`    AS 'GDP per capita, PPP (current
              international $)'
47          ,`NE.CON.GOVT.ZS`               AS 'General government final
              consumption expenditure (% of GDP)'
48          ,`NE.CON.GOVT.KD.ZG`    AS 'General government final consumption
              expenditure (annual % growth)'
49          ,`NY.GNP.MKTP.CD`               AS 'GNI (current US$)'
50          ,`NY.GNP.MKTP.KD.ZG`    AS 'GNI growth (annual %)'
51          ,`NY.GNP.PCAP.KD.ZG`    AS 'GNI per capita growth (annual %)'
52          ,`NY.GNP.PCAP.PP.CD`    AS 'GNI per capita, PPP (current
              international $)'
53          ,`NE.GDI.TOTL.ZS`               AS 'Gross capital formation (% of
              GDP)'
54          ,`NE.GDI.TOTL.KD.ZG`    AS 'Gross capital formation (annual %
              growth)'
55          ,`NE.GDI.TOTL.CD`               AS 'Gross capital formation (
              current US$)'
56          ,`NY.GDS.TOTL.ZS`               AS 'Gross domestic savings (% of
              GDP)'
57          ,`NE.GDI.FTOT.ZS`               AS 'Gross fixed capital formation
              (% of GDP)'
58          ,`NE.GDI.FTOT.KD.ZG`    AS 'Gross fixed capital formation (annual %
               growth)'
59          ,`NE.GDI.FTOT.CD`               AS 'Gross fixed capital formation (
              current US$)'
60          ,`NE.GDI.FPRV.ZS`               AS 'Gross fixed capital formation,
              private sector (% of GDP)'
61          ,`NE.DAB.TOTL.ZS`               AS 'Gross national expenditure (%
              of GDP)'
62          ,`NE.DAB.TOTL.CD`               AS 'Gross national expenditure (
              current US$)'
63          ,`NY.GNS.ICTR.ZS`               AS 'Gross savings (% of GDP)'
64          ,`NY.GNS.ICTR.GN.ZS`    AS 'Gross savings (% of GNI)'
65          ,`NY.GNS.ICTR.CD`               AS 'Gross savings (current US$)'
66          ,`NY.GDP.FCST.CD`               AS 'Gross value added at factor
              cost (current US$)'
67          ,`TX.VAL.TECH.MF.ZS`    AS 'High-technology exports (% of
              manufactured exports)'
68          ,`NE.CON.PRVT.PC.KD`    AS 'Household final consumption expenditure
               per capita (constant 2010 US$)'
69          ,`NE.CON.PRVT.PC.KD.ZG` AS 'Household final consumption expenditure
               per capita growth (annual %)'
70          ,`NE.CON.PETC.ZS`               AS 'Household final consumption
              expenditure, etc. (% of GDP)'
71          ,`NE.CON.PETC.KD.ZG`    AS 'Household final consumption expenditure
              , etc. (annual % growth)'
72          ,`NE.CON.PRVT.KD.ZG`    AS 'Household final consumption expenditure
               (annual % growth)'
73          ,`BX.GSR.CCIS.ZS`               AS 'ICT service exports (% of
              service exports, BoP)'
74          ,`IQ.CPA.IRAI.XQ`               AS 'IDA resource allocation index
              (1=low to 6=high)'
75          ,`NE.IMP.GNFS.ZS`               AS 'Imports of goods and services
              (% of GDP)'
```

```
76          ,`NE.IMP.GNFS.KD.ZG`    AS 'Imports of goods and services (annual %
                growth)'
77          ,`NE.IMP.GNFS.CD`               AS 'Imports of goods and services (
                current US$)'
78          ,`NV.IND.TOTL.ZS`               AS 'Industry, value added (% of GDP
                )'
79          ,`NV.IND.TOTL.KD.ZG`    AS 'Industry, value added (annual % growth)
                '
80          ,`FP.CPI.TOTL.ZG`               AS 'Inflation, consumer prices (
                annual %)'
81          ,`TX.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
                commercial service exports)'
82          ,`TM.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
                commercial service imports)'
83          ,`FR.INR.LNDP`                  AS 'Interest rate spread (lending
                rate minus deposit rate, %)'
84          ,`FR.INR.LEND`                  AS 'Lending interest rate (%)'
85          ,`CM.MKT.LDOM.NO`               AS 'Listed domestic companies,
                total [CM.MKT.LDOM.NO]'
86          ,`TM.VAL.MANF.ZS.UN`    AS 'Manufactures imports (% of merchandise
                imports)'
87          ,`TX.VAL.MANF.ZS.UN`    AS 'Manufactures exports (% of merchandise
                exports)'
88          ,`NV.IND.MANF.ZS`               AS 'Manufacturing, value added (%
                of GDP)'
89          ,`NV.IND.MANF.KD.ZG`    AS 'Manufacturing, value added (annual %
                growth)'
90          ,`CM.MKT.LCAP.GD.ZS`    AS 'Market capitalization of listed
                domestic companies (% of GDP)'
91          ,`TG.VAL.TOTL.GD.ZS`    AS 'Merchandise trade (% of GDP)'
92          ,`MS.MIL.XPND.GD.ZS`    AS 'Military expenditure (% of GDP)'
93          ,`TX.VAL.MMTL.ZS.UN`    AS 'Ores and metals exports (% of
                merchandise exports)'
94          ,`TM.VAL.MMTL.ZS.UN`    AS 'Ores and metals imports (% of
                merchandise imports)'
95          ,`IQ.SCI.OVRL`                  AS 'Overall level of statistical
                capacity (scale 0 - 100)'
96          ,`IQ.SCI.PRDC`                  AS 'Periodicity and timeliness
                assessment of statistical capacity (scale 0 - 100)'
97          ,`BX.TRF.PWKR.DT.GD.ZS` AS 'Personal remittances, received (% of
                GDP)'
98          ,`SP.POP.GROW`                  AS 'Population growth (annual %)'
99          ,`EN.URB.LCTY.UR.ZS`    AS 'Population in the largest city (% of
                urban population)'
100         ,`IC.CRD.PRVT.ZS`               AS 'Private credit bureau coverage
                (% of adults)'
101         ,`SG.GEN.PARL.ZS`               AS 'Proportion of seats held by
                women in national parliaments (%)'
102         ,`IC.CRD.PUBL.ZS`               AS 'Public credit registry coverage
                 (% of adults)'
103         ,`FR.INR.RINR`                  AS 'Real interest rate (%)'
104         ,`FR.INR.RISK`                  AS 'Risk premium on lending (
                lending rate minus treasury bill rate, %)'
105         ,`SP.RUR.TOTL.ZS`               AS 'Rural population (% of total
                population)'
```

```
106         ,'SP.RUR.TOTL.ZG'               AS 'Rural population growth (annual
                %)'
107         ,'CM.MKT.INDX.ZG'               AS 'S&P Global Equity Indices (
                annual % change)'
108         ,'NV.SRV.TETC.ZS'               AS 'Services, etc., value added (%
                of GDP)'
109         ,'NV.SRV.TETC.KD.ZG'    AS 'Services, etc., value added (annual %
                growth)'
110         ,'IQ.SCI.SRCE'                  AS 'Source data assessment of
                statistical capacity (scale 0 - 100)'
111         ,'CM.MKT.TRAD.GD.ZS'    AS 'Stocks traded, total value (% of GDP)'
112         ,'CM.MKT.TRNR'                  AS 'Stocks traded, turnover ratio
                of domestic shares (%)'
113         ,'IC.TAX.PAYM'                  AS 'Tax payments (number)'
114         ,'IC.WRH.DURS'                  AS 'Time required to build a
                warehouse (days)'
115         ,'IC.LGL.DURS'                  AS 'Time required to enforce a
                contract (days)'
116         ,'IC.PRP.DURS'                  AS 'Time required to register
                property (days)'
117         ,'IC.REG.DURS'                  AS 'Time required to start a
                business (days)'
118         ,'IC.TAX.DURS'                  AS 'Time to prepare and pay taxes (
                hours)'
119         ,'IC.TAX.TOTL.CP.ZS'    AS 'Total tax rate (% of commercial profits
                )'
120         ,'NE.TRD.GNFS.ZS'               AS 'Trade (% of GDP)'
121         ,'BG.GSR.NFSV.GD.ZS'    AS 'Trade in services (% of GDP)'
122         ,'TX.VAL.TRAN.ZS.WT'    AS 'Transport services (% of commercial
                service exports)'
123         ,'TM.VAL.TRAN.ZS.WT'    AS 'Transport services (% of commercial
                service imports)'
124         ,'TX.VAL.TRVL.ZS.WT'    AS 'Travel services (% of commercial
                service exports)'
125         ,'TM.VAL.TRVL.ZS.WT'    AS 'Travel services (% of commercial
                service imports)'
126         ,'SP.URB.TOTL.IN.ZS'    AS 'Urban population (% of total)'
127         ,'SP.URB.GROW'                  AS 'Urban population growth (annual
                %)'
128         ,'FP.WPI.TOTL'                  AS 'Wholesale price index (2010 =
                100)'
129 FROM TBL_COUNTRY
130 where 'CountryName' = 'Brazil'
```

Linsting A.4: Query para mineração de um País no decorrer dos anos

```
1  SELECT
2          'Time'
3       ,'CountryName'
4          ,'SP.POP.DPND'                  AS 'Age dependency ratio (% of
               working-age population)'
5          ,'SP.POP.DPND.OL'               AS 'Age dependency ratio, old (% of
                working-age population)'
6          ,'TM.VAL.AGRI.ZS.UN'    AS 'Agricultural raw materials imports (%
               of merchandise imports)'
```

```
7         ,`TX.VAL.AGRI.ZS.UN`     AS 'Agricultural raw materials export (% of
              mechandise exports)'
8         ,`EA.PRD.AGRI.KD`                AS 'Agriculture value added per
              worker (constant 2010 US$)'
9         ,`NV.AGR.TOTL.ZS`                AS 'Agriculture, value added (% of
              GDP)'
10        ,`NV.AGR.TOTL.KD.ZG`     AS 'Agriculture, value added (annual %
              growth)'
11        ,`FB.BNK.CAPA.ZS`                AS 'Bank capital to assets ratio
              (%)'
12        ,`FD.RES.LIQU.AS.ZS`     AS 'Bank liquid reserves to bank assets
              ratio (%)'
13        ,`FB.AST.NPER.ZS`                AS 'Bank nonperforming loans to
              total gross loans (%)'
14        ,`FM.LBL.BMNY.GD.ZS`     AS 'Broad money (% of GDP)'
15        ,`FM.LBL.BMNY.ZG`                AS 'Broad money growth (annual %)'
16        ,`FM.LBL.BMNY.IR.ZS`     AS 'Broad money to total reserves ratio'
17        ,`IC.BUS.DISC.XQ`                AS 'Business extent of disclosure
              index (0=less disclosure to 10=more disclosure)'
18        ,`FM.AST.CGOV.ZG.M3`     AS 'Claims on central government (annual
              growth as % of broad money)'
19        ,`FS.AST.CGOV.GD.ZS`     AS 'Claims on central government, etc. (%
              GDP)'
20        ,`FS.AST.DOMO.GD.ZS`     AS 'Claims on other sectors of the domestic
               economy (% of GDP)'
21        ,`FM.AST.DOMO.ZG.M3`     AS 'Claims on other sectors of the domestic
               economy (annual growth as % of broad money)'
22        ,`FM.AST.PRVT.ZG.M3`     AS 'Claims on private sector (annual growth
               as % of broad money)'
23        ,`TX.VAL.OTHR.ZS.WT`     AS 'Computer, communications and other
              services (% of commercial service exports)'
24        ,`TM.VAL.OTHR.ZS.WT`     AS 'Computer, communications and other
              services (% of commercial service imports)'
25        ,`FP.CPI.TOTL`                   AS 'Consumer price index (2010 =
              100)'
26        ,`IC.REG.COST.PC.ZS`     AS 'Cost of business start-up procedures (%
               of GNI per capita)'
27        ,`BN.CAB.XOKA.GD.ZS`     AS 'Current account balance (% of GDP)'
28        ,`FR.INR.DPST`                   AS 'Deposit interest rate (%)'
29        ,`FD.AST.PRVT.GD.ZS`     AS 'Domestic credit to private sector by
              banks (% of GDP)'
30        ,`FS.AST.PRVT.GD.ZS`     AS 'Domestic credit to private sector (% of
               GDP)'
31        ,`NE.EXP.GNFS.ZS`                AS 'Exports of goods and services
              (% of GDP)'
32        ,`NE.EXP.GNFS.KD.ZG`     AS 'Exports of goods and services (annual %
               growth)'
33        ,`NE.RSB.GNFS.ZS`                AS 'External balance on goods and
              services (% of GDP)'
34        ,`NE.CON.TETC.ZS`                AS 'Final consumption expenditure,
              etc. (% of GDP)'
35        ,`NE.CON.TETC.KD.ZG`     AS 'Final consumption expenditure, etc. (
              annual % growth)'
36        ,`TX.VAL.FOOD.ZS.UN`     AS 'Food exports (% of merchandise exports)
              '
```

```
37          ,`TM.VAL.FOOD.ZS.UN`     AS 'Food imports (% of merchandise imports)
              '
38          ,`BX.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net inflows
              (% of GDP)'
39          ,`BM.KLT.DINV.WD.GD.ZS` AS 'Foreign direct investment, net outflows
              (% of GDP)'
40          ,`TX.VAL.FUEL.ZS.UN`     AS 'Fuel exports (% of merchandise exports)
              '
41          ,`TM.VAL.FUEL.ZS.UN`     AS 'Fuel imports (% of merchandise imports)
              '
42          ,`NY.GDP.MKTP.CD`                AS 'GDP (current US$)'
43          ,`NY.GDP.MKTP.KD.ZG`     AS 'GDP growth (annual %)'
44          ,`NY.GDP.PCAP.CD`                AS 'GDP per capita (current US$)'
45          ,`NY.GDP.PCAP.KD.ZG`     AS 'GDP per capita growth (annual %)'
46          ,`NY.GDP.PCAP.PP.CD`     AS 'GDP per capita, PPP (current
              international $)'
47          ,`NE.CON.GOVT.ZS`                AS 'General government final
              consumption expenditure (% of GDP)'
48          ,`NE.CON.GOVT.KD.ZG`     AS 'General government final consumption
              expenditure (annual % growth)'
49          ,`NY.GNP.MKTP.CD`                AS 'GNI (current US$)'
50          ,`NY.GNP.MKTP.KD.ZG`     AS 'GNI growth (annual %)'
51          ,`NY.GNP.PCAP.KD.ZG`     AS 'GNI per capita growth (annual %)'
52          ,`NY.GNP.PCAP.PP.CD`     AS 'GNI per capita, PPP (current
              international $)'
53          ,`NE.GDI.TOTL.ZS`                AS 'Gross capital formation (% of
              GDP)'
54          ,`NE.GDI.TOTL.KD.ZG`     AS 'Gross capital formation (annual %
              growth)'
55          ,`NE.GDI.TOTL.CD`                AS 'Gross capital formation (
              current US$)'
56          ,`NY.GDS.TOTL.ZS`                AS 'Gross domestic savings (% of
              GDP)'
57          ,`NE.GDI.FTOT.ZS`                AS 'Gross fixed capital formation
              (% of GDP)'
58          ,`NE.GDI.FTOT.KD.ZG`     AS 'Gross fixed capital formation (annual %
               growth)'
59          ,`NE.GDI.FTOT.CD`                AS 'Gross fixed capital formation (
              current US$)'
60          ,`NE.GDI.FPRV.ZS`                AS 'Gross fixed capital formation,
              private sector (% of GDP)'
61          ,`NE.DAB.TOTL.ZS`                AS 'Gross national expenditure (%
              of GDP)'
62          ,`NE.DAB.TOTL.CD`                AS 'Gross national expenditure (
              current US$)'
63          ,`NY.GNS.ICTR.ZS`                AS 'Gross savings (% of GDP)'
64          ,`NY.GNS.ICTR.GN.ZS`     AS 'Gross savings (% of GNI)'
65          ,`NY.GNS.ICTR.CD`                AS 'Gross savings (current US$)'
66          ,`NY.GDP.FCST.CD`                AS 'Gross value added at factor
              cost (current US$)'
67          ,`TX.VAL.TECH.MF.ZS`     AS 'High-technology exports (% of
              manufactured exports)'
68          ,`NE.CON.PRVT.PC.KD`     AS 'Household final consumption expenditure
               per capita (constant 2010 US$)'
69          ,`NE.CON.PRVT.PC.KD.ZG` AS 'Household final consumption expenditure
               per capita growth (annual %)'
```

```
70          ,`NE.CON.PETC.ZS`               AS 'Household final consumption
              expenditure, etc. (% of GDP)'
71          ,`NE.CON.PETC.KD.ZG`    AS 'Household final consumption expenditure
              , etc. (annual % growth)'
72          ,`NE.CON.PRVT.KD.ZG`    AS 'Household final consumption expenditure
               (annual % growth)'
73          ,`BX.GSR.CCIS.ZS`               AS 'ICT service exports (% of
              service exports, BoP)'
74          ,`IQ.CPA.IRAI.XQ`               AS 'IDA resource allocation index
              (1=low to 6=high)'
75          ,`NE.IMP.GNFS.ZS`               AS 'Imports of goods and services
              (% of GDP)'
76          ,`NE.IMP.GNFS.KD.ZG`    AS 'Imports of goods and services (annual %
               growth)'
77          ,`NE.IMP.GNFS.CD`               AS 'Imports of goods and services (
              current US$)'
78          ,`NV.IND.TOTL.ZS`               AS 'Industry, value added (% of GDP
              )'
79          ,`NV.IND.TOTL.KD.ZG`    AS 'Industry, value added (annual % growth)
              '
80          ,`FP.CPI.TOTL.ZG`               AS 'Inflation, consumer prices (
              annual %)'
81          ,`TX.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
              commercial service exports)'
82          ,`TM.VAL.INSF.ZS.WT`    AS 'Insurance and financial services (% of
              commercial service imports)'
83          ,`FR.INR.LNDP`                  AS 'Interest rate spread (lending
              rate minus deposit rate, %)'
84          ,`FR.INR.LEND`                  AS 'Lending interest rate (%)'
85          ,`CM.MKT.LDOM.NO`               AS 'Listed domestic companies,
              total [CM.MKT.LDOM.NO]'
86          ,`TM.VAL.MANF.ZS.UN`    AS 'Manufactures imports (% of merchandise
              imports)'
87          ,`TX.VAL.MANF.ZS.UN`    AS 'Manufactures exports (% of merchandise
              exports)'
88          ,`NV.IND.MANF.ZS`               AS 'Manufacturing, value added (%
              of GDP)'
89          ,`NV.IND.MANF.KD.ZG`    AS 'Manufacturing, value added (annual %
              growth)'
90          ,`CM.MKT.LCAP.GD.ZS`    AS 'Market capitalization of listed
              domestic companies (% of GDP)'
91          ,`TG.VAL.TOTL.GD.ZS`    AS 'Merchandise trade (% of GDP)'
92          ,`MS.MIL.XPND.GD.ZS`    AS 'Military expenditure (% of GDP)'
93          ,`TX.VAL.MMTL.ZS.UN`    AS 'Ores and metals exports (% of
              merchandise exports)'
94          ,`TM.VAL.MMTL.ZS.UN`    AS 'Ores and metals imports (% of
              merchandise imports)'
95          ,`IQ.SCI.OVRL`                  AS 'Overall level of statistical
              capacity (scale 0 - 100)'
96          ,`IQ.SCI.PRDC`                  AS 'Periodicity and timeliness
              assessment of statistical capacity (scale 0 - 100)'
97          ,`BX.TRF.PWKR.DT.GD.ZS` AS 'Personal remittances, received (% of
              GDP)'
98          ,`SP.POP.GROW`                  AS 'Population growth (annual %)'
99          ,`EN.URB.LCTY.UR.ZS`    AS 'Population in the largest city (% of
              urban population)'
```

```
100         ,`IC.CRD.PRVT.ZS`               AS 'Private credit bureau coverage
              (% of adults)'
101         ,`SG.GEN.PARL.ZS`               AS 'Proportion of seats held by
              women in national parliaments (%)'
102         ,`IC.CRD.PUBL.ZS`               AS 'Public credit registry coverage
              (% of adults)'
103         ,`FR.INR.RINR`                  AS 'Real interest rate (%)'
104         ,`FR.INR.RISK`                  AS 'Risk premium on lending (
              lending rate minus treasury bill rate, %)'
105         ,`SP.RUR.TOTL.ZS`               AS 'Rural population (% of total
              population)'
106         ,`SP.RUR.TOTL.ZG`               AS 'Rural population growth (annual
              %)'
107         ,`CM.MKT.INDX.ZG`               AS 'S&P Global Equity Indices (
              annual % change)'
108         ,`NV.SRV.TETC.ZS`               AS 'Services, etc., value added (%
              of GDP)'
109         ,`NV.SRV.TETC.KD.ZG`    AS 'Services, etc., value added (annual %
              growth)'
110         ,`IQ.SCI.SRCE`                  AS 'Source data assessment of
              statistical capacity (scale 0 - 100)'
111         ,`CM.MKT.TRAD.GD.ZS`    AS 'Stocks traded, total value (% of GDP)'
112         ,`CM.MKT.TRNR`                  AS 'Stocks traded, turnover ratio
              of domestic shares (%)'
113         ,`IC.TAX.PAYM`                  AS 'Tax payments (number)'
114         ,`IC.WRH.DURS`                  AS 'Time required to build a
              warehouse (days)'
115         ,`IC.LGL.DURS`                  AS 'Time required to enforce a
              contract (days)'
116         ,`IC.PRP.DURS`                  AS 'Time required to register
              property (days)'
117         ,`IC.REG.DURS`                  AS 'Time required to start a
              business (days)'
118         ,`IC.TAX.DURS`                  AS 'Time to prepare and pay taxes (
              hours)'
119         ,`IC.TAX.TOTL.CP.ZS`    AS 'Total tax rate (% of commercial profits
              )'
120         ,`NE.TRD.GNFS.ZS`               AS 'Trade (% of GDP)'
121         ,`BG.GSR.NFSV.GD.ZS`    AS 'Trade in services (% of GDP)'
122         ,`TX.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
              service exports)'
123         ,`TM.VAL.TRAN.ZS.WT`    AS 'Transport services (% of commercial
              service imports)'
124         ,`TX.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
              service exports)'
125         ,`TM.VAL.TRVL.ZS.WT`    AS 'Travel services (% of commercial
              service imports)'
126         ,`SP.URB.TOTL.IN.ZS`    AS 'Urban population (% of total)'
127         ,`SP.URB.GROW`                  AS 'Urban population growth (annual
              %)'
128         ,`FP.WPI.TOTL`                  AS 'Wholesale price index (2010 =
              100)'
129 FROM TBL_COUNTRY
130 where `CountryName` = `Brazil`
```

Apêndice B

Logs de saída

Linsting B.1: Log de saída para o indicador *GDP Per Capita (Current US$)* com 2 *bins* e frequência igual.

```
1  === Run information ===
2
3  Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2 -batch-size 500
4  Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
       weka.filters.unsupervised.attribute.Remove-R6
       ,11,16-17,23-24,26-28,30-31,35-36,39-40,43-44,46-50,53-54,57-58,60-61,67,72,75,85,87-
       weka.filters.unsupervised.attribute.Remove-R71-weka.filters.unsupervised
       .attribute.Remove-R37-38-weka.filters.unsupervised.attribute.Remove-R27-
       weka.filters.unsupervised.attribute.Discretize-F-B2-M-1.0-R26
5  Instances:    320
6  Attributes:   82
7                Age dependency ratio (% of working-age population)
8                Age dependency ratio, old (% of working-age population)
9                Agricultural raw materials imports (% of merchandise imports)
10               Agricultural raw materials export (% of mechandise exports)
11               Agriculture value added per worker (constant 2010 US$)
12               Agriculture, value added (annual % growth)
13               Bank capital to assets ratio (%)
14               Bank liquid reserves to bank assets ratio (%)
15               Bank nonperforming loans to total gross loans (%)
16               Broad money growth (annual %)
17               Broad money to total reserves ratio
18               Business extent of disclosure index (0=less disclosure to 10=
                     more disclosure)
19               Claims on central government (annual growth as % of broad
                     money)
20               Claims on other sectors of the domestic economy (annual
                     growth as % of broad money)
21               Claims on private sector (annual growth as % of broad money)
22               Computer, communications and other services (% of commercial
                     service exports)
23               Computer, communications and other services (% of commercial
                     service imports)
24               Consumer price index (2010 = 100)
25               Deposit interest rate (%)
26               Exports of goods and services (annual % growth)
27               Final consumption expenditure, etc. (annual % growth)
```

```
28                Food exports (% of merchandise exports)
29                Food imports (% of merchandise imports)
30                Fuel exports (% of merchandise exports)
31                Fuel imports (% of merchandise imports)
32                GDP per capita (current US$)
33                General government final consumption expenditure (annual %
                      growth)
34                Gross capital formation (annual % growth)
35                Gross capital formation (current US$)
36                Gross fixed capital formation (annual % growth)
37                Gross fixed capital formation (current US$)
38                Gross national expenditure (current US$)
39                Gross savings (current US$)
40                Gross value added at factor cost (current US$)
41                High-technology exports (% of manufactured exports)
42                Household final consumption expenditure, etc. (annual %
                      growth)
43                Household final consumption expenditure (annual % growth)
44                ICT service exports (% of service exports, BoP)
45                IDA resource allocation index (1=low to 6=high)
46                Imports of goods and services (annual % growth)
47                Imports of goods and services (current US$)
48                Industry, value added (annual % growth)
49                Inflation, consumer prices (annual %)
50                Insurance and financial services (% of commercial service
                      exports)
51                Insurance and financial services (% of commercial service
                      imports)
52                Interest rate spread (lending rate minus deposit rate, %)
53                Lending interest rate (%)
54                Listed domestic companies, total [CM.MKT.LDOM.NO]
55                Manufactures imports (% of merchandise imports)
56                Manufactures exports (% of merchandise exports)
57                Manufacturing, value added (annual % growth)
58                Ores and metals exports (% of merchandise exports)
59                Ores and metals imports (% of merchandise imports)
60                Overall level of statistical capacity (scale 0 - 100)
61                Periodicity and timeliness assessment of statistical capacity
                      (scale 0 - 100)
62                Population growth (annual %)
63                Population in the largest city (% of urban population)
64                Private credit bureau coverage (% of adults)
65                Proportion of seats held by women in national parliaments (%)
66                Public credit registry coverage (% of adults)
67                Real interest rate (%)
68                Risk premium on lending (lending rate minus treasury bill
                      rate, %)
69                Rural population (% of total population)
70                Rural population growth (annual %)
71                S&P Global Equity Indices (annual % change)
72                Services, etc., value added (annual % growth)
73                Source data assessment of statistical capacity (scale 0 -
                      100)
74                Stocks traded, turnover ratio of domestic shares (%)
75                Tax payments (number)
76                Time required to build a warehouse (days)
```

```
77                Time required to enforce a contract (days)
78                Time required to register property (days)
79                Time required to start a business (days)
80                Time to prepare and pay taxes (hours)
81                Total tax rate (% of commercial profits)
82                Transport services (% of commercial service exports)
83                Transport services (% of commercial service imports)
84                Travel services (% of commercial service exports)
85                Travel services (% of commercial service imports)
86                Urban population (% of total)
87                Urban population growth (annual %)
88                Wholesale price index (2010 = 100)
89 Test mode:    10-fold cross-validation
90
91 === Classifier model (full training set) ===
92
93 J48 pruned tree
94 ------------------
95
96 Age dependency ratio, old (% of working-age population) <= 15.25902: '(-inf
       -22714.718895]' (121.0/1.0)
97 Age dependency ratio, old (% of working-age population) > 15.25902
98 |   Inflation, consumer prices (annual %) <= 5.57775
99 |   |   Agriculture value added per worker (constant 2010 US$) <=
       12386.45763
100 |   |   |   Gross national expenditure (current US$) <= 256430287619.844:
       '(-inf-22714.718895]' (18.77/1.64)
101 |   |   |   Gross national expenditure (current US$) > 256430287619.844:
       '(22714.718895-inf)' (4.01/0.12)
102 |   |   Agriculture value added per worker (constant 2010 US$) >
       12386.45763
103 |   |   |   Imports of goods and services (current US$) <= 84481244511.3537
104 |   |   |   |   Private credit bureau coverage (% of adults) <= 95.7:
       '(22714.718895-inf)' (5.36)
105 |   |   |   |   Private credit bureau coverage (% of adults) > 95.7: '(-inf
       -22714.718895]' (2.56)
106 |   |   |   Imports of goods and services (current US$) > 84481244511.3537:
        '(22714.718895-inf)' (147.1)
107 |   Inflation, consumer prices (annual %) > 5.57775: '(-inf-22714.718895]'
       (19.19)
108
109 Number of Leaves  :     7
110
111 Size of the tree :     13
112
113
114 Time taken to build model: 0.02 seconds
115
116 === Stratified cross-validation ===
117 === Summary ===
118
119 Correctly Classified Instances         304               95.5975 %
120 Incorrectly Classified Instances        14                4.4025 %
121 Kappa statistic                          0.9119
122 Mean absolute error                      0.0523
123 Root mean squared error                  0.2041
```

```
Relative absolute error                 10.4685 %
Root relative squared error             40.8222 %
Total Number of Instances               318
Ignored Class Unknown Instances                 2

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
                    ROC Area  PRC Area  Class
                 0,943     0,031     0,968       0,943     0,955      0,912
                    0,953     0,924     '(-inf-22714.718895]'
                 0,969     0,057     0,945       0,969     0,957      0,912
                    0,958     0,947     '(22714.718895-inf)'
Weighted Avg.    0,956     0,044     0,956       0,956     0,956      0,912
    0,955     0,935

=== Confusion Matrix ===

   a   b   <-- classified as
 150   9 |   a = '(-inf-22714.718895]'
   5 154 |   b = '(22714.718895-inf)'
```

Linsting B.2: Log de saída para o indicador *GDP Per Capita Growth (Annual %)* com 3 *bins* e frequência diferente.

```
1  === Run information ===
2
3  Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
4  Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
       weka.filters.unsupervised.attribute.Remove-R6
       ,11,16-17,23-24,26-28,30-31,35-36,39-40,43-44,46-50,53-54,57-58,60-61,67,72,75,85,87-
       weka.filters.unsupervised.attribute.Remove-R71-weka.filters.unsupervised
       .attribute.Remove-R37-38-weka.filters.unsupervised.attribute.Remove-R26-
       weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R26
5  Instances:    320
6  Attributes:   82
7                Age dependency ratio (% of working-age population)
8                Age dependency ratio, old (% of working-age population)
9                Agricultural raw materials imports (% of merchandise imports)
10               Agricultural raw materials export (% of mechandise exports)
11               Agriculture value added per worker (constant 2010 US$)
12               Agriculture, value added (annual % growth)
13               Bank capital to assets ratio (%)
14               Bank liquid reserves to bank assets ratio (%)
15               Bank nonperforming loans to total gross loans (%)
16               Broad money growth (annual %)
17               Broad money to total reserves ratio
18               Business extent of disclosure index (0=less disclosure to 10=
                     more disclosure)
19               Claims on central government (annual growth as % of broad
                     money)
20               Claims on other sectors of the domestic economy (annual
                     growth as % of broad money)
21               Claims on private sector (annual growth as % of broad money)
22               Computer, communications and other services (% of commercial
                     service exports)
```

```
23              Computer, communications and other services (% of commercial
                    service imports)
24              Consumer price index (2010 = 100)
25              Deposit interest rate (%)
26              Exports of goods and services (annual % growth)
27              Final consumption expenditure, etc. (annual % growth)
28              Food exports (% of merchandise exports)
29              Food imports (% of merchandise imports)
30              Fuel exports (% of merchandise exports)
31              Fuel imports (% of merchandise imports)
32              GDP per capita growth (annual %)
33              General government final consumption expenditure (annual %
                    growth)
34              Gross capital formation (annual % growth)
35              Gross capital formation (current US$)
36              Gross fixed capital formation (annual % growth)
37              Gross fixed capital formation (current US$)
38              Gross national expenditure (current US$)
39              Gross savings (current US$)
40              Gross value added at factor cost (current US$)
41              High-technology exports (% of manufactured exports)
42              Household final consumption expenditure, etc. (annual %
                    growth)
43              Household final consumption expenditure (annual % growth)
44              ICT service exports (% of service exports, BoP)
45              IDA resource allocation index (1=low to 6=high)
46              Imports of goods and services (annual % growth)
47              Imports of goods and services (current US$)
48              Industry, value added (annual % growth)
49              Inflation, consumer prices (annual %)
50              Insurance and financial services (% of commercial service
                    exports)
51              Insurance and financial services (% of commercial service
                    imports)
52              Interest rate spread (lending rate minus deposit rate, %)
53              Lending interest rate (%)
54              Listed domestic companies, total [CM.MKT.LDOM.NO]
55              Manufactures imports (% of merchandise imports)
56              Manufactures exports (% of merchandise exports)
57              Manufacturing, value added (annual % growth)
58              Ores and metals exports (% of merchandise exports)
59              Ores and metals imports (% of merchandise imports)
60              Overall level of statistical capacity (scale 0 - 100)
61              Periodicity and timeliness assessment of statistical capacity
                    (scale 0 - 100)
62              Population growth (annual %)
63              Population in the largest city (% of urban population)
64              Private credit bureau coverage (% of adults)
65              Proportion of seats held by women in national parliaments (%)
66              Public credit registry coverage (% of adults)
67              Real interest rate (%)
68              Risk premium on lending (lending rate minus treasury bill
                    rate, %)
69              Rural population (% of total population)
70              Rural population growth (annual %)
71              S&P Global Equity Indices (annual % change)
```

```
72                Services, etc., value added (annual % growth)
73                Source data assessment of statistical capacity (scale 0 -
                      100)
74                Stocks traded, turnover ratio of domestic shares (%)
75                Tax payments (number)
76                Time required to build a warehouse (days)
77                Time required to enforce a contract (days)
78                Time required to register property (days)
79                Time required to start a business (days)
80                Time to prepare and pay taxes (hours)
81                Total tax rate (% of commercial profits)
82                Transport services (% of commercial service exports)
83                Transport services (% of commercial service imports)
84                Travel services (% of commercial service exports)
85                Travel services (% of commercial service imports)
86                Urban population (% of total)
87                Urban population growth (annual %)
88                Wholesale price index (2010 = 100)
89 Test mode:     10-fold cross-validation
90
91 === Classifier model (full training set) ===
92
93 J48 pruned tree
94 ------------------
95
96 Gross fixed capital formation (annual % growth) <= -6.29428
97 |   Industry, value added (annual % growth) <= -1.15125: '(-inf--1.27109]'
       (35.44/1.33)
98 |   Industry, value added (annual % growth) > -1.15125:
       '(-1.27109-6.16451]' (3.16/0.14)
99 Gross fixed capital formation (annual % growth) > -6.29428
100 |   Population in the largest city (% of urban population) <= 5.61147:
       '(6.16451-inf)' (10.88)
101 |   Population in the largest city (% of urban population) > 5.61147
102 |   |   Gross fixed capital formation (annual % growth) <= 15.97391
103 |   |   |   Industry, value added (annual % growth) <= 7.53006
104 |   |   |   |   Household final consumption expenditure (annual % growth)
       <= 0.64504
105 |   |   |   |   |   Industry, value added (annual % growth) <= -3.89137:
       '(-inf--1.27109]' (3.11/0.1)
106 |   |   |   |   |   Industry, value added (annual % growth) > -3.89137
107 |   |   |   |   |   |   S&P Global Equity Indices (annual % change) <=
       -49.04018: '(-inf--1.27109]' (3.03/1.02)
108 |   |   |   |   |   |   S&P Global Equity Indices (annual % change) >
       -49.04018: '(-1.27109-6.16451]' (37.36/1.25)
109 |   |   |   |   Household final consumption expenditure (annual % growth) >
        0.64504: '(-1.27109-6.16451]' (188.61/1.23)
110 |   |   |   Industry, value added (annual % growth) > 7.53006
111 |   |   |   |   Time required to enforce a contract (days) <= 588:
       '(-1.27109-6.16451]' (9.26)
112 |   |   |   |   Time required to enforce a contract (days) > 588:
       '(6.16451-inf)' (6.86/0.81)
113 |   |   Gross fixed capital formation (annual % growth) > 15.97391
114 |   |   |   Services, etc., value added (annual % growth) <= 6.94524:
       '(-1.27109-6.16451]' (11.68/1.6)
```

```
|   |   |   Services, etc., value added (annual % growth) > 6.94524: '(6.16451-inf)' (10.61/0.1)

Number of Leaves  :      11

Size of the tree :      21


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         291               90.9375 %
Incorrectly Classified Instances        29                9.0625 %
Kappa statistic                          0.7474
Mean absolute error                      0.072
Root mean squared error                  0.2376
Relative absolute error                 28.9942 %
Root relative squared error             67.7154 %
Total Number of Instances              320

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,805     0,025     0,825       0,805     0,815      0,788    0,901     0,679     '(-inf--1.27109]'
                 0,952     0,239     0,933       0,952     0,942      0,732    0,871     0,924     '(-1.27109-6.16451]'
                 0,700     0,017     0,808       0,700     0,750      0,728    0,886     0,721     '(6.16451-inf)'
Weighted Avg.    0,909     0,191     0,907       0,909     0,908      0,738    0,876     0,873

=== Confusion Matrix ===

   a   b   c   <-- classified as
  33   8   0 |   a = '(-inf--1.27109]'
   7 237   5 |   b = '(-1.27109-6.16451]'
   0   9  21 |   c = '(6.16451-inf)'
```

Linsting B.3: Log de saída para o indicador *External Balance On Goods And Services (% Of GDP)* com 2 *bins* e frequência igual.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute.Remove-R30-weka.filters.unsupervised.attribute.Discretize-F-B2-M-1.0-R24
Instances:    320
Attributes:   124
              [list of attributes omitted]
Test mode:    10-fold cross-validation

```

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Gross national expenditure (% of GDP) <= 100.01074
|   Merchandise trade (% of GDP) <= 101.41412
|   |   Population in the largest city (% of urban population) <= 33.13758
|   |   |   Time required to register property (days) <= 6.5
|   |   |   |   Business extent of disclosure index (0=less disclosure to
    10=more disclosure) <= 7.4: '(-0.549795-inf)' (4.0)
|   |   |   |   Business extent of disclosure index (0=less disclosure to
    10=more disclosure) > 7.4: '(-inf--0.549795]' (2.0)
|   |   |   Time required to register property (days) > 6.5: '(-0.549795-
    inf)' (108.03/0.52)
|   |   Population in the largest city (% of urban population) > 33.13758
|   |   |   Gross fixed capital formation (% of GDP) <= 22.01535
|   |   |   |   Food exports (% of merchandise exports) <= 53.96196:
    '(-0.549795-inf)' (27.87/1.0)
|   |   |   |   Food exports (% of merchandise exports) > 53.96196: '(-inf
    --0.549795]' (4.13/0.13)
|   |   |   Gross fixed capital formation (% of GDP) > 22.01535: '(-inf
    --0.549795]' (7.0)
|   Merchandise trade (% of GDP) > 101.41412
|   |   Public credit registry coverage (% of adults) <= 6.1: '(-inf
    --0.549795]' (7.0)
|   |   Public credit registry coverage (% of adults) > 6.1: '(-0.549795-
    inf)' (3.0)
Gross national expenditure (% of GDP) > 100.01074
|   Food exports (% of merchandise exports) <= 2.66646: '(-0.549795-inf)'
    (9.4/0.4)
|   Food exports (% of merchandise exports) > 2.66646
|   |   Gross savings (% of GDP) <= 22.58732: '(-inf--0.549795]'
    (111.13/1.15)
|   |   Gross savings (% of GDP) > 22.58732
|   |   |   Transport services (% of commercial service exports) <=
    21.94048: '(-inf--0.549795]' (25.44/1.34)
|   |   |   Transport services (% of commercial service exports) > 21.94048
|   |   |   |   Gross national expenditure (% of GDP) <= 102.27473:
    '(-0.549795-inf)' (5.0)
|   |   |   |   Gross national expenditure (% of GDP) > 102.27473: '(-inf
    --0.549795]' (2.0)

Number of Leaves  :     13

Size of the tree :      25


Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         284               89.8734 %
Incorrectly Classified Instances        32               10.1266 %
Kappa statistic                          0.7975
```

```
Mean absolute error                      0.1115
Root mean squared error                  0.3053
Relative absolute error                 22.3038 %
Root relative squared error             61.0654 %
Total Number of Instances              316
Ignored Class Unknown Instances                 4

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
                    ROC Area  PRC Area  Class
                 0,892     0,095     0,904       0,892     0,898      0,798
                    0,914     0,847     '(-inf--0.549795]'
                 0,905     0,108     0,894       0,905     0,899      0,798
                    0,914     0,914     '(-0.549795-inf)'
Weighted Avg.    0,899     0,101     0,899       0,899     0,899      0,798
    0,914     0,880

=== Confusion Matrix ===

   a   b   <-- classified as
 141  17 |   a = '(-inf--0.549795]'
  15 143 |   b = '(-0.549795-inf)'
```

Linsting B.4: Log de saída para o indicador *Industry, Value Added (% Of GDP)* com 2 *bins* e frequência diferente.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
    weka.filters.unsupervised.attribute.Remove-R76-weka.filters.unsupervised
    .attribute.Remove-R6,84,104-weka.filters.unsupervised.attribute.
    Discretize-B2-M-1.0-R74-unset-class-temporarily
Instances:    320
Attributes:   121
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Age dependency ratio (% of working-age population) <= 37.60639: '(37.99396-
    inf)' (16.0/1.0)
Age dependency ratio (% of working-age population) > 37.60639
|   Manufactures exports (% of merchandise exports) <= 13.77189
|   |   Final consumption expenditure, etc. (% of GDP) <= 75.52909:
    '(37.99396-inf)' (19.23/0.11)
|   |   Final consumption expenditure, etc. (% of GDP) > 75.52909: '(-inf
    -37.99396]' (12.57/1.0)
|   Manufactures exports (% of merchandise exports) > 13.77189
|   |   GNI per capita growth (annual %) <= 4.49773: '(-inf-37.99396]'
    (206.15)
|   |   GNI per capita growth (annual %) > 4.49773
```

```
|   |   |   Merchandise trade (% of GDP) <= 119.51691: '(-inf-37.99396]' (39.06/1.89)
|   |   |   Merchandise trade (% of GDP) > 119.51691: '(37.99396-inf)' (3.0)

Number of Leaves  :     6

Size of the tree :      11


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         279               94.2568 %
Incorrectly Classified Instances        17                5.7432 %
Kappa statistic                          0.7463
Mean absolute error                      0.0637
Root mean squared error                  0.2295
Relative absolute error                 27.0023 %
Root relative squared error             67.124  %
Total Number of Instances              296
Ignored Class Unknown Instances                24

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,973    0,250    0,961      0,973   0,967      0,747    0,771     0,901     '(-inf-37.99396]'
                 0,750    0,027    0,811      0,750   0,779      0,747    0,870     0,616     '(37.99396-inf)'
Weighted Avg.    0,943    0,220    0,941      0,943   0,942      0,747    0,784     0,862

=== Confusion Matrix ===

   a   b   <-- classified as
 249   7 |   a = '(-inf-37.99396]'
  10  30 |   b = '(37.99396-inf)'
```

Linsting B.5: Log de saída para o indicador *Industry, Value Added (Annual % Growth)* com 2 *bins* e frequência diferente.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupervised.attribute.Remove-R6,63,85,105-weka.filters.unsupervised.attribute.Remove-R5-6,38-42,45-48,83,102-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-R63
Instances:    320
Attributes:   108
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

```
9
10 === Classifier model (full training set) ===
11
12 J48 pruned tree
13 ------------------
14
15 Gross fixed capital formation (annual % growth) <= -1.94321
16 |   Household final consumption expenditure, etc. (annual % growth) <=
       -1.06454: '(-inf--1.21294]' (26.09/0.09)
17 |   Household final consumption expenditure, etc. (annual % growth) >
       -1.06454
18 |   |   Consumer price index (2010 = 100) <= 111.00833
19 |   |   |   Foreign direct investment, net inflows (% of GDP) <= 0.54228:
       '(-1.21294-inf)' (6.0/1.0)
20 |   |   |   Foreign direct investment, net inflows (% of GDP) > 0.54228
21 |   |   |   |   Foreign direct investment, net inflows (% of GDP) <=
       4.23021: '(-inf--1.21294]' (17.83)
22 |   |   |   |   Foreign direct investment, net inflows (% of GDP) > 4.23021
23 |   |   |   |   |   S&P Global Equity Indices (annual % change) <=
       -7.44441: '(-1.21294-inf)' (2.33)
24 |   |   |   |   |   S&P Global Equity Indices (annual % change) > -7.44441:
       '(-inf--1.21294]' (4.67/0.67)
25 |   |   Consumer price index (2010 = 100) > 111.00833: '(-1.21294-inf)'
       (6.29/0.17)
26 Gross fixed capital formation (annual % growth) > -1.94321
27 |   General government final consumption expenditure (% of GDP) <=
       19.76479: '(-1.21294-inf)' (187.41/2.0)
28 |   General government final consumption expenditure (% of GDP) > 19.76479
29 |   |   S&P Global Equity Indices (annual % change) <= -41.04159: '(-inf
       --1.21294]' (5.0)
30 |   |   S&P Global Equity Indices (annual % change) > -41.04159
31 |   |   |   Rural population (% of total population) <= 14.637
32 |   |   |   |   Bank capital to assets ratio (%) <= 4.7: '(-inf--1.21294]'
       (2.0)
33 |   |   |   |   Bank capital to assets ratio (%) > 4.7: '(-1.21294-inf)'
       (13.0/2.0)
34 |   |   |   Rural population (% of total population) > 14.637: '(-1.21294-
       inf)' (31.39)
35
36 Number of Leaves  :     11
37
38 Size of the tree :     21
39
40
41 Time taken to build model: 0.06 seconds
42
43 === Stratified cross-validation ===
44 === Summary ===
45
46 Correctly Classified Instances         274               90.7285 %
47 Incorrectly Classified Instances        28                9.2715 %
48 Kappa statistic                          0.6974
49 Mean absolute error                      0.1094
50 Root mean squared error                  0.2969
51 Relative absolute error                 34.2239 %
52 Root relative squared error             74.4022 %
```

```
53 Total Number of Instances               302
54 Ignored Class Unknown Instances                 18
55
56 === Detailed Accuracy By Class ===
57
58                  TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
                       ROC Area  PRC Area  Class
59                  0,717     0,045     0,796       0,717     0,754      0,699
                       0,813     0,597     '(-inf--1.21294]'
60                  0,955     0,283     0,931       0,955     0,943      0,699
                       0,798     0,871     '(-1.21294-inf)'
61 Weighted Avg.    0,907     0,236     0,905       0,907     0,905      0,699
    0,801     0,816
62
63 === Confusion Matrix ===
64
65    a   b   <-- classified as
66   43  17 |   a = '(-inf--1.21294]'
67   11 231 |   b = '(-1.21294-inf)'
```

Linsting B.6: Log de saída para o indicador *Inflation, Consumer Prices (Annual %)* com 2 *bins* e frequência diferente.

```
1 === Run information ===
2
3 Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
4 Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
    weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-R77
5 Instances:    309
6 Attributes:   125
7               [list of attributes omitted]
8 Test mode:    10-fold cross-validation
9
10 === Classifier model (full training set) ===
11
12 J48 pruned tree
13 ------------------
14
15 GNI per capita, PPP (current international $) <= 20360
16 |   Age dependency ratio, old (% of working-age population) <= 16.24988
17 |   |   Time required to register property (days) <= 9.5: '(6.326845-inf)'
    (10.0/1.0)
18 |   |   Time required to register property (days) > 9.5
19 |   |   |   Periodicity and timeliness assessment of statistical capacity (
    scale 0 - 100) <= 70
20 |   |   |   |   Gross capital formation (current US$) <= 3465623362.67075:
    '(-inf-6.326845]' (2.0)
21 |   |   |   |   Gross capital formation (current US$) > 3465623362.67075:
    '(6.326845-inf)' (9.0)
22 |   |   |   Periodicity and timeliness assessment of statistical capacity (
    scale 0 - 100) > 70
23 |   |   |   |   Gross capital formation (annual % growth) <= 19.33313
24 |   |   |   |   |   Deposit interest rate (%) <= 6.27521: '(-inf-6.326845]'
    (56.55/2.4)
25 |   |   |   |   |   Deposit interest rate (%) > 6.27521
26 |   |   |   |   |   |   Gross savings (% of GDP) <= 19.29532
```

```
|   |   |   |   |   |   |   Bank nonperforming loans to total gross loans (%) <= 2.8548: '(6.326845-inf)' (3.0)
|   |   |   |   |   |   |   Bank nonperforming loans to total gross loans (%) > 2.8548: '(-inf-6.326845]' (14.19)
|   |   |   |   |   |   Gross savings (% of GDP) > 19.29532: '(6.326845-inf)' (6.98/0.38)
|   |   |   |   Gross capital formation (annual % growth) > 19.33313
|   |   |   |   |   Broad money growth (annual %) <= 11.48299: '(-inf-6.326845]' (2.19)
|   |   |   |   |   Broad money growth (annual %) > 11.48299: '(6.326845-inf)' (6.09/0.09)
|   Age dependency ratio, old (% of working-age population) > 16.24988: '(6.326845-inf)' (14.0)
GNI per capita, PPP (current international $) > 20360: '(-inf-6.326845]' (183.0/1.0)

Number of Leaves  :     11

Size of the tree :      21


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         277               90.228 %
Incorrectly Classified Instances        30                9.772 %
Kappa statistic                          0.6473
Mean absolute error                      0.1105
Root mean squared error                  0.3011
Relative absolute error                 39.6378 %
Root relative squared error             80.8902 %
Total Number of Instances              307
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,941     0,294     0,941       0,941     0,941      0,647    0,788     0,907     '(-inf-6.326845]'
                 0,706     0,059     0,706       0,706     0,706      0,647    0,781     0,543     '(6.326845-inf)'
Weighted Avg.    0,902     0,255     0,902       0,902     0,902      0,647    0,787     0,846

=== Confusion Matrix ===

   a   b   <-- classified as
 241  15 |   a = '(-inf-6.326845]'
  15  36 |   b = '(6.326845-inf)'
```

Linsting B.7: Log de saída para o indicador *Military Expenditure (% Of Gdp)* com 3 *bins* e frequência diferente.

```
1  === Run information ===
2
3  Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
4  Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
       weka.filters.unsupervised.attribute.Remove-R39-43,46-49-weka.filters.
       unsupervised.attribute.Discretize-B3-M-1.0-R80
5  Instances:    309
6  Attributes:   116
7                [list of attributes omitted]
8  Test mode:    10-fold cross-validation
9
10 === Classifier model (full training set) ===
11
12 J48 pruned tree
13 ------------------
14
15 Population in the largest city (% of urban population) <= 45.04268
16 |   Gross national expenditure (current US$) <= 10480941661983.1
17 |   |   Time required to enforce a contract (days) <= 1210
18 |   |   |   Time required to build a warehouse (days) <= 35.4:
       '(2.468583-4.937167]' (10.0)
19 |   |   |   Time required to build a warehouse (days) > 35.4
20 |   |   |   |   Rural population growth (annual %) <= 0.75768: '(-inf
       -2.468583]' (221.93/4.0)
21 |   |   |   |   Rural population growth (annual %) > 0.75768
22 |   |   |   |   |   Cost of business start-up procedures (% of GNI per
       capita) <= 17.4: '(-inf-2.468583]' (17.0)
23 |   |   |   |   |   Cost of business start-up procedures (% of GNI per
       capita) > 17.4
24 |   |   |   |   |   |   Wholesale price index (2010 = 100) <= 84.53665: '(-
       inf-2.468583]' (3.0/1.0)
25 |   |   |   |   |   |   Wholesale price index (2010 = 100) > 84.53665:
       '(2.468583-4.937167]' (14.0)
26 |   |   Time required to enforce a contract (days) > 1210
27 |   |   |   Business extent of disclosure index (0=less disclosure to 10=
       more disclosure) <= 8: '(-inf-2.468583]' (2.97/1.0)
28 |   |   |   Business extent of disclosure index (0=less disclosure to 10=
       more disclosure) > 8: '(2.468583-4.937167]' (10.0)
29 |   Gross national expenditure (current US$) > 10480941661983.1:
       '(2.468583-4.937167]' (10.1/0.1)
30 Population in the largest city (% of urban population) > 45.04268
31 |   Business extent of disclosure index (0=less disclosure to 10=more
       disclosure) <= 5: '(-inf-2.468583]' (10.0)
32 |   Business extent of disclosure index (0=less disclosure to 10=more
       disclosure) > 5: '(4.937167-inf)' (10.0)
33
34 Number of Leaves  :     10
35
36 Size of the tree :     19
37
38
39 Time taken to build model: 0.07 seconds
40
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         290               93.8511 %
Incorrectly Classified Instances        19                6.1489 %
Kappa statistic                          0.8066
Mean absolute error                      0.0483
Root mean squared error                  0.1917
Relative absolute error                 22.1632 %
Root relative squared error             58.3667 %
Total Number of Instances              309

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC
                    ROC Area  PRC Area  Class
                 0,968     0,183     0,956       0,968     0,962       0,800
                    0,901     0,954     '(-inf-2.468583]'
                 0,780     0,023     0,867       0,780     0,821       0,790
                    0,892     0,819     '(2.468583-4.937167]'
                 1,000     0,007     0,833       1,000     0,909       0,910
                    0,997     0,833     '(4.937167-inf)'
Weighted Avg.    0,939     0,152     0,938       0,939     0,938       0,802
    0,902     0,928

=== Confusion Matrix ===

   a   b   c   <-- classified as
 241   6   2 |   a = '(-inf-2.468583]'
  11  39   0 |   b = '(2.468583-4.937167]'
   0   0  10 |   c = '(4.937167-inf)'
```

Linsting B.8: Log de saída para o indicador *Real interest rate (%)* com 2 *bins* e frequência diferente.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1-2-
    weka.filters.unsupervised.attribute.Remove-R81-weka.filters.unsupervised
    .attribute.Discretize-B2-M-1.0-R99
Instances:    320
Attributes:   124
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Transport services (% of commercial service imports) <= 52.07669
|   Time to prepare and pay taxes (hours) <= 792
|   |   Military expenditure (% of GDP) <= 0
|   |   |   Inflation, consumer prices (annual %) <= 4.50407: '(11.164855-
    inf)' (2.0)
```

```
|   |   |   Inflation, consumer prices (annual %) > 4.50407: '(-inf
    -11.164855]' (8.0)
|   |   Military expenditure (% of GDP) > 0: '(-inf-11.164855]' (158.94)
|   Time to prepare and pay taxes (hours) > 792
|   |   Food exports (% of merchandise exports) <= 17.0616: '(-inf
    -11.164855]' (13.52/1.0)
|   |   Food exports (% of merchandise exports) > 17.0616: '(11.164855-inf)
    ' (12.42/0.48)
Transport services (% of commercial service imports) > 52.07669
|   General government final consumption expenditure (annual % growth) <=
    4.07315: '(11.164855-inf)' (8.08/1.04)
|   General government final consumption expenditure (annual % growth) >
    4.07315: '(-inf-11.164855]' (4.04/0.02)

Number of Leaves  :     7

Size of the tree :      13


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         193               93.2367 %
Incorrectly Classified Instances        14                6.7633 %
Kappa statistic                          0.6291
Mean absolute error                      0.0843
Root mean squared error                  0.2557
Relative absolute error                 43.5798 %
Root relative squared error             82.924  %
Total Number of Instances              207
Ignored Class Unknown Instances                113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                    ROC Area  PRC Area  Class
                 0,968    0,364    0,957      0,968    0,962      0,630
                    0,590     0,630     '(-inf-11.164855]'
                 0,636    0,032    0,700      0,636    0,667      0,630
                    0,860     0,428     '(11.164855-inf)'
Weighted Avg.    0,932    0,328    0,930      0,932    0,931      0,630
    0,619     0,609

=== Confusion Matrix ===

   a   b   <-- classified as
 179   6 |   a = '(-inf-11.164855]'
   8  14 |   b = '(11.164855-inf)'
```

Referências

[1] Conceitos processo etl. https://danielteofilo.wordpress.com/2016/02/03/conceitos-processo-etl/. acessado em 07/02/2017. ix, 12

[2] Nações unidas no brasil. https://nacoesunidas.org/agencia/bancomundial/. acessado em 03/09/2016. 63

[3] World bank group. http://www.worldbank.org/. acessado em 03/09/2016. 1, 25

[4] O'BRIEN James A. *Sistemas De Informação E As Decisões Gerenciais Na Era Da Internet.* Editora Saraiva, São Paulo, 3 edition, 2010. 5

[5] Neeraj Bhargava, Girja Sharma, Ritu Bhargava, and Manish Mathuria. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 2013. 46

[6] Manuel. CASTELLS. *A Era da Informação: Economia, Sociedade e Cultura: A sociedade em Rede.* Editora Paz e Terra, São Paulo, 1 edition, 1999. 4

[7] L. DAVENORT, T. H.; PRUSAK. *Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual.* Editora Campus, Rio de Janeiro, 1998. ix, xi, 4, 5

[8] Peter. DRUCKER. *Desafios Gerenciais para o Século XXI.* Thompson Learning, São Paulo, 1999. 3

[9] R. Navathe. S. e Elmasri. *Sistemas de Banco de Dados.* Prentice Hall, São Paulo, 6 edition, 2010. 8, 9

[10] Nonaka e Takeuchi. *Criação de Conhecimento na Empresa.* Editora Campus, Rio de Janeiro, 1997. 7

[11] Fabrício Augusto. Ferrari. *Crie um banco de dados em MYSQL.* Digeratti Books, São Paulo, 1 edition, 2007. 8, 9

[12] André Ponce de Leon; FACELI Katti; LORENA Ana Carolina; OLIVEIRA Marcia. GAMA, João; CARVALHO. *Extração de conhecimento de dados: data mining.* Edições Sílabo, 2 edition, 2015. ix, 17

[13] J. Han and M. Kamber. *Data mining : concepts and techniques.* Kaufmann, San Francisco, 2005. 11, 15, 25

[14] E. Frank I. H. Witten. *Data Mining: Practical Machine Learning Tools and Technique.* Morgan Kaufmann, 2 edition, 2005. 19

[15] E. Frank I. H. Witten and M. A. Hell. *Data Mining: Practical Machine Learning Tools and Technique.* Morgan Kaufmann, 3 edition, 2011. 15, 19, 21, 42

[16] Willian H INMON. *DW 2.0: The Architecture for the Next Generation of Data Warehousing.* Morgan Kaufmann, Massachusetts, 1 edition, 2008. 11

[17] SANTOS. Maribel Yasmina; RAMOS. Isabel. *Business Intelligence : tecnologias da informação na gestão de conhecimento.* FCA Editora de Informática, Lisboa, 2006. 12

[18] Micheline Kamber Jiawei Han and Jian Pei. *Data mining : concepts and techniques.* Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA, 3 edition, 2012. 1

[19] Dessloch S. Jorg, T. *Towards generating ETL processes for incremental loading.* IDEAS, 2008. 14

[20] Caserta J. Kimball, R. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* John Wiley Sons, 2004. 13

[21] Ralph. Kimball. *The Data Warehouse Lifecycle Toolkit: pratical techniques for building dimensional data Warehouse.* John Wiley Sons, 1996. 10, 11

[22] A. Korth, H.F. e Silberschatz. *Sistemas de Banco de Dados.* Makron Books, 2 edition, 1994. 8

[23] Jane Price. LAUDON, Kenneth C.; LAUDON. *Sistemas de Informação.* Editora LTC, Rio de Janeiro, 4 edition, 1999. 7

[24] BRAGA. Everaldo Miranda. *Mineração de Dados de Posição Geográfica e Compras.* Departamento de Ciência da Computação,Universidade de Brasília, Brasília, 2012. 15

[25] J Ross Quinlan. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90, 1996. 46

[26] M. Ross W. Thornthwaite. R. Kimbal, L. Reeves. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying DataWarehouses.* John Wiley Sons, 1998. 11, 12

[27] Denis Alcides. REZENDE. *Tecnologia da informação aplicada a sistemas de informação empresariais.* Editora Atlas S.A, São Paulo, 4 edition, 2006. 5

[28] J. Minds Searle. *Brains Science: the 1984 Reith Lectures.* Penguin Books, New York, 1991. 6

[29] V.W. Setzer. Dado, informação, conhecimento e competência. *Os Meios Eletrônicos e a Educação: Uma Visão alternativa. Datagrama*, 10, 2001. 5, 7

[30] P. Smyth U. M. Fayyad, G. Piatetsky-Shapiro and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ix, 16

[31] WEKA Weka. 3: data mining software in java. *University of Waikato, Hamilton, New Zealand (www. cs. waikato. ac. nz/ml/weka)*, 2011. 19, 52