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

Mineração de dados socioeconômicos do WorldBank usando Python e a biblioteca Scikit-Learn

Luthiery Costa Cavalcante
Fernando Ferreira Cordeiro

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

Orientador
Prof. Dr. Jan Mendonça Correa

Brasília
2024

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

Mineração de dados socioeconômicos do WorldBank usando Python e a biblioteca Scikit-Learn

Luthiery Costa Cavalcante
Fernando Ferreira Cordeiro

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

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

Prof. Dr. Donald Knuth
Stanford University

Dr. Leslie Lamport
Microsoft Research

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 10 de Setembro de 2024

Dedicatória

Dedico esse trabalho aos meus pais por sempre me apoiarem no meu trajeto acadêmico, e que em nenhum momento deixaram de me incentivar a seguir os meus sonhos.

- Fernando Cordeiro

Dedico este trabalho à Luzia e ao Luiz, meus pais. Ao Lindberg e ao Lucas, meus irmãos. E também à Lianna, à Allana e ao Arthur, sobrinhos que tenho com tanto carinho. Todos vocês, tendo noção disso ou não, me inspiram e me impedem de desistir.

- Luthiery Costa Cavalcante

Agradecimentos

Nos *agradecimentos*, o autor se dirige a pessoas ou instituições que contribuíram para elaboração do trabalho apresentado. Por exemplo: *Agradeço aos gigantes cujos ombros me permitiram enxergar mais longe. E a Google e Wikipédia.*

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

resumo: fazer por ultimo

Palavras-chave: Python, Pandas, Scikit-Learn, Aprendizado de Máquina, Mineração de Dados, WorldBank, Banco Mundial

Abstract

abstract: last thing to do

Keywords: Python, Pandas, Scikit-Learn, Machine Learning, Data Mining, WorldBank

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AM Aprendizado de Máquina.

API *Application Programming Interface.*

BI *Business Inteligence.*

CART *Classification and Regression Trees.*

CIC Departamento de Ciência da Computação.

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

CSV *Comma-separated Values.*

DSA *Data Staging Area.*

DW *Data Warehouse.*

ETL *Extract, Transform, Load.*

FN Falso Negativo.

FP Falso Positivo.

IA Inteligência Artificial.

KDD *Knowledge Discovery in Databases.*

KNN *K-Nearest Neighbors.*

ML *Machine Learning.*

PCA Análise de Componentes Principais.

PIB Produto Interno Bruto.

PLN Processamento de Linguagem Natural.

RNA Rede Neural Artificial.

SGBD Sistema Gerenciador de Banco de Dados.

SI Sistema de Informação.

TI Tecnologia de Informação.

UnB Universidade de Brasília.

VN Verdadeiro Negativo.

VP Verdadeiro Positivo.

WDI Indicadores de Desenvolvimento Mundial.

Capítulo 1

Introdução

1.1 Problema

É de grande interesse social e acadêmico, em áreas como Economia, Ciências Sociais e Políticas Públicas, assim como na Ciência de Dados em particular, a criação de métodos de obtenção de conhecimento preditivo sobre a tendência geral de desenvolvimento de um determinado território. Existe um senso comum sobre quais são os fatores que mais contribuem para o crescimento ou decrescimento da riqueza de um Estado ou território. No entanto, com milhares de indicadores socioeconômicos disponíveis e aferidos sobre esses Estados ao longo do tempo, é um problema recorrer ao senso comum para tentar responder esse questionamento - para isso, mostra-se necessário o uso de análise e mineração de dados desses indicadores.

1.2 Objeto

O Banco Mundial [1] é uma instituição financeira internacional que oferece empréstimos e assistência técnica a países em desenvolvimento. Fundado em 1944, o Banco Mundial tem como objetivo reduzir a pobreza e promover o desenvolvimento sustentável em todo o mundo. Ele trabalha em parceria com governos, organizações não governamentais e setor privado para implementar projetos nas áreas de saúde, educação, infraestrutura, meio ambiente e economia.

O Banco Mundial disponibiliza dados de desenvolvimento global por várias interfaces dentro de sua plataforma: o **DataCatalog** (usada no projeto), onde é possível baixar bases de dados inteiras em lote, em diversos formatos, a **API de Dados** e o **DataBank**. Todos esses métodos, e o motivo para escolha do primeiro, são mais elaborados na seção 4.2.1.

Entre muitas bases de dados disponibilizadas pelo Banco Mundial, usaremos a de **Indicadores de Desenvolvimento Mundial** ou *World Development Indicators* (WDI). Trata-se de uma fonte abrangente de informações socioeconômicas e demográficas de países em todo o mundo [2]. As bases de dados do Banco Mundial são amplamente utilizadas no apoio à decisão por pesquisadores, formuladores de políticas e profissionais da área econômica.

1.3 Justificativa

Propomos este trabalho como uma continuidade de trabalhos anteriores que se propunham à descoberta de padrões com mineração de dados usando o mesmo conjunto de ferramentas (vide seção 4.1) e/ou a mesma base de dados objeto. Nesses trabalhos, em particular os que abrangiam bases de dados com muitos atributos, era comum o treinamento dos modelos a partir da seleção manual das características mais relevantes, a cargo dos autores, em detrimento de uma seleção algorítmica.

1.4 Objetivos

Demonstrar o poder da linguagem de programação Python em conjunto com a biblioteca Scikit-Learn para a análise de dados descritiva, modelagem com Aprendizado de Máquina e seleção de atributos. Em outras palavras, demonstrar sua adequação para as mais diversas tarefas de mineração de dados, incluindo particularmente o caso de tarefas de mineração sobre bases de dados com muitos registros e características.

1.4.1 Objetivos específicos

- Agrupar os dados coletados a partir da aplicação web do Banco Mundial em uma base de dados única;
- Realizar tratamento de valores vazios na base, através da exclusão e inferência de valores, para deixá-la mais rica;
- Reduzir a quantidade de atributos da base selecionando os mais importantes para a modelagem de forma algorítmica;
- Por fim, construir um modelo capaz de aproximar, ou predizer, o indicador (atributo) de "Crescimento anual do PIB (em porcentagem)"a partir do conjunto de treinamento com os indicadores remanescentes.

1.5 Metodologia

O estudo de caso envolve, inicialmente, a coleta dos dados contendo países e seus indicadores a partir da aplicação web do Banco Mundial, de forma manual. Por meio das ferramentas abordadas na seção 4.1, serão construídos *scripts* capazes de:

- unir os arquivos coletados em um único arquivo, que será então exportado para uma base de dados única;
- realizar o pré-processamento de dados;
- selecionar os atributos mais relevantes;
- construir um modelo de Aprendizado de Máquina (AM) com fim de realizar predições conforme os objetivos do trabalho;
- gerar gráficos para a visualização da base de dados e do modelo gerado;

Os *scripts* possuem parâmetros definidos nas suas linhas iniciais, que impactam a etapa de pré-processamento e a geração do modelo. O modelo pode ser gerado algumas vezes, com parâmetros diferentes, a fim de comparação de resultados.

Os gráficos gerados, assim como a avaliação numérica (*score*) da qualidade dos modelos gerados, serão utilizados para avaliar o resultado do modelo e fazer conclusões sobre as hipóteses levantadas.

1.5.1 Hipóteses

Propomos as seguintes hipóteses como forma de expandir o entendimento sobre a capacidade de modelagem das ferramentas, assim como expandir o conhecimento já obtido pelos estudos relacionados:

Hipótese 1: Os modelos gerados pelos algoritmos do Scikit-Learn possibilitam uma valoração numérica da sua qualidade; a derivação de uma grande quantidade de informação sobre a base de dados; o reconhecimento de padrões a partir da visualização do modelo gerado; e a seleção e visualização dos atributos mais importantes.

Hipótese 2: Utilizar a biblioteca Scikit-Learn para selecionar (filtrar) atributos de forma algorítmica, em uma base de dados com uma quantidade suficientemente grande de atributos, é mais eficiente ao analista e gera um modelo melhor do que aquele gerado com uma seleção manual de atributos.

1.6 Estrutura do Trabalho

Os próximos capítulos deste trabalho estão estruturados da seguinte forma:

Capítulo 2 : Abrange definições úteis para o entendimento do trabalho a partir de pesquisa bibliográfica.

Capítulo 3 : Discorre sobre trabalhos similares encontrados na bibliografia que se relacionam com o presente trabalho em seus objetivos e objetos de estudo.

Capítulo 4 : O corpo do trabalho, indicando o passo-a-passo do desenvolvimento, desde a coleta de dados até a aferição dos resultados das modelagens.

Capítulo 5 : É onde fazemos uma análise e avaliação sobre o conhecimento obtido com o estudo.

Capítulo 6 : Onde refletimos sobre a qualidade do próprio estudo e retomamos as hipóteses para concluir se as mesmas foram observadas.

Ao fim, nos apêndices, estão inclusos todos os *scripts* confeccionados neste estudo, assim como informação adicional sobre a base de dados objeto.

Capítulo 2

Revisão Teórica

Diversos conceitos relacionados à mineração de dados possuem grau de importância para o entendimento do projeto e precisam ser contextualizados para o entendimento deste trabalho, em uma ordem lógica. Na seção 2.1, é abordada a própria definição de dado e como obter conhecimento a partir de dados. As seções 2.1.2 e 2.2 discorrem sobre estruturas de armazenamento e agrupamento de dados, na forma de Bancos de Dados e *Data Warehouses*. A seção 2.3 define o conceito de Mineração de Dados, seus processos e algumas de suas aplicações mais comuns. A seção 2.4 aborda a teoria inicial sobre Inteligência Artificial, que serve de base para a formulação do Aprendizado de Máquina (AM), descrito na seção 2.5. Nessa mesma seção, são contextualizados e diferenciados o aprendizado supervisionado do não-supervisionado, exemplos de algoritmos em cada categoria - em particular os que serão utilizados no estudo de caso - assim como o conceito de redução de dimensionalidade.

2.1 Dado, Informação e Conhecimento

Castro e Ferrari [3] fornecem definições sucintas para esses conceitos: **dados** são símbolos ou sinais não estruturados, sem significado inerente; a **informação** é obtida nas descrições ou manipulações sobre esses dados, agregando significado e utilidade; já o **conhecimento** é algo que permite uma tomada de decisão para agregar valor de negócio ao que está sendo trabalhado, a partir da interpretação das informações obtidas.

Dados são sequências de fatos ainda não analisados - armazenados em um formato arbitrário, seja uma tabela em um banco de dados, uma planilha, um caderno físico etc. - representativos de eventos que ocorrem em um certo contexto organizacional, antes de terem sido organizados e dispostos de forma que as pessoas possam entendê-los e usá-los. Informação é um conjunto de dados que já foi organizado e modelado em um formato significativo útil para as pessoas. A transformação de dado em informação pode ser feita

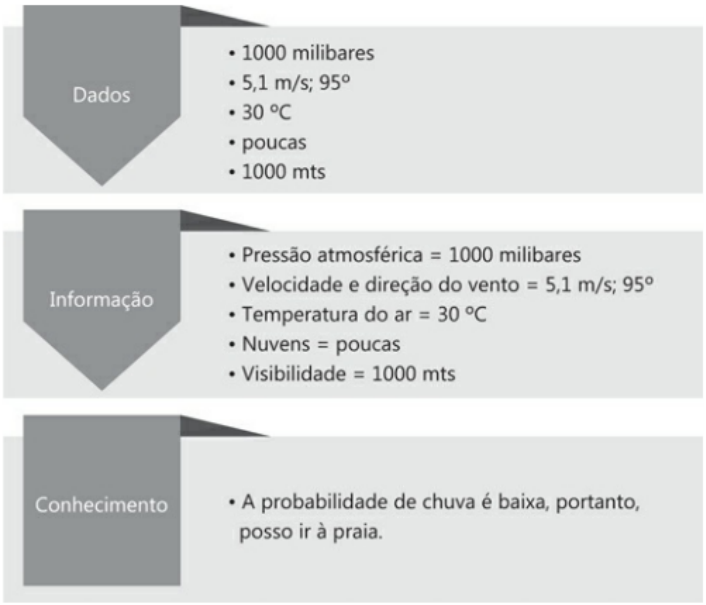

Figura 2.1: Exemplo abstrato de transformação de dado em conhecimento (Fonte: [3]).

através de um Sistema de Informação (SI) [4]. A geração de conhecimento é a etapa (geralmente) humana de analisar e interpretar as informações de forma que seja útil para as decisões de negócio de uma organização ou, em um contexto de análise experimental, para a descoberta de padrões na base de dados, por exemplo.

Elmasri e Navathe [5] fornecem definições complementares: a de conhecimento *dedutivo*, que deduz novas informações com base na aplicação de regras lógicas previamente especificadas sobre os dados; e a de conhecimento *indutivo*, que descobre novas regras e padrões a partir dos dados.

2.1.1 Sistemas de Informação

Em um contexto organizacional, um Sistema de Informação (SI) pode ser definido como um conjunto de componentes inter-relacionados que coletam, processam, armazenam e distribuem informações destinadas a apoiar a tomada de decisões, a coordenação e o controle em uma organização. Além disso, os sistemas de informação também auxiliam os gerentes e trabalhadores a analisar problemas, visualizar assuntos complexos e criar novos produtos. Um SI também contém informações sobre pessoas, locais e itens significativos para a organização ou ambiente que os cerca [4].

Há três atividades em um SI responsáveis por gerar as informações necessárias às decisões, operações, análise de problemas e o desenvolvimento de novas soluções [4]. São elas:

entrada: que coleta ou recupera dados brutos dentro de uma organização ou de seu ambiente externo;

processamento: que converte, manipula e mapeia esses dados em uma forma mais significativa (informação);

saída: que transfere as informações processadas às pessoas que as utilizarão ou às atividades nas quais serão empregadas.

O termo Sistema de Informação (SI) por vezes é confundido com o de Tecnologia de Informação (TI). Embora os SI informatizados utilizem a tecnologia de computadores para processar dados brutos e transformá-los em informações inteligíveis, existe uma diferença entre hardware e software, de um lado, e um SI, de outro. Os computadores são os equipamentos que armazenam e processam as informações. Os programas de computador ou software são conjuntos de instruções operacionais que dirigem e controlam o processamento do computador. Eles são apenas parte de um Sistema de Informação. Além do emprego de Tecnologia de Informação, os SI compreendem também uma natureza organizacional e humana [4].

2.1.2 Bancos de dados

Elmasri e Navathe [5] definem um **banco de dados** (tradução de *database*, do inglês) como uma coleção de dados relacionados logicamente, estruturados com algum significado inerente, o qual é projetado, construído e populado para uma finalidade específica. Um banco de dados tem alguma fonte da qual se derivam os dados presentes, um grau de interação fidedigna com eventos do mundo real (esses eventos compõem o que são chamados de *minimundo* ou *universo de discurso* do banco de dados) e um público ativamente interessado em seu conteúdo, que pode consultá-lo ou fazer alterações no mesmo.

Bancos de Dados podem ser gerados e mantidos manualmente (em uma agenda, por exemplo, onde o processo de alteração ou inclusão de dados é lento) ou de forma computadorizada [5]. No mundo digital em que vivemos, o termo *banco de dados* quase sempre se referem àqueles controlados de forma computadorizada através de um Sistema Gerenciador de Banco de Dados (SGBD) - eles permitem com que os bancos de dados cresçam exponencialmente em volume de dados armazenados sem que isso impossibilite o trabalho de consulta e alteração dos mesmos.

Um Sistema Gerenciador de Banco de Dados é definido como uma coleção de programas que facilita os processos de:

Definição do banco de dados - especificar os tipos, estruturas, domínios e restrições dos dados a serem armazenados, especificação que também é armazenada no SGBD com o nome de *metadados*;

Construção do banco de dados - especificar o armazenamento dos dados em um meio controlado pelo SGBD;

Manipulação do banco de dados - inclui funcionalidades de consulta para recuperar, atualizar (por meio de transações) ou gerar relatórios relativos a esses dados;

Compartilhamento do banco de dados entre diversos usuários, programas e sistemas da web de forma simultânea;

Proteção do sistema de banco de dados contra falhas de hardware (geralmente por meio de redundância) assim como proteção de segurança (por meio de controle de acesso);

Manutenção do sistema, permitindo que o banco de dados tenha um ciclo de vida longo e evolua conforme seus requisitos de negócio mudam.

Um outro termo que vemos ser usado por autores como Castro e Ferrari [3] é o de **base de dados**, definida pelos mesmos como: "coleção organizada de dados, ou seja, valores quantitativos ou qualitativos referentes a um conjunto de itens, que permite uma recuperação eficiente dos dados- definição bastante similar à do primeiro parágrafo desta seção. Ainda segundo os autores, "conceitualmente, os dados podem ser entendidos como o nível mais básico de abstração a partir do qual informação e, depois, conhecimento, podem ser extraídos", processo ilustrado na Figura 2.1.

Na bibliografia de bancos de dados, autores como Elmasri e Navathe [5] e Silberschatz et al. [6] não se preocupam em fazer distinção entre os termos *banco de dados* e *base de dados* - de fato usando apenas o primeiro. Nessa concepção, base de dados seria um banco de dados que, todavia, não está arquitetado sob nenhum SGBD, tendo uma estrutura arbitrária (mais flexível) [1] . Para os que fazem a distinção entre os termos, um banco de dados estaria intrinsecamente ligado ao uso de SGBD, sendo necessário usar outro termo para coleções lógicas digitais de dados que não estão estruturados sob eles [7].

O termo base de dados costuma ser mais utilizado em contextos de análise exploratória de dados e artigos relacionados aos conceitos de *Big Data* (vide Cavique [8]) e Inteligência Artificial (IA), nos quais o conhecimento a ser extraído desses dados - e como extrair - é a

[1] A confusão na terminologia é agravada já que *base de dados* também é uma tradução aceitável para *database*. No contexto deste trabalho, onde não são usados SGBDs, a diferença conceitual entre os termos acaba não sendo relevante. Portanto, eles serão usados de forma sinônima, podendo ainda, no lugar, ser usado o termo *conjunto de dados*.

parte crucial, independente da estrutura em que os mesmos estão implementados. Alguns exemplos de artigos nesse escopo são os trabalhos do Capítulo 3.

2.2 Data Warehouse

Um armazém de dados - ou *Data Warehouse* (DW) - oferece armazenamento, funcionalidade e responsividade a consultas além das capacidades de um banco de dados tradicional. Surgiu como consequência do crescente poder de processamento e a sofisticação das ferramentas e técnicas analíticas [5], motivados pela tecnologia de armazenamento dos SGBDs, o advento das aplicações online e a "bagunça"que essa inundação de dados criou nas organizações [9].

Existe uma distinção clara entre um banco de dados tradicional e um *Data Warehouse*: ambos são coleções logicamente coerentes de informações, acompanhadas por softwares de suporte e gerenciamento. Contudo, os bancos de dados são **transacionais** - sejam eles com implementação relacional, orientada a objetos, em rede ou hierárquica - enquanto os DW servem principalmente para aplicações de **apoio à decisão**, sendo otimizados para recuperação de dados, em vez de processamento de transações de rotina [5].

Enquanto SGBDs relacionais são otimizados para processar consultas que envolvem uma parte pequena do banco de dados e transações que lidam com inserções e atualizações em uma relação (tabela) por vez, *Data Warehouse*s são projetados para dar suporte à extração, processamento e apresentação eficientes para fins analíticos e de tomada de decisão. Eles contêm quantidades muito grandes de dados de várias fontes, que podem incluir até bancos de dados de diferentes modelos, assim como arquivos adquiridos de sistemas e plataformas independentes [5].

Dessa forma, Elmasri e Navathe [5] caracterizam os *Data Warehouse* como uma coleção de tecnologias de apoio à decisão, visando a habilitar o trabalhador do conhecimento (p.ex. executivo, gerente, analista) a tomar decisões melhores e mais rápidas. Porém, não existe uma única definição canônica para esse termo. Outros autores, como Inmon et al. [9] definem um *Data Warehouse* como uma coleção de dados para processamento de informações, a qual, além de cumprir a função de apoio à decisão, deve estritamente possuir as seguintes propriedades:

Integrado : os dados são reunidos no DW a partir de diversas fontes, de naturezas e formatos arbitrários, agrupados a fim de gerar uma consistência coerente;

Orientado a assunto : esse agrupamento traz informações sobre um domínio específico e determinado;

Não-volátil : não deve haver mutabilidade nos dados. Dessa forma, são adicionados e agregados novos dados sem alterar os registros anteriores, funcionando como um registro temporal para a organização;

Variável ao longo do tempo : Os dados são identificados a cada período determinado de tempo.

2.2.1 Business Inteligence

Vários autores, como Kimball et al. [10] colocam os *Data Warehouse*s como parte fundamental e inseparável da Inteligência de Negócio, ou *Business Inteligence* (BI).

De acordo com Santana e Carvalho [11], O BI abrange conjuntos de ferramentas e técnicas que se dedicam à obtenção, análise, organização, compartilhamento e monitoramento de dados dentro dos bancos de dados, oferecendo suporte à decisão e à gestão de negócios. Essas tecnologias são capazes de suportar uma enorme quantidade de dados, mesmo desestruturados.

Segundo Kimball et al. [10], o termo BI teria começado a ser usado nos anos 1990, referindo-se à construção de relatórios e à análise dos dados guardados nos armazéns - no início, existiam inúmeras organizações que usavam os *warehouses* como meros depósitos de registros brutos, sem consideração ao processo de entregar dados e conhecimentos organizados aos usuários do negócio de forma útil, levando à cunhagem desse novo termo para se referir à entrega de valor por meio dos DWs.

Na indústria, alguns tratam o BI como uma parte (ou etapa) do *data warehousing*, que seria o processo como um todo, enquanto há outros que fazem o contrário: consideram os DWs a camada central de dados e processos relacionados, contidos dentro de uma estratégia maior de *Business Inteligence*. Por isso, Kimball et al. [10] propõem o termo conjunto DW/BI para se referir a todo o conjunto, reforçando sua interdependência; todos os dados recuperáveis no sistema DW/BI estão no *Data Warehouse* do negócio, enquanto as ferramentas de análise e geração de valor são *aplicações de BI*.

2.2.2 ETL

Uma das bases que estruturam o *data warehousing*, ou DW/BI, é o *Extract, Transform, Load* (ETL) - inglês para Extração, Transformação e Carga. Em uma analogia, Kimball et al. [10] definem o ETL como a "cozinha"que mantém o "restaurante"do *BI* em funcionamento. Dados brutos são extraídos das fontes de dados organizacionais e levados à "cozinha", onde eles são transformados em informação útil ao negócio. Tal espaço precisa ser arquitetado bem antes de haverem dados para trafegar; ele é projetado para entregar o maior vazão de dados possível; também é desenhado para entregar um produto final com

a maior qualidade possível; os insumos que chegam a ele devem ser checados quanto à sua integridade e qualidade; e os próprios critérios de qualidade são frequentemente revisados.

Santana e Carvalho [11] descrevem os processos de extração, transformação e carga, da seguinte forma:

Extração : Rotinas de extração são executadas sobre as diversas fontes de dados, podendo ser sistemas de bancos de dados, sistemas operacionais ou até planilhas e arquivos de texto. Procura-se a capacidade de ler e extrair dados em diversos formatos, além de integrá-los, sem causar perda de informações em relação às fontes originais.

Transformação : Os dados extraídos são propagados para a chamada *Data Staging Area* (DSA), onde os dados serão manipulados sem a necessidade de consultas às bases originais. A transformação consiste em adequar os dados às necessidades e restrições do modelo DW em questão, de forma a garantir a qualidade, consistência e limpeza dos dados.

Carga : Rotinas de carga são responsáveis por entregar os dados já integrados ao *Data Warehouse* em si, respeitando restrições de integridade e criando uma visão concreta unificada dos dados extraídos.

É necessário notar que não serão usadas aplicações de DW/BI no presente trabalho, optando-se por uma estratégia de armazenamento de dados mais direta e simples usando arquivos em formato CSV, e realização de manipulação sobre essa massa usando scripts na linguagem Python. Contudo, no tratamento inicial da massa de dados objeto são aplicadas técnicas análogas a um ETL (processo detalhado na seção 4.2.1), de forma que se fez útil a definição de todos os conceitos relacionados dentro desta seção.

2.3 Mineração de dados

Mineração de Dados é o processo que tem como objetivo descobrir padrões e tendências em grandes conjuntos de dados [12] através de algoritmos e outros sistemas de informação. O termo é uma alusão ao conceito "original"de mineração, que é a extração de minerais valiosos, como ouro e pedras preciosas, a partir de uma fonte - uma mina. No contexto computacional, é um campo vasto e multidisciplinar que envolve áreas como bancos de dados, estatística, Aprendizado de Máquina (AM), engenharia, Sistemas de Informação (SI), Redes Neurais Artificiais (RNA), processamento de sinais e visualização de dados [3], entre outras.

No contexto histórico mundial atual, há uma quantidade enorme e exponencialmente crescente de informação sendo coletada por sistemas corporativos, governamentais, entre muitos outros. Entretanto, segundo Larose e Larose [12], essa informação não necessariamente se converte inteiramente em conhecimento, devido à relativa escassez de analistas de dados no mercado. A demanda por profissionais especializados em dados foi criada, segundo os mesmos autores, por essa oferta gigante de dados coletados, a evolução da tecnologia em sistemas de bancos de dados, o aumento no volume de acesso a esses sistemas por vários usuários simultaneamente e a própria evolução do poder de processamento e de armazenamento dos computadores atuais.

2.3.1 KDD

A mineração de dados está englobada no processo de Descoberta de Conhecimento em Bancos de Dados (ou *Knowledge Discovery in Databases* (KDD)), como uma de suas etapas. O processo de KDD sempre recebe um banco de dados e, a partir de regras de negócio e objetivos específicos dos analistas, submetem essa base a uma série de processos para obter um determinado conjunto de conhecimentos ao final. Sejam as seis etapas do KDD, a seguir, descritas por Elmasri e Navathe [5]:

Seleção de dados : filtragem da amostra de registros (linhas) [2] e/ou atributos (colunas) [3] específicos;

Limpeza de dados : onde ocorre o tratamento de atributos inválidos e/ou exclusão de registros com dados incorretos (segundo as regras de negócio determinadas);

Enriquecimento : melhora a base de dados, por exemplo, inserindo atributos processados (derivados) a partir de outros atributos e fontes de dados externas ao banco;

Transformação de dados : pode ser usado para reduzir a quantidade de dados, por meio do mapeamento (ou codificação) de um domínio extenso de valores de um atributo em um domínio menor.

Mineração de dados : etapa em que os dados são submetidos a técnicas que processam esses dados com o objetivo de descobrir padrões e obter conhecimento (algumas técnicas serão detalhadas na seção 2.3.3);

Exibição : etapa onde o conhecimento obtido é organizado, condensado e exposto em um formato logicamente agradável ao público interessado nos conhecimentos obtidos.

[2] Os termos "registro", "instância", "tupla", "dado", "objeto"e "linha"são usados de forma sinônima e intercambiável na bibliografia.

[3] Os termos "atributo", "característica", "variável"e "coluna"são usados de forma sinônima e intercambiável na bibliografia.

2.3.2 CRISP-DM

O *Cross-Industry Standard Process for Data Mining* (CRISP-DM) é um padrão multi-industrial, independente de ferramentas e aplicações, desenvolvido por analistas para representar as etapas da mineração de dados. Segundo Larose e Larose [12], esse padrão serve o objetivo de resolver problemas dentro de uma organização (ou projeto de pesquisa) através do uso de mineração de dados, sendo um padrão não-proprietário e gratuito.

Um projeto de mineração de dados qualquer, segundo o CRISP-DM, tem um ciclo de vida iterativo e adaptativo baseado em seis etapas. Isso significa que a decisão de qual será a próxima etapa por vezes depende do resultado da etapa anterior - por vezes, os projetos precisam voltar a etapas anteriores para serem refinados antes de prosseguir para os próximos passos.

Assim seguem as etapas do CRISP-DM para projetos de mineração de dados [12]:

Entendimento do negócio (ou pesquisa) : Enunciar os objetivos, escopo e requisitos do projeto em termos de decisões negociais ou dos pesquisadores. A partir disso, traduzir tais requisitos para uma definição de um problema de mineração de dados e preparar uma estratégia preliminar para o cumprimento dos objetivos.

Entendimento dos dados : Coletar os dados em si. Em seguida, usar análise exploratória de dados para se familiarizar com os mesmos. Por fim, avaliar a qualidade dos dados e, se for o caso, detectar subconjuntos de dados que sejam mais úteis aos objetivos negociais.

Preparação dos dados : Aqui entram todos os aspectos de preparação da base de dados final, a partir da coleta bruta inicial dos dados. É necessário selecionar casos, variáveis e subconjuntos que são mais apropriados para a análise; realizar transformações em certas variáveis (atributos) para tornar mais fácil a manipulação; e filtrar o conjunto de dados para maximizar a utilidade ferramentas de modelagem.

Modelagem : Escolher, aplicar e calibrar as técnicas de modelagem que serão usadas no problema, de forma a otimizar os resultados. Por vezes, várias técnicas podem ser aplicadas em um mesmo processo de mineração (aqui entra a natureza adaptativa dos projetos - a base de dados pode precisar voltar à fase de preparação para se adaptar a cada modelo particular).

Avaliação : Os *modelos* gerados na etapa anterior precisam ser avaliados segundo sua qualidade, acurácia e efetividade. É necessário também certificar-se que o modelo atende os requisitos definidos na 1ª etapa. Caso não os atenda, pode ser necessário voltar e escolher outra técnica de modelagem.

Implantação : O modelo gerado não é a resposta do problema, mas sim o uso desse modelo para gerar conhecimento - essa é a etapa final chamada implantação. Um exemplo muito comum é gerar e expor relatórios sobre os resultados obtidos. Outro exemplo seria usar esse modelo como insumo para outros projetos relacionados de mineração de dados.

2.3.3 Tarefas em mineração de dados

As tarefas - ou funcionalidades - da mineração de dados especificam os tipos de informações que podem ser obtidas pela mineração. Segundo Castro e Ferrari [3], as tarefas podem ser separadas em duas categorias: descritivas e preditivas. As descritivas dispõem sobre propriedades gerais da base de dados e dos seus atributos; as preditivas fazem inferência a partir dos dados almejando fazer predições de valores. Frequentemente, esse segunda categoria de tarefas está associada ao uso de algoritmos de Aprendizado de Máquina (exemplificados na seção 2.5). Em muitos casos, é incerto ao analista a escolha de qual funcionalidade aplicar ao problema de mineração, tornando importante a capacidade dessas ferramentas de se adaptar e encontrar conhecimento por vários caminhos.

Castro e Ferrari [3] e Han et. al [13] descrevem as principais tarefas de mineração de dados:

Análise descritiva de dados

É uma funcionalidade que não requer nível elevado de complexidade algorítmica. Consiste de ferramentas capazes de medir e descrever características intrínsecas aos dados. Particularmente, obtém-se medidas estatísticas como a distribuição de frequência, medidas de centro (média, mediana e moda), variância, medidas de posição relativa e associação dos dados, além de técnicas elementares para visualização de dados como a *plotagem*[4] de gráficos.

As análises descritivas permitem sumarizar e compreender os objetos presentes na base de dados e seus atributos. É possível, a partir de um exemplo (ou registro) da base, determinar o quão semelhante ele é da média de todos os registros - e outras deduções estatísticas. Também é possível detectar padrões a partir da visualização dos registros em gráficos de diversos tipos, como histogramas (distribuições de frequência).

[4]Termo técnico que, na Ciência de Dados e outras áreas relacionadas à Ciência da Computação, se refere ao ato de gerar gráficos de forma algorítmica e parametrizada a partir de uma fonte de dados.

Predição - classificação e regressão

Predição é a terminologia usada para se referir à construção de modelos preditivos - seja para determinar o rótulo (ou classe[5]) de um objeto (registro) não-rotulado, no qual a tarefa é chamada de classificação, ou para estimar o valor de um ou mais atributos de um objeto, no qual a tarefa é denominada regressão ou estimação. Temos nessas duas tarefas os principais tipos de problemas de predição, sendo que a classificação é usada para predizer valores discretos (não necessariamente numéricos), enquanto a regressão é usada para predizer valores contínuos (comumente numéricos).

Um exemplo simples usado por Castro e Ferrari [3] para distinguir os dois tipos de tarefas é o de crédito pessoal: um cliente se dirige a uma instituição bancária com objetivo de obter um financiamento. O cliente, dessa forma, é um objeto (ou registro) da base de dados da instituição; usa-se a classificação para determinar se ele deve ou não ter acesso a crédito (particularmente, trata-se de uma classificação *binária*, com apenas dois valores possíveis); além de um modelo de regressão para determinar qual o valor, em dinheiro, que tal cliente deve obter.

O uso de mineração de dados neste projeto, que visa responder perguntas sobre a base de dados do Banco Mundial, consiste em um problema de regressão. Na seção 2.5, veremos que os modelos criados a partir de Aprendizado de Máquina para resolver problemas de classificação e regressão se encaixam dentro da categoria chamada de aprendizado *supervisionado*.

Agrupamento

O agrupamento (em inglês *clustering*) é o nome dado à tarefa de separar, particionar ou segmentar um conjunto de objetos em grupos (ou *clusters*), que são montados com base nas semelhanças entre os atributos dos objetos. Um *cluster* pode ser definido como um subconjunto de objetos similares uns aos outros e diferentes dos objetos pertencentes a outros *clusters*. O objetivo é determinar classes de objetos similares, separados em subconjuntos.

Na seção 2.5, veremos que os modelos feitos com Aprendizado de Máquina para executar tarefas de agrupamento se encaixam dentro da categoria chamada de aprendizado *não-supervisionado*.

[5] Ambos os termos "rótulo", "classe", "alvo"e "classe-alvo"são usados de forma sinônima e intercambiável na bibliografia

Associação

Em tarefas preditivas e de agrupamento, o objetivo é encontrar relações - sejam de similaridade, classes ou estimativas - entre objetos de uma base. Entretanto, em tarefas de associação, o objetivo é determinar relações entre os próprios atributos (variáveis) presentes na base, ao invés de seus registros (tuplas).

A análise por associação, também chamada de mineração de regras de associação, se encarrega da descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em um conjunto de dados, ou seja, atributos que podem ser determinados por (ou variam em função de) outros atributos.

Há dois aspectos centrais para a construção dessas regras: a proposição (construção) eficiente das regras e a quantificação da significância das mesmas. Um bom algoritmo de mineração de regras de associação precisa ser capaz de propor associações que sejam estatisticamente relevantes dentro do universo representado pela base de dados.

Detecção de anomalias

Objetos ou tuplas que não seguem um comportamento ou característica comum entre os dados da base, ou que possuem valores de atributos muito díspares da média, são chamados de anomalias, exceções, ruído, valores discrepantes ou *outliers*. Como eles geralmente não conseguem ser representados pelo modelo gerado, a maioria das ferramentas de mineração os descarta - caso contrário, a capacidade de predição, associação ou descrição do modelo pode ser impactada negativamente. Há outras tarefas, no entanto, que têm o objetivo de ressaltar (e não deletar) as anomalias encontradas, como em algoritmos de detecção de fraude bancária.

As anomalias podem ser detectadas de várias formas, incluindo métodos estatísticos que obtém uma distribuição ou modelo de probabilidade dos dados, ou medidas de distância por meio das quais objetos muito distantes dos demais são considerados anomalias.

2.3.4 Pré-processamento de dados

O pré-processamento de dados, ou tratamento de dados, abrange tarefas que se enquadram dentro das fases de Entendimento dos Dados e de Preparação dos Dados dentro do padrão CRISP-DM [12].

Muitos dos dados brutos coletados nos bancos de dados são incompletos e contém informação ruidosa que não será aproveitada. Por exemplo, essas bases podem conter: campos que são obsoletos e redundantes; valores faltantes; *outliers*; dados em formato não apropriado ou compatível com os modelos de modelagem de dados; e valores inconsistentes com o observado na realidade ou fora de conformidade com as regras de negócio [12]. O

pré-processamento de dados é necessário para resolver tais problemas e atribuir **qualidade** aos dados. A qualidade de um banco de dados consiste de propriedades como a acurácia, completude, consistência, temporalidade, credibilidade e interpretabilidade [13].

A seguir veremos técnicas como a remoção ou inferência de valores vazios, redução do ruído nos dados, identificação de *outliers* e correção de inconsistência, que são definidas como técnicas de **limpeza de dados** [3].

Valores vazios

Um valor ausente numa base de dados constitui em um valor que, por algum motivo, foi ignorado ou não foi observado no momento de coleta. Ele costuma ser representado por um código de ausência, que pode ser um valor específico, como um caractere específico (por ex., `"?"`), o número 0, uma implementação de `NaN`[6] ou um valor em branco [3].

É necessário à maioria dos algoritmos de modelagem, uma grande massa de dados para seu correto funcionamento. Ao mesmo tempo, entretanto, é frequente encontrar uma grande proporção de valores ausentes - particularmente em grandes bancos de dados, com muitos dados e atributos [12].

Muitos dos algoritmos de modelagem não funcionam corretamente ao se deparar com valores ausentes, sendo assim é necessário algum tratamento - os mais frequentes são a remoção de dados (seja de registros ou de atributos) e também a inferência ou imputação de dados, que consiste em preencher os valores ausentes seguindo alguma fórmula consistente, a fim de manter o banco de dados populado com valores significantes [3].

A imputação de valores ausentes assume que a ausência de valores implicam em perda de informação relevante para a eficácia da mineração. Assim, os valores que serão inferidos não devem somar nem subtrair informação da base - isto é, não deve enviesá-la [3]. Já a remoção de tuplas ou de atributos com valores ausentes, segundo Han et al. [13], é uma técnica trivial, não muito benéfica visando a eficácia do algoritmo, visto que pode acabar sacrificando dados válidos úteis consigo. Ela pode ser útil, no entanto, para descartar atributos que sabidamente não interferirão no resultado da mineração.

São estratégias comuns para a inferência ou imputação de dados [3] [13]:

- Preencher valores ausentes manualmente, segundo o conhecimento do próprio analista - tarefa inviável em cenários de grandes bases de dados;
- Preencher todos os valores ausentes com um único valor constante, como o número 0: método simples, mas que pode prejudicar o resultado da mineração ao analisar tal valor como tendo um significado;

[6] Abreviação de "Not a Number", conceito computacional para a representação de números não definidos ou inválidos.

- Preencher valores com alguma medida de tendência central (como a média ou mediana) correspondente a cada atributo;
- Preencher valores usando medidas de tendência central dos registros que possuem a mesma classe do registro em questão;
- Preencher valores usando como base os valores válidos dos objetos mais similares - a similaridade pode ser medida usando uma comparação categórica ou medidas de distância entre objetos (método chamado de imputação *hot-deck*).
- Preencher valores a partir da última observação - atribui os valores válidos logo anteriores encontrados na base (parte da premissa que a base está ordenada sob algum critério lógico)
- Preencher valores ausentes usando o valor mais *provável* - para isso, são empregadas diversas técnicas preditivas como regressão, inferência estatística a partir da teoria Bayesiana ou indução com árvores de decisão. Esse é o método mais popular, por usar uma maior quantidade de informação sobre a própria base para inferir valores.

Dados ruidosos

Problemas na coleta de dados podem também gerar dados com erros. Frequentemente, esses erros se tornam parte indissociável dos dados e não podem ser removidos facilmente. O acúmulo desses erros e distorções, além da presença de *outliers*, que são dados fiéis ao observado porém discrepante dos padrões, é chamado de ruído [3] [13].

Existem métodos para "suavizar"os dados na base e reduzir o ruído, apesar de não haver um padrão consistente que permita a identificação dos mesmos, o que ocasiona um mínimo irredutível de ruído [3]. Alguns deles são:

- Encaixotamento *(binning)*: distribuir os valores de um atributo em "caixas", de forma que elas tenham o mesmo tamanho, ou abranjam um mesmo intervalo;
- Agrupamento *(clustering)*: encontrar grupos de objetos similares e passar a referir-se aos objetos pelo seu grupo, ou por um objeto específico do mesmo;
- Aproximação: aproximar os dados por alguma função ou modelo paramétrico.

Dados inconsistentes

A falta da propriedade de consistência, em uma base de dados, influencia na validade, utilidade e integridade de uma aplicação de mineração de dados. Um exemplo é o uso de nomes diferentes para se referir a uma mesma característica, que, aos olhos do algoritmo

de mineração, seriam tratados como rótulos diferentes. Outro exemplo ocorre quando valores apresentados não condizem com o domínio dos atributos dados. Soluções frequentes envolvem rotinas manuais desenvolvidas pelos analistas responsáveis [3].

Além das técnicas de limpeza de dados vistas acima, são parte do pré-processamento as tarefas de **integração de dados** (que consistem em unir dados oriundos de diversas fontes e assegurar a integridade e consistência dessa junção, observando a remoção de redundâncias, duplicidades e conflitos), assim como as de **redução de dados** e **redução de dimensionalidade** [3] [13], que serão mais detalhadas na seção 2.5.4.

2.4 Inteligência Artificial

A Inteligência Artificial (IA) é a área da Ciência da Computação que se propõe, não apenas a compreender a **inteligência** do ser humano, mas a construir entidades que consigam emular o processo humano do pensamento e, junto dele, toda sua capacidade de perceber, compreender, prever e manipular o mundo ao seu redor [14].

Nota-se que, ao longo da história da IA, existe uma divergência entre acadêmicos a respeito da definição dessa área do conhecimento: se deve ser o estudo dos sistemas que pensam (ou agem) *como seres humanos*; ou que pensam (ou agem) *racionalmente*. A diferença é sutil - pensar racionalmente (i.e. de forma ideal, ótima ou perfeita) pode por vezes ser antônima ao comportamento humano, que é imperfeito. Uma abordagem centrada nos seres humanos deve ser em parte uma ciência empírica e cognitiva, com aspectos biológicos e neuropsicológicos, hipóteses e confirmação experimental; uma abordagem racionalista, por sua vez, envolve uma combinação de lógica, matemática e engenharia [14].

Um dos trabalhos pioneiros da IA, elaborado entre os anos 1940 e 1950, foi o *Teste de Turing*, construído por Alan Turing, no qual ele formula uma definição objetiva para *inteligência*. O teste consiste em submeter um programa computacional a uma conversa por texto com uma pessoa - o interrogador - que lhe faz perguntas. Tal programa passa no teste e, portanto, é classificado inteligente, se o interrogador não conseguir distinguir se teve uma conversa com uma pessoa ou com um computador [14].

Para passar no Teste de Turing, um programa (i.e. máquina), precisaria possuir algumas capacidades. É possível traçar uma correspondência entre essas capacidades e as principais sub-áreas de conhecimento (ou disciplinas) dentro da Inteligência Artificial atualmente [14]:

Processamento de Linguagem Natural - o programa precisa se comunicar com sucesso em um idioma natural, logo precisa saber tanto perceber quanto emitir mensagens;

Representação de conhecimento - para possibilitar o armazenamento das informações percebidas;

Raciocínio automatizado - para usar informações com a finalidade de responder a perguntas e tirar novas conclusões; e

Aprendizado de Máquina - que dá ao programa a habilidade de se adaptar a novas circunstâncias, assim como detectar e extrapolar padrões.

Trabalhos posteriores ao de Turing evoluíram o teste para considerar ainda a dimensão física de uma máquina que se passa por humano [14]. Além de saber entender e escrever se comunicar por texto, no chamado *Teste de Turing Total* o programa precisa possuir capacidades extras de:

Visão Computacional - que confere à maquina a capacidade de perceber objetos físicos; e da

Robótica - que a possibilita se movimentar e manipular objetos conforme solicitado.

Tendo em mente o escopo da aplicação da IA no presente projeto, será trabalhado mais a fundo particularmente o conceito do Aprendizado de Máquina (AM), também conhecido em inglês como *Machine Learning (ML)*, na seção 2.5.

2.5 Apredizado de Máquina

De acordo com Russell e Norvig [14], diz-se que um agente (de qualquer natureza) **aprende** quando se torna capaz de melhorar seu desempenho em tarefas de predição de padrões e comportamentos após fazer observações sobre o mundo, construindo um conhecimento interno através do processamento de suas próprias percepções anteriores.

No contexto da computação e da inteligência artificial, os agentes inteligentes podem ser empregados em problemas complexos de predição de comportamentos (por exemplo, tarefas de mineração de dados como descritas na seção 2.3.3, graças aos seus poderes de processamento, armazenando conhecimento para utilizá-los em suas diferentes aplicações.

Em um problema de mineração de dados, o mapeamento explícito entre todas as entradas e todas as saídas possíveis é inviável ou mesmo impossível, para muitas aplicações na vida real. Isso pode ocorrer por vários motivos: os conjuntos de entradas e/ou saídas podem ser infinitos ou suficientemente grandes a ponto de inviabilizar uma representação direta; mesmo que seja possível representar todo o domínio e contradomínio do problema, frequentemente o projetista não conhece a função que o resolve, pois determiná-la foge da

capacidade humana em tempo viável; além disso, projetar um programa para um conjunto exaustivo de cenários removeria sua capacidade de aprendizado e adaptação [14].

Géron [15], em raciocínio parecido, sumariza problemas e aplicações ideais para o uso de Aprendizado de Máquina: resolver problemas onde as soluções existentes exigem muita configuração manual, listas extensas de regras de inferência e códigos longos; problemas complexos para os quais não existe uma boa solução conhecida com a abordagem tradicional; problemas em ambientes flutuantes que constantemente tem os dados alterados; e problemas com um enorme volume de dados. Muitas vezes, os problemas apresentam várias dessas características ao mesmo tempo.

Esses algoritmos devem ser capazes de inferir conhecimento sobre o conjunto de entradas para que ele aprenda a resolver o problema da forma esperada independentemente da entrada que lhe for dada ao longo do tempo, delegando-lhes a tarefa de aprender a resolver problemas complexos, ao invés de descrever explicitamente as regras da solução [14] [15].

Posto isso, Géron [15] retoma duas definições amplamente utilizadas para o Aprendizado de Máquina (AM):

> [Aprendizado de Máquina é o] campo de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programado. [16]

> Diz-se que um programa de computador aprende pela experiência E em relação a algum tipo de tarefa T e alguma medida de desempenho P se o seu desempenho em T, conforme medido por P, melhora com a experiência E. [17]

Uma das primeiras soluções mundialmente aplicadas de Aprendizado de Máquina foram os filtros de *spam* [15], capazes de ler o conteúdo de um e-mail dentro da caixa de entrada de um usuário e determinar se ele deve ser marcado ou não como *spam*. Nesse exemplo simples, aplicando a definição de Mitchell [17], a tarefa **T** seria atribuir ou não o rótulo de *spam*; a experiência **E** seria o dado de treinamento (o conjunto de textos de e-mail que ele armazenou anteriormente, ou que foi entregue a ele para processar); e a medida de desempenho **P** poderia ser, por exemplo, a proporção entre e-mails marcados corretamente (dentro da categoria *spam* ou fora dela) e o total de e-mails processados. Essa medida específica é chamada de *acurácia* [15] e será aprofundada, junto com outras métricas, na seção 2.5.5.

Os algoritmos de Aprendizado de Máquina podem ser classificados de acordo com o tipo de supervisão que recebem durante o treinamento. Existem quatro categorias principais de aprendizado: supervisionado, não supervisionado, semi-supervisionado e por reforço [15]. Dependendo de sua categorização, os algoritmos podem ser considerados mais adequados a tarefas específicas de mineração de dados (vide seção 2.3.3).

2.5.1 Aprendizado Supervisionado e Não-Supervisionado

Sistemas de Aprendizado de Máquina precisam de conjuntos de dados de treinamento para adquirir experiência - é a partir deles que o sistema adquire conhecimento para fazer predições com novos dados. A diferença entre as duas abordagens consiste nas informações presentes nesse conjunto de treinamento.

Aprendizado Supervisionado

No aprendizado supervisionado, esses dados de treinamento fornecidos incluem as soluções desejadas - ou seja, cada tupla (registro) tem seu rótulo (classe) conhecido(a) [15].

Retomando a seção 2.3.3, é sabido que as tarefas de classificação e de regressão são típicos problemas de aprendizado supervisionado. O filtro de *spam* é um bom exemplo disso [15]: ele é treinado com muitos exemplos de e-mails onde a informação de que se trata de *spam* ou não já é sabida; a partir desse conhecimento obtido, ele deve ser capaz de classificar corretamente novos e-mails. Nessa mesma seção, o problema da concessão de crédito pessoal por uma instituição bancária é um problema de regressão onde o algoritmo precisa ser treinado com rótulos de clientes anteriores (no caso, o valor do crédito concedido), para então ser capaz de atribuir um valor de crédito a clientes futuros.

Alguns dos algoritmos mais importantes do *estado da arte* da literatura de aprendizado supervisionado são citados por Castro e Ferrari [3], categorizados por sua estrutura:

- os baseados em conhecimento (regras lógicas);
- os baseados em distância, como o *k-Nearest-Neighbors (KNN)*
- os conexionistas, tendo como grande exemplo a Rede Neural Artificial (RNA) e suas diversas variações e evoluções;
- os baseados em funções (parametrizadas);
- os probabilísticos, como o *Naive Bayes*; e
- os baseados em árvores, como as Árvores de Decisão e as Florestas Aleatórias (*Random Forests*).

Essa categorização se aplica tanto aos problemas de classificação quanto aos de regressão. Existem problemas de regressão implementados com algoritmos conexionistas, com algoritmos estatísticos paramétricos, com modelos baseados em árvores, entre outros. A modelagem do problema deste trabalho é feita por uma Floresta Aleatória. Sendo assim, detalharemos os dois algoritmos baseados em árvores supracitados nas seções 2.5.2 e 2.5.3, respectivamente.

Aprendizado Não-supervisionado

No aprendizado não-supervisionado, os dados de treinamento não possuem rótulo (classe). Logo, o algoritmo não deve prever rótulos a partir de rótulos conhecidos e, em vez disso, tenta obter conhecimento sobre a base de dados de forma autônoma - geralmente, descobrindo semelhanças e correlações dentro desses dados.

Segundo Géron [15], as tarefas de mineração mais comumente realizadas com esses algoritmos são as de agrupamento ou *clustering* (com os algoritmos *k-Means* e *Clustering Hierárquico*, por exemplo), mineração de regras de associação (com algoritmos como *Apriori* e *Eclat*) e a de redução de dimensionalidade, tendo como principal expoente o algoritmo Análise de Componentes Principais (PCA), descrito mais detalhadamente na seção 2.5.4.

2.5.2 Árvores de Decisão

As árvores de decisão são algoritmos versáteis de Aprendizado de Máquina supervisionado capazes de executar tarefas de classificação, regressão ou mesmo tarefas *multioutput*[7] [15]. Uma árvore de decisão é uma estrutura de fluxograma em formato de árvore, que possui nós e ramificações, representando o fluxo da classificação. Cada nó interno (que não é folha, ou seja, que possui ramificações saindo do mesmo) denota um teste condicional sobre um conjunto de atributos; cada ramificação representa uma saída (ou caminho) possível a partir da resposta dos testes; e cada nó-folha (ou terminal) representa a atribuição de um rótulo [13].

Para cada instância de uma base de dados, ela deve ser inserida na árvore a partir do nó-raiz (na profundidade 0, marcando o início do algoritmo). Então, tal instância deve ser confrontada com a árvore de decisão, passando sucessivamente por todos os nós (testes condicionais) até que chegue em um nó-folha de maior profundidade e receba um rótulo. Cada nó pode testar apenas um atributo, como no caso da Figura 2.2, ou vários deles [13]. É intuitivo concluir que instâncias com valores suficientemente similares seguirão o mesmo caminho pela árvore, sendo classificadas com o mesmo rótulo, e vice-versa.

Existem cenários onde, em uma mesma base de treinamento, tuplas que possuem o mesmo rótulo podem seguir caminhos diferentes na árvore. A árvore é construída através do exemplo dos dados da base, por isso tal base deve ser rica e variada, abrangendo todos os cenários. Se dentro desse banco de dados houverem "buracos", a construção da árvore e predições serão deficitárias [12]. Pode, ainda, haver ruído na base que impacta o formato da árvore. Para resolver isso, podem ser realizadas as técnicas de "poda"(*pruning*) [13].

[7] Em alguns casos, pode ser necessário que o classificador atribua várias classes para uma mesma instância. Essas tarefas denominam-se tarefas *multilabel*, essas múltiplas classes são binárias, ou *multioutput*, no caso generalizado [15].

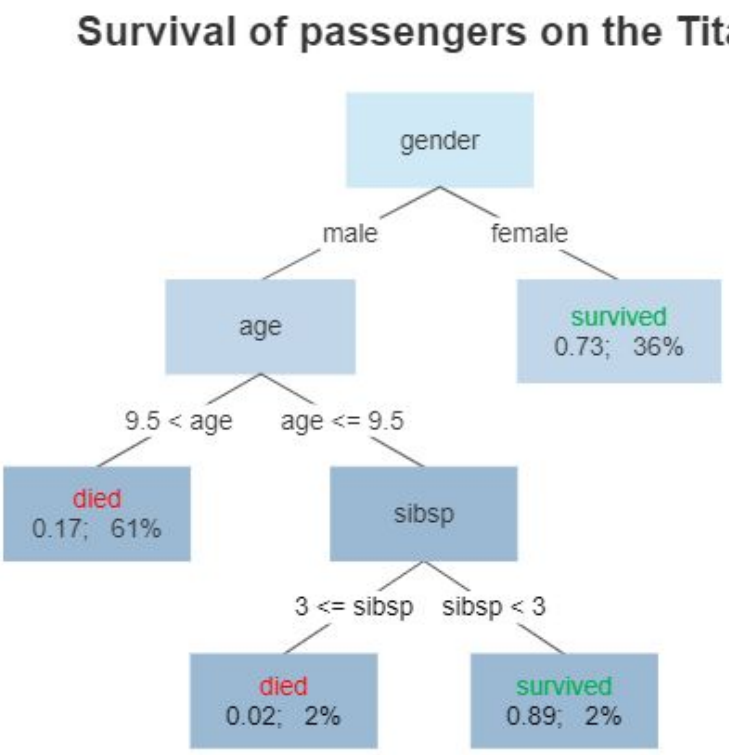

Figura 2.2: Exemplo de uma árvore de decisão aplicada na base de dados do Titanic, com a tarefa de determinar se uma pessoa passageira sobreviveu a partir dos seus dados (em inglês) (Fonte: [18, 19]).

Com respeito à construção das árvores, destacam-se os algoritmos `ID3`, `C4.5` e *Classification and Regression Trees* (CART). Tais algoritmos adotam uma estratégia gulosa (*greedy*) por divisão e consquista, construindo árvores de forma indutiva. Esses algoritmos (particularmente o CART), assim como os critérios usados para a criação dos nós das árvores, são explicados em detalhe por Han et al. [13] e Larose e Larose [12]. O CART é o algoritmo utilizado pelo *Scikit-Learn* para treinar suas árvores de decisão [15].

2.5.3 Florestas Aleatórias

Floresta Aleatória (ou *Random Forest*) é um tipo de método de agregação (*ensemble*). Os *ensembles* consistem de uma combinação de k modelos de AM (chamados de classificadores-base) $M_1, M_2, ..., M_k$. Cada M_k é treinado com um subconjunto D_k distinto da base de dados D. O modelo resultante M' executa a previsão para uma nova tupla com base em "votos"dos k modelos - cada modelo processa e atribui uma classe para

a tupla e, após isso, seguindo algum critério (por ex., por voto da maioria), se obtém a classe resultante [13].

Usar um conjunto de classificadores/estimadores geralmente resulta em modelos mais eficazes do que a utilização de um modelo individual - não por acaso, as florestas aleatórias são alguns dos algoritmos mais poderosos de AM disponíveis atualmente [15]. Métodos *ensemble* performam melhor quando existe uma diversidade significativa nos dados - i.e. com pouca correlação entre os modelos-base [13] [15].

Uma Floresta Aleatória, como o nome pode sugerir, é um conjunto de árvores de decisão. Cada árvore M_i recebe uma base D_i derivada da base de dados original D. Os conjuntos de tuplas D_i são geralmente feitos com o método de *bagging*, que consiste em criar amostras aleatórias com substituição[8] do mesmo tamanho de D. Cada árvore, em cada nó, seleciona aleatoriamente uma quantia bem inferior de atributos em relação aos atributos originais, para determinar as decisões e ramificações. Ás árvores são construídas com o algoritmo CART. O desempenho de uma floresta aleatória depende da eficácia das árvores individuais, em conjunto com a medida de correlação (dependência) entre os mesmos [13].

Em caso de problemas de classificação, um critério comum para se definir a classe resultante de uma tupla, após ser classificada por todas as árvores individuais, é a votação simples majoritária. Já em problemas de regressão, é comum utilizar uma média simples dos valores retornados por cada árvore [13].

Uma grande qualidade das florestas aleatórias é facilitar a medição da importância relativa de cada atributo. É possível, na ferramenta *Scikit-Learn*, medir a importância de uma característica analisando o quanto os nós de árvores que a utilizam reduzem a impureza, em média [15].

2.5.4 Redução de Dimensionalidade

Muitos problemas de Aprendizado de Máquina envolvem milhares (como no caso da base de dados deste projeto) ou mesmo milhões de características para cada objeto ou instância de treinamento. À primeira vista, parece um ponto positivo por agregar mais informações à base de dados. No entanto, isso não apenas torna o treinamento extremamente lento, mas também dificulta encontrar uma boa solução. Esse problema é referido como a **maldição da dimensionalidade**. Felizmente, é possível transformar um problema insolúvel (em tempo viável) em um problema tratável por algoritmos de AM ao reduzir consideravelmente o número de características abordadas [15].

[8] A técnica de *bagging* (*bootstrap aggregation*) constrói amostras com substituição, o que significa que uma tupla, após incluída na amostra, tem a mesma chance de ser selecionada novamente. Isso pode ocasionar a duplicação de tuplas nas amostras, e a omissão de outras. Tais amostras são chamadas de amostras *bootstrap* [13].

O aumento do número de objetos e, principalmente, de dimensões, podem fazer com que os dados fiquem esparsos e as medidas matemáticas e estatísticas usadas na análise se tornem instáveis. Além disso, como esperado, uma base muito grande em tuplas e atributos pode tornar os algoritmos de aprendizado, assim como os modelos gerados pelos mesmos, muito complexos [3].

Em contextos como o de *Big Data*, entre outros, onde o volume de dados gerado é grande, dinâmico e não-estruturado, a resposta para contornar a base e reduzir a complexidade dos algoritmos pode estar na redução de dimensionalidade dos dados [8].

Castro e Ferrari [3] destacam alguns métodos de redução de dimensionalidade de dados, dos quais o primeiro é particularmente relevante nesse projeto:

Seleção de Atributos (ou *Feature Selection*): técnica que remove atributos pouco relevantes ou redundantes.

Compressão de atributos: emprega técnicas de codificação (ou transformação) de dados, resultando em uma nova base com atributos mais relevantes, em vez de seleção.

Redução de dados: os dados (tanto tuplas quanto atributos) podem ser removidos, estimados ou substituídos por representações menores, como modelos paramétricos (que armazenam apenas parâmetros em vez dos dados em si) e os não-paramétricos, como o agrupamento, a amostragem e o uso de histogramas.

Discretização: os valores de atributos são substituídos por intervalos ou níveis conceituais de abstração mais elevados, reduzindo a quantidade final e o domínio dos atributos.

Ao reduzir a dimensionalidade, perdemos informação (por exemplo, em uma compressão de arquivo de imagem). Embora acelere o treinamento, também pode fazer com que o modelo, o aprendizado e as predições funcionem de forma pior [15]. É importante que os métodos de redução de dimensionalidade de dados preservem a integridade e características dos dados originais, de forma a manter a eficácia da tarefa de mineração, produzindo modelos e resultados igualmente confiáveis com menos esforço computacional [3].

Seleção de Atributos

A Seleção de Atributos ou *Feature Selection* é uma técnica que reduz o tamanho do conjunto de dados ao remover atributos redundantes ou irrelevantes. Seu objetivo é encontrar um conjunto mínimo de atributos tal que a distribuição de probabilidade dos rótulos seja a mais próxima possível da distribuição de probabilidade original com todos os atributos [13].

Para n atributos, existem 2^n subconjuntos de atributos possíveis, o que tornaria inviável uma busca exaustiva pelo melhor subconjunto quando n é muito grande. Assim, a abordagem mais utilizada envolve métodos heurísticos que trabalham em uma parte restrita dos dados: de forma *forward*, onde o conjunto de atributos inicia vazio e iterativamente o "melhor"atributo fora do conjunto é incluído no mesmo; de forma *backward*, onde os "piores"atributos são removidos sucessivamente do conjunto original de atributos; ou uma junção de ambos os métodos. O critério de parada desses algoritmos pode se basear no tamanho do subconjunto (quando o usuário quiser selecionar um número k de atributos) ou quando a métrica usada atingir um certo limiar [13].

Mais importante que o funcionamento do algoritmo de seleção, entretanto, é saber a definição de "melhor"e de "pior". Segundo Han et al. [13] essas métricas envolvem o uso de testes estatísticos (como de variância e correlação) ou cálculos de ganho de informação (entropia).

As métricas providas pela biblioteca *Scikit-Learn*, assim como a decisão estratégica de qual número k de atributos selecionar na etapa de *feature selection* desse projeto são abordadas na seção 4.6.

Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é um dos métodos mais populares de compressão de atributos encontrados no estado-da-arte do Aprendizado de Máquina. É um procedimento estatístico que converte um conjunto de objetos com presença de alguma correlação em um outro conjunto de objetos com variáveis linearmente descorrelacionados, os *componentes principais* [3].

O número de componentes principais é sempre menor ou igual ao número de atributos originais; a transformação é feita de forma que os componentes resultantes tenham a maior variância possível, realizando um mapeamento linear dos dados (chamado de projeção) em um espaço de dimensão menor [3]. Se os dados originais puderem ser reconstruídos com a aplicação inversa do algoritmo, sem perda de informação, houve uma compressão "sem perda"*(lossless)*; caso contrário, a compressão foi "com perda"*(lossy)* [13].

Essa projeção, que essencialmente transforma a base em uma nova base com novos atributos, difere da técnica de seleção de atributos, à medida em que não preserva os valores originais. Em um experimento, Traskas [20] se propõe a ilustrar a diferença entre os algoritmos PCA e *feature selection* na performance de modelos de predição - aplicando ambos na etapa de pré-processamento dos dados. O estudo se baseia nas respectivas implementações dos algoritmos na biblioteca *Scikit-Learn*.

2.5.5 Avaliação de desempenho

A avaliação do desempenho de um algoritmo de Aprendizado de Máquina corresponde à aferição da qualidade da modelagem. Determina se o modelo aproxima a solução ideal ou, em outras palavras, se desenvolveu aprendizado. É a última etapa do processo de construção e aplicação de um modelo preditivo , pertencendo à etapa de Avaliação dentro do processo CRISP-DM (vide a seção 2.3.2). As medidas de desempenho se propõem em responder o quão bem o modelo generalizará para dados fora da base de treinamento. No caso do aprendizado supervisionado, elas são baseadas em cálculos de acerto e erro entre a saída fornecida pelo modelo e a saída desejada [3].

Avaliando o desempenho de classificadores: a Matriz de Confusão

O desempenho de um algoritmo de classificação depende de sua flexibilidade (*bias*) e da qualidade do treinamento (variância). A forma mais comum de avaliar é simplesmente calcular o percentual de classificação correta, mais conhecida como **acurácia**, assim como seu complemento, o **erro**. Por padrão, a acurácia não considera o custo de uma predição incorreta - qualquer erro, para qualquer classe, possui o mesmo peso [3]. A Acurácia, apesar de ser uma medida específica, também se refere genericamente à qualidade de predição de um modelo [13].

Os problemas binários (com dois valores possíveis para a classe-alvo) são um caso particular de grande interesse dentro dos problemas de classificação. Vários problemas reais podem ser mapeados em classificação binária, como a concessão de crédito ou o classificador de spam [3]. O alvo recebe o nome de classe positiva para o rótulo `Verdadeiro`, e de classe negativa para o rótulo `Falso`. A partir dessas definições, são definidos alguns conceitos os quais são atribuídos a tuplas, e que constroem a matriz de confusão de um classificador binário [3]:

- Verdadeiro Positivo (VP): objeto de classe positiva classificado como `Verdadeiro`;
- Verdadeiro Negativo (VN): objeto de classe negativa classificado como `Falso`;
- Falso Positivo (FP): objeto de classe negativa classificado como `Verdadeiro` - conhecido como "alarme falso"ou Erro do Tipo 1;
- Falso Negativo (FN): objeto de classe positiva classificado como `Falso` - conhecido como Erro do Tipo 2;

A partir desses conceitos , definem-se as principais métricas de avaliação [3] [13]:

		Classe predita	
		Positiva	**Negativa**
Classe original	**Positiva**	VP	FN
	Negativa	FP	VN

Tabela 2.1: Matriz de confusão de um classificador binário

$$Acc = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

(Acurácia)

$$E = 1 - Acc = \frac{FP + FN}{VP + FP + VN + FN} \tag{2.2}$$

(Erro)

$$Pr = \frac{VP}{VP + FP} \tag{2.3}$$

(Precisão)

$$Re = \frac{VP}{VP + FN} \tag{2.4}$$

(*Recall*, Revocação, Sensitividade)

$$Esp = \frac{VN}{VN + FP} \tag{2.5}$$

(Especificidade)

$$F_1 = \frac{2 * Pr * Re}{Pr + Re} \tag{2.6}$$

(*F-score*, Medida F)

Validação cruzada

Na aprendizagem supervisionada, os modelos precisam atingir um equilíbrio (conhecido como o *Dilema bias-variância*), de forma a serem flexíveis para se aproximarem ao máximo da solução com os dados de treinamento disponíveis, de forma a evitar o *underfitting*[9] e o *overfitting*[10].

A validação cruzada é o método sistemático mais comum usado para atingir esse equilíbrio. Ela consiste em executar testar o modelo em um subconjunto de dados de

[9]Diz respeito à incapacidade e inflexibilidade do modelo de se adaptar aos dados de treinamento. É também conhecido como erro de representação [3].

[10]Ocorre quando o modelo é treinado em excesso e absorve ruídos da base de treinamento, tornando-se ineficaz para a predição de novos dados. Chamado também de erro de generalização [3].

teste não usados no treinamento. Com o passar das iterações, se o erro nas predições do conjunto de teste começar a aumentar de forma consistente, é a hora mais indicada para parar o treinamento . Outra finalidade da validação cruzada é a de aferir o desempenho dos algoritmos [3].

Um algoritmo comum é a *validação cruzada em k-pastas*, que consiste em dividir a base de dados em k subconjuntos, de forma que, em k iterações, o subconjunto i ($i \in 1...k$) será usado para teste, enquanto os outros $k-1$ são usados para treinamento. O método da separação de pastas influencia o resultado final, por isso é comum que o algoritmo acima seja executado k vezes, cada vez com uma estratégia diferente de separação [3].

Desempenho de modelos de regressão

Tarefas de classificação podem ser vistas como casos particulares de regressão, nas quais a saída é discreta (ou categórica). Assim, as métricas vistas e o conceito de validação cruzada são válidos em ambos tipos de modelo. A saída de um estimador é um valor numérico contínuo o qual se deseja ser o mais próximo possível do valor observado. A diferença entre os dois valores fornecem medidas de erro de estimação do algoritmo [3].

Para cada tupla j em uma base de dados com n tuplas, seja d_j o valor desejado (real), y_j o valor estimado pelo modelo, e $e_j = d_j - y_j$ o erro observado. A partir disso definem-se várias métricas de natureza contínua [3] [12]. De maior relevância ao projeto, temos:

$$SSE = \sum_{j=1}^{n} (e_j)^2 \tag{2.7}$$

(Soma dos erros quadráticos)

$$MSE = \frac{1}{n} * \sum_{j=1}^{n} (e_j)^2 \tag{2.8}$$

(Erro quadrático médio)

$$s = \sqrt{\frac{1}{n} * \sum_{j=1}^{n} (e_j)^2} \tag{2.9}$$

(Raiz do erro quadrático médio; erro padrão da estimativa)

$$R^2 = 1 - \frac{\sum_{j=1}^{n} (e_j)^2}{\sum_{j=1}^{n} (d_j - \overline{d})^2} \tag{2.10}$$

(Coeficiente de determinação)

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são listadas as referências acadêmicas que foram analisadas e revisadas para refinar e embasar tema, escopo, hipótese e objetivo deste projeto.

Nos últimos tempos, um enorme número de estudos foram criados com o intuito de aplicar a mineração de dados nos mais variados universos de dados. Neste trabalho, nos dedicamos a analisar os objetivos e resultados de quatro estudos em particular, explicados respectivamente nas seções 3.1, 3.2, 3.3 e 3.4.

Esses estudos são trabalhos de graduação recentes de graduandos do Departamento de Ciência da Computação (CIC) da UnB, cujo orientador também foi o profº Jan Mendonça Corrêa. De certa forma, este trabalho de conclusão de curso pode ser visto como uma tentativa de continuidade da pesquisa e hipóteses apresentadas nesses trabalhos - em especial os apresentados na seção 3.3 e 3.4, onde há similaridades, respectivamente, com relação a objeto e ferramentas usadas.

Diferenças notáveis em relação a esses estudos incluem a decisão estratégica de adotar uma IDE analítica e controle de versão para o armazenamento e visualização dos dados em vez de utilizar persistência de dados com Sistema Gerenciador de Banco de Dados (SGBD), além da ausência do uso de ferramentas de mineração de dados externas ao ecossistema *Python* como *Weka*, *Pentaho* e *Orange*, por entender que as ferramentas listadas na seção 4.1 continham todos os recursos computacionais e analíticos necessários ao estudo.

3.1 Hirano e Beserra (2018)

Hirano e Beserra [21] se propuseram a analisar os dados públicos de votações do Congresso Nacional do Brasil, em particular do Senado brasileiro. Segundo os mesmos, esses dados valiosos de interesse da sociedade civil, apesar de publicados on-line, se encontravam em um formato pouco amigável e sem tratamento. Dessa forma, aplicaram mineração de

dados em busca da obtenção de conhecimento por meio de informações e padrões, de forma a colaborar para o acesso público à informação.

Para isso, utilizaram ferramentas como o *Weka*, *Pentaho*, *Orange* e um SGBD específico: o *MySQL*, aplicando de forma elaborada e profunda um projeto de data warehousing, abrangendo os dados brutos (extraídos manualmente de arquivos pdf e armazenados no SGBD) e um processo de transformação e carga para um segundo ambiente analítico, a partir do qual foram realizadas as tarefas de mineração.

O desafio com as bases de dados desse estudo não era a cardinalidade dos atributos da base, nem a quantidade de registros (que representam cada parlamentar e seu voto na proposição) mas sim a quantidade enorme de bases de dados distintas a serem agregadas, pois cada proposição legislativa analisada era uma base de dados distinta (em seu próprio arquivo pdf), sem uma padronização clara, adicionando um enorme desafio para a extração e consolidação da base de dados.

Para o objetivo último do estudo, que era a descoberta de padrões de semelhança entre parlamentares (e partidos) baseados nas suas votações, assim como a identificação de anomalias dentro de um partido qualquer, foi empregada a tarefa de mineração de agrupamento (*clusterização*), em vez de classificação ou regressão, como é o caso deste projeto. No entanto, como análise acessória ao estudo, foi realizada uma classificação usando árvores de decisão na ferramenta *Weka*, demonstrando a eficácia dessa categoria de algoritmo, o que influenciou na escolha das árvores de decisão também neste trabalho.

3.2 Alves (2021)

O trabalho de Alves [22] foi motivado pelo volume abundante de dados de transações e instituições financeiras trafegados pelos sistemas do Banco Central do Brasil continuamente. Em particular, o estudo se propôs a aplicar mineração de dados para a predição dos números de inadimplência total das instituições financeiras - valores numéricos, constituindo assim um problema de regressão.

As ferramentas usadas envolveram *Weka*, *Orange* e um SGDB (dessa vez o *PostgreSQL*), assim como Hirano e Beserra [21]. A novidade neste estudo é que a linguagem *Python* e a ferramenta *Pandas* foram utilizadas na etapa de pré-processamento.

As bases de dados de indicadores de instituições financeiras estavam fragmentadas em diversas fontes. No entanto, como os registros sempre diziam respeito a uma instituição específica, o *Python* foi usado para agregar todos esses dados distintos realizando um processo análogo ao `JOIN` dos bancos de dados, assim como tratamentos nos tipos de dados e tratamento de valores ausentes. Em seguida, a base consolidada foi carregada no SGBD, pronta para ser consumida pelo *Weka* e *Orange*.

O trabalho de Alves [22] foi de grande utilidade neste trabalho ao comprovar a utilidade da mineração de dados para problemas de regressão prevendo valores de um indicador específico dentre uma série de indicadores, com a diferença que as entidades do mundo real observadas eram instituições financeiras brasileiras, ao invés de países e regiões geo-econômicas do mundo. Nota-se também a inspiração do uso de *Python* para o pré-processamento de dados.

3.3 Santana e Carvalho (2017)

De todos os trabalhos aqui elencados, o de Santana e Carvalho [11] é o único que se debruça sobre os dados do Banco Mundial e da base de dados Indicadores de Desenvolvimento Mundial (WDI), servindo de inspiração para esse projeto. Apesar disso, os objetivos da mineração de dados no nosso estudo são diferentes dos propostos por eles.

O resultado observado no processo de mineração, ou seja, o conhecimento obtido, envolveu o agrupamento de países em *clusters* de acordo com a similaridade dos seus indicadores, que se encaixam em tarefas de aprendizado não-supervisionado, assim como tarefas de regressão usando árvores de decisão para predizer diversos indicadores, como o crescimento do Produto Interno Bruto (PIB), inflação, PIB *per capita*, entre outros.

Um ponto notável nesse estudo é a estratégia "manual"de filtragem dos melhores indicadores. Houve, após uma análise manual, uma simples filtragem "automática"de indicadores com base na quantidade de valores ausentes, reduzindo-os a 125, assim como também ocorreu uma filtragem significativa no número de países, reduzindo bastante a quantidade de registros na base - filtro esse implementado de forma manual e subjetiva, mantendo os de mais "importância"continental e que possuíam mais valores não-nulos.

Essa redução de dados rígida tanto de países quanto de indicadores, somada ao pequeno intervalo temporal abrangido (os 10 anos mais recentes à época), reduziram a riqueza da base de dados.

Em contraste, nosso projeto se propõe a melhorar esse processo aplicando uma filtragem algorítmica de indicadores (parte importante do objetivo do estudo) e removendo apenas os países com mais dados ausentes. Com essa estratégia, a quantidade superlativa de indicadores não é um impeditivo à eficiência da modelagem, nem à análise dos programadores. Também propomos um passo adicional de inferência dos valores ausentes que restarem, também usando modelos preditivos - detalhado na seção 4.4.2.

Assim como os outros trabalhos mencionados acima, Santana e Carvalho [11] também utilizaram ferramentas como *Weka*, *Pentaho* e *Orange*, assim como o *MySQL* para a persistência de dados, sem fazer o uso da linguagem *Python*.

3.4 Porto (2019)

Porto [23] propõe em seu estudo empregar a mineração de dados para descobrir conhecimento, padrões, e fatores que explicariam, ou influenciariam, seu desempenho no Exame Nacional do Ensino Médio (conhecido pelo acrônimo Enem), a partir de dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), correspondente particularmente à aplicação do exame em todo o território brasileiro em 2017.

A grande diferença do trabalho de Porto é utilizar exclusivamente o *Python* e recursos correlatos (*Pandas*, *Jupyter* e *Matplotlib*) para a realização da modelagem, importando a base de dados diretamente em formato CSV sem um ambiente formal de persistência de dados, sendo a maior inspiração para nosso estudo nesse sentido.

O estudo dedica atenção especial à análise exploratória dos dados, que não envolve diretamente a aplicação de modelos de Aprendizado de Máquina (AM), mas sim uma análise estatística apoiada pela manipulação dos *dataframes* e pela geração de tabelas e gráficos para visualização dos dados. Tal abordagem demonstrou a eficácia do *Pandas* e da biblioteca *Matplotlib* para a geração de gráficos úteis (melhorando a visualização de dados por parte dos usuários do modelo) e de descoberta de padrões relevantes.

Porto [23], dessa forma, aplicou a mineração de dados com resultados notáveis na detecção de padrões e análises agrupadas por diversos tipos de indicadores, como localização geográfica, gênero, idade dos estudantes, entre outros, sem utilizar diretamente um algoritmo de AM para a predição de valores em nenhum momento.

Capítulo 4

Estudo de caso: mineração de dados de Indicadores de Desenvolvimento Mundial

O estudo de caso envolve, inicialmente, a coleta dos dados contendo países e seus indicadores a partir da aplicação web do Banco Mundial, de forma manual.

Os outros artefatos a serem construídos são *scripts* na linguagem de programação **Python**. Um deles é responsável por processar o arquivo CSV, transformar os dados para um formato mais apropriado à modelagem e exportar esses dados resultantes para um novo arquivo CSV. Outro, o mais extenso deles, é responsável por importar esse arquivo resultante para um dataframe Pandas e, a partir dele, realizar a construção de um modelo de Aprendizado de Máquina (AM) capaz de minerar dados e obter conhecimento sobre essa base.

O segundo *script* também é responsável por renderizar gráficos para uma visualização analítica descrevendo características a respeito da base de dados, assim como a visualização do resultado do modelo - dessa forma, mesmo não gerando um artefato físico, o algoritmo constrói um modelo que é reproduzível a partir dos *scripts* e exibido a partir dos gráficos gerados.

Dentro dos *scripts*, o modelo será treinado a partir de um conjunto de treinamento e avaliado a partir de um conjunto de teste. Essa modelagem depende diretamente da etapa de pré-processamento dos dados, que dispõe de alguns parâmetros. Ao fim da modelagem, será possível determinar sua qualidade de predição com base nos resultados da sua aplicação sobre o conjunto de teste. Será possível, também, visualizar quais atributos são considerados os mais determinantes pelo modelo para realizar as predições.

Ao longo das seções, são mostrados alguns trechos de códigos do script da modelagem, mostrados nas listagens 4.1, 4.2, 4.3, 4.4. Todos eles fazem parte do código exibido

integralmente no Apêndice B.

4.1 Ferramentas

Python

Python é uma linguagem de programação amplamente utilizada em ciência de dados e mineração de dados. Sua sintaxe simples e legibilidade tornam-na uma escolha popular para análise e modelagem de dados. Além disso, a vasta comunidade de desenvolvedores e a disponibilidade de bibliotecas especializadas facilitam a implementação de algoritmos de aprendizado de máquina e a manipulação de dados [24]. A versão usada é a **3.11.5**.

Pandas

O Pandas é uma biblioteca Python para análise de dados que oferece estruturas de dados flexíveis, como DataFrames e Series. Ele permite a leitura, manipulação e limpeza eficiente de dados tabulares. Com o Pandas, é possível realizar operações como filtragem, agregação, transformação e visualização de dados de maneira eficaz [25].

Scikit-Learn

O Scikit-Learn (ou sklearn) é uma biblioteca de aprendizado de máquina em Python. Ele fornece uma ampla variedade de algoritmos de classificação, regressão, agrupamento e pré-processamento de dados. O Scikit-Learn é amplamente utilizado para construir, treinar e avaliar modelos de aprendizado de máquina, tornando-o essencial para projetos de mineração de dados [25].

Spyder

O Spyder é um ambiente de desenvolvimento integrado (IDE) específico para ciência de dados e análise numérica em Python. Ele oferece recursos como edição de código, depuração, visualização de variáveis e integração com bibliotecas populares, incluindo Pandas e Scikit-Learn. O Spyder é uma escolha conveniente para desenvolvedores que desejam trabalhar com eficiência em projetos de mineração de dados [26]. A versão usada é a **5.4.3**. Todas as imagens de visualização de Dataframes foram obtidas através desta ferramenta.

4.2 Base de dados do Banco Mundial

Na base de dados dos Indicadores de Desenvolvimento Mundial (WDI) estão contidos os dados de indicadores de todos os países cadastrados no mundo, assim como dados agregados por regiões geoeconômicas.

Para cada país ou região, dispostos como linhas, constam todos os indicadores socioeconômicos, também dispostos em linhas. Já as colunas representam os anos de aferição de dados, ou seja, a progressão anual de cada um desses indicadores, desde 1960 até 2023 (período de 64 anos). Em outras palavras, cada objeto nessa base de dados é composto por: país/região e sua abreviação; nome e código do indicador; valor do indicador para esse país/região em 1960; valor do indicador para esse país/região em 1961; e assim sucessivamente.

A base de dados WDI, portanto, é composta de **396.872 linhas** (produto dos 266 países/regiões pelos 1492 indicadores) e **68 colunas** (sendo 4 para a descrição dos países/regiões e indicadores; e mais 64 correspondentes aos valores anuais).

Todos os nomes de países, regiões e indicadores, assim como todos os metadados, se encontram em língua inglesa.

4.2.1 Coleta dos dados

A base de dados dos Indicadores de Desenvolvimento Mundial (WDI) está distribuída em arquivos *Comma-separated Values* (CSV), que foram extraídos de um arquivo zipado, obtido manualmente a partir da página do DataCatalog (Catálogo de Dados) do endereço web do Banco Mundial [27].

A base em si é apenas um dos 6 arquivos CSV presentes no arquivo zipado - os outros fornecem metadados diversos, incluindo sobre os países/regiões e sobre os indicadores, que serão úteis em momento oportuno. A base possui tem tamanho aproximado de 190 *megabytes*; já a totalidade dos arquivos têm tamanho aproximado de 265 *megabytes*.

Alternativamente, o Banco Mundial oferece mais possibilidades de obtenção das bases: uma é de forma computadorizada, por uma *Application Programming Interface* (API). Tal opção não foi explorada neste trabalho devido à dificuldade para aprender o funcionamento e contrato da API para obter os dados nos parâmetros desejados - exigindo um esforço significativamente maior que o *download* da base diretamente do DataCatalog, feito com um clique.

Outra opção, que inicialmente foi utilizada antes do DataCatalog, é o DataBank, ferramenta de visualização que contém coleções de dados diversas além da WDI. O DataBank oferece uma flexibilidade muito maior para, por ex., filtrar indicadores, países e períodos temporais e decidir o formato mais adequado para exportar. No entanto, uma limitação

que impossibilitou seu uso foi o fato de não ser possível baixar a base WDI inteira sem que a aplicação "quebrasse": teria que ser feito o *download* manual uma vez para cada ano disponível. Inevitavelmente, obter a base pelo DataBank resultaria num arquivo menos populada e rica.

4.2.2 Indicadores socioeconômicos

Os registros da base de dados WDI são caracterizados por 1492 indicadores socioeconômicos.

Entre os indicadores mais importantes, a base inclui dados sobre indicadores como Produto Interno Bruto (PIB), expectativa de vida, taxa de alfabetização, desigualdade, pobreza, entre outros. Esses dados são coletados de fontes oficiais e permitem análises comparativas e estudos de tendências globais.

Todos os 1492 dentro das bases de dados, são referidos por um código - uma abreviatura de até 25 caracteres, assim como por seu nome por extenso, todos em inglês.

Os indicadores estão divididos em alguns grupos temáticos [2], descritos a seguir. Dentre cada categoria, eles podem aparecer replicados com diferentes clivagens, mantendo relativa semelhança ou dependência entre si - por ex., um indicador para cada gênero; valor absoluto e em porcentagem anual; medições em diferentes unidades; entre outros.

Pobreza e desigualdade: Pobreza, prosperidade, consumo, distribuição de renda, entre outros;

População: Dinâmicas de população, educação, trabalho, saúde e gênero, entre outros;

Ambiente: Agricultura, mudanças climáticas, energia, biodiversidade, saneamento, entre outros;

Economia: Crescimento econômico, estrutura econômica, renda, poupanças, comércio, entre outros;

Mercados: Negócios, mercados de ações, comunicações, transporte, tecnologia, indústria militar, entre outros;

Relações internacionais: Dívida externa, comércio externo, ajuda humanitária, turismo, migração, entre outros;

A tarefa da modelagem desse projeto é predizer os valores para um indicador específico dentro do tema da economia - o indicador **Crescimento Anual do PIB** (*GDP growth (annual %)*), de código `NY.GDP.MKTP.KD.ZG`. Essa predição será feita usando todos

indicadores de todos os grupos temáticos para o treinamento do modelo, excluindo-se do treinamento a própria variável-alvo.

No Apêndice C, há uma lista exaustiva de todos os indicadores da base WDI, com seu código e seu nome em inglês[1].

4.2.3 Metadados

Junto à base dos Indicadores de Desenvolvimento Mundial (WDI), foram extraídos vários arquivos de metadados, ou seja, bases de dados que fornecem contexto em relação à base principal:

WDICountry: Lista os países, regiões e informações contextuais intrínsecas aos mesmos como, por exemplo, a região a que pertence, nome da moeda corrente, ano-base para o cálculo de indicadores e o último ano de realização de censos demográficos;

WDISeries: Lista todos os indicadores socioeconômicos com abreviaturas, nomes completos, textos de descrição, unidades com os quais são medidos, metodologias para a aferição, entre vários outros;

WDIcountry-series: Mapeia relacionamentos entre países e indicadores, denotando as fontes dos valores de cada indicador para cada país;

WDIseries-time: Contém relacionamentos entre alguns países e alguns anos, adicionando informação contextual relevante;

WDIfootnote: Notas de rodapé com observações relevantes para algumas aferições de alguns países, como, por exemplo, o grau de incerteza.

4.3 Transformação e carga dos dados

A base de dados WDI dispõe os anos como colunas, estruturada numa espécie de progressão temporal para cada indicador, que são dispostos como linhas, para cada país ou região. Como tarefas de mineração geralmente procuram descobrir padrões e conhecimento sobre as características (colunas) de uma base, este formato somente seria útil se a tarefa fosse prever a progressão dos indicadores para um ano específico (por exemplo, no futuro). Para que tal tarefa resultasse na geração de conhecimento útil, seria necessário restringir bastante o conjunto de indicadores e/ou países e, então, executar predições temporais.

[1] alguns dos nomes dos indicadores foram abreviados para fins de adequação ao tamanho da tela e para prevenir o excesso de páginas.

Ao invés disso, um dos objetivos centrais do estudo é determinar quais são os indicadores que mais impactam a modelagem na predição de uma característica específica. Sendo assim, foi necessário transformar a base de dados de forma que os indicadores fossem as colunas, enquanto os anos de aferição passariam a ser linhas.

Tal transformação pode ser feita com o Pandas submetendo o *dataframe* original (mostrado na Figura 4.1), aos métodos `melt` e `pivot`, sucessivamente (vide Apêndice A):

- O `melt` tem a função de "rebaixar"as colunas dos anos para linhas. É criado um novo atributo que recebe o valor observado em cada ano, sendo 64 linhas criadas para cada registro original. O restante dos atributos, como "Nome do País"e "Nome do Indicador", continuam inalterados.
- O `pivot` faz o inverso: escolhe um ou mais atributos e mapeia todos os seus valores observados em novas colunas - nesse caso, escolhemos o atributo "Código do Indicador". Uma observação é que nesse passo é perdido o atributo "Nome do indicador", mas tal perda não impacta o estudo pois essas informações podem ser recuperadas novamente, a partir do código, pela base de metadados.

Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963
Albania	ALB	Population ages 65-69, male (% of male population)	SP.POP.6569.MA.5Y	1.45798	1.4766	1.52173	1.57066
Albania	ALB	Population ages 70-74, female (% of female population)	SP.POP.7074.FE.5Y	1.78195	1.7001	1.61665	1.54833
Albania	ALB	Population ages 70-74, male (% of male population)	SP.POP.7074.MA.5Y	1.25739	1.1799	1.10039	1.03302
Albania	ALB	Population ages 75-79, female (% of female population)	SP.POP.7579.FE.5Y	1.2925	1.29623	1.29487	1.28154
Albania	ALB	Population ages 75-79, male (% of male population)	SP.POP.7579.MA.5Y	0.846119	0.84596	0.843368	0.833143
Albania	ALB	Population ages 80 and above, female (% of female population)	SP.POP.80UP.FE.5Y	1.48591	1.47107	1.45318	1.4331
Albania	ALB	Population ages 80 and above, male (% of male population)	SP.POP.80UP.MA.5Y	0.868032	0.854566	0.840647	0.827879
Albania	ALB	Population density (people per sq. km of land area)	EN.POP.DNST	nan	60.5766	62.4569	64.3292
Albania	ALB	Population growth (annual %)	SP.POP.GROW	nan	3.12086	3.05673	2.95375
Albania	ALB	Population in largest city	EN.URB.LCTY	134761	137714	139561	141434
Albania	ALB	Population in the largest city (% of urban population)	EN.URB.LCTY.UR.ZS	27.2805	26.8139	26.2943	25.8125
Albania	ALB	Population in urban agglomerations of more than 1 million	EN.URB.MCTY	nan	nan	nan	nan
Albania	ALB	Population in urban agglomerations of more than 1 million (% …	EN.URB.MCTY.TL.ZS	nan	nan	nan	nan
Albania	ALB	Population living in areas where elevation is below 5 meters …	EN.POP.EL5M.ZS	nan	nan	nan	nan
Albania	ALB	Population living in slums (% of urban population)	EN.POP.SLUM.UR.ZS	nan	nan	nan	nan

Figura 4.1: Trecho do *Dataframe* original, obtido a partir da base WDI original (visualizado pela interface do Spyder).

A nova base de dados transformada é então carregada (armazenada) no mesmo formato CSV em um arquivo com nome `WDItratado`[2]. Tal arquivo será o objeto dos algoritmos de modelagem em todos os próximos passos.

Essa nova base de dados é composta de **17.024 linhas** (produto dos 266 países/regiões por 64 anos) e **1495 colunas** (sendo 3 para a descrição dos países/regiões e indicadores; e mais 1492 indicadores).

[2]A partir desse ponto, quando nos referirmos à base de dados Indicadores de Desenvolvimento Mundial (WDI), estamos nos referindo na verdade à base WDItratado.

Country Name	Country Code	Year	AG.AGR.TRAC.NO	AG.CON.FERT.PT.ZS	AG.CON.FERT.ZS	AG.LND.AGRI.K2	AG.LND.AGRI.ZS	AG.LND.ARBL.HA
Arab World	ARB	2018	nan	14.584	55.9942	5.23625e+06	39.9697	nan
Arab World	ARB	2019	nan	14.0102	55.5936	5.22804e+06	39.907	nan
Arab World	ARB	2020	nan	13.0246	54.6098	5.23672e+06	39.9733	nan
Arab World	ARB	2021	nan	13.0369	55.0926	5.23639e+06	39.9707	nan
Arab World	ARB	2022	nan	nan	nan	nan	nan	nan
Arab World	ARB	2023	nan	nan	nan	nan	nan	nan
Argentina	ARG	1960	nan	nan	nan	nan	nan	nan
Argentina	ARG	1961	120000	523.903	0.873313	1.37829e+06	50.3634	1.8597e+07
Argentina	ARG	1962	130000	211.85	0.662722	1.36434e+06	49.8537	1.918e+07
Argentina	ARG	1963	140000	421.6	1.07551	1.34875e+06	49.284	1.96e+07
Argentina	ARG	1964	150000	590	1.475	1.33297e+06	48.7074	2e+07
Argentina	ARG	1965	155000	810	1.97938	1.3178e+06	48.1531	2.0461e+07
Argentina	ARG	1966	160000	1413.88	2.4328	1.30948e+06	47.849	2.1608e+07
Argentina	ARG	1967	165000	604.183	3.06432	1.29703e+06	47.3941	2.2341e+07

Figura 4.2: Trecho do *Dataframe* tratado com a aplicação das funções melt e pivot (visualizado pela interface do Spyder).

4.4 Pré-processamento de dados

A base de dados WDI possui uma quantidade enorme de valores vazios ao longo dos anos e dos indicadores. Ao todo, são 16.572.594 aferições de indicadores vazias, resultando numa proporção de aproximadamente 66% da quantidade total (vide Figura 4.3). Isso torna consideravelmente difícil construir qualquer modelo preditivo com uma boa acurácia.

Sendo assim, é indispensável, na etapa de pré-processamento dos dados, buscar reduzir esse "buraco"na base. Isso será feito com dois métodos diferentes, que são explicados nas subseções seguintes.

Ao longo dessa seção, a base de dados será ilustrada através de figuras com análise da base de dados a partir do Spyder e do Matplotlib, representando a base antes e depois do pré-processamento.

4.4.1 Remoção de valores vazios

A redução de dados será feita com diferentes estratégias, visando reduzir a quantidade de valores vazios. Ao mesmo tempo, é preciso lembrar que essa "poda"muitas vezes tem como efeito colateral a perda de dados reais, úteis, não-nulos. Sendo assim, é necessário encontrar um certo equilíbrio de forma a manter a qualidade da base de dados.

A primeira etapa de redução de dados é remover todos os registros em que o valor da variável-alvo **Crescimento Anual do PIB** é vazio. Como se trata de um algoritmo de aprendizado supervisionado, todos os dados de treinamento precisam de rótulos informa-

```
In [3]: total_nan = wdi.isna().sum().sum()

In [4]: total_nan
Out[4]: 16752594

In [5]: total_values = total_countries * total_indicators * total_years

In [6]: total_values
Out[6]: 25304320

In [7]: total_nan / total_values
Out[7]: 0.6620448208053012
```

Figura 4.3: Cálculo da quantidade de valores nulos no *Dataframe* original, antes do processo de redução de dados (visualizado pelo console do Spyder).

dos (no caso numéricos). Logo, não faz sentido manter na base as tuplas onde o rótulo desejado é desconhecido.

Esse é o caso de 3.173 registros (18,6% do total); após esse tratamento, a base de dados é reduzida de 17.024 para 13.851 registros.

Em seguida, serão aplicadas três etapas de redução de dados: redução de anos, redução de países e redução de indicadores, todas baseadas em eliminar aqueles com mais valores vazios.

Depois, iremos retirar da base todos os registros pertencentes a anos com menos dados válidos. É possível visualizar a quantidade de valores nulos agrupados por cada ano na Figura 4.4, na qual nota-se que os dados mais antigos são os mais propensos a estarem ausentes, assim como o ano mais recente (2023). A quantidade de anos a remover da base é determinada pelo parâmetro `YEARS TO DROP`.

Em seguida, iremos remover do *Dataframe* todos os registros dos países mais "vazios", isto é, com a maior proporção de valores ausentes. A lista desses países é vista na Figura 4.5, onde é possível perceber a prevalência de pequenas ilhas não-soberanas e países de menor território. A quantidade de países a remover é determinada pelo parâmetro `COUNTRIES TO DROP`.

Por último, outro jeito de reduzir os dados será "podar"os indicadores com dados mais vazios, efetivamente reduzindo a quantidade de atributos a serem considerados no treinamento - diferentes das reduções anteriores em que a base perdia registros ao invés de atributos.

Isso pode ser feito de duas formas: remover uma quantidade parametrizada de indicadores com mais dados ausentes; ou estabelecer um limiar no qual todos os indicadores que tiverem uma quantidade dados válidos menor que esse valor são eliminados da base. Neste projeto foi adotado o segundo caminho. O limiar de valores não-nulos a ser respei-

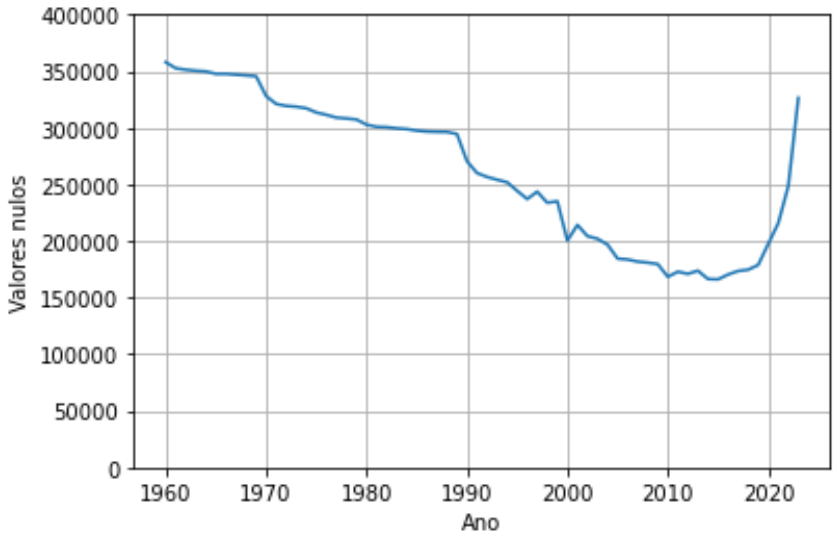

Figura 4.4: Quantidade de valores nulos observados em cada ano dentro da base original, considerando todos os países/regiões e indicadores.

tado pelos indicadores para se manterem na base filtrada é determinada pelo parâmetro `INDICATORS_NOT_NAN_THRESHOLD`.

Nas figuras Figura 4.6 são mostrados os indicadores mais vazios e, na Figura 4.7, um histograma mostrando a frequência de valores vazios, onde percebemos que a grande maioria dos indicadores ultrapassa 8500 valores nulos, ou seja, mais da metade do total de 17.024 registros, aproximadamente.

Ao longo de todas as etapas de redução de dados, os três parâmetros acima são de suma importância e podem alterar drasticamente a base de dados resultante, em qualidade e quantidade de dados (existem ainda outros parâmetros para a modelagem, a serem usados em outras partes do estudo, como nas seções 4.5 e 4.6. Executamos o pré-processamento nesse estudo com os seguintes valores para os parâmetros mostrado no código 4.1:

```
1 YEARS_TO_DROP = 16                    # 1/4 do total de anos
2 COUNTRIES_TO_DROP = 28                # aprox. 10% dos 266 países e regiõ
    es
3 INDICATORS_NOT_NAN_THRESHOLD = 0.6    # exclui todo indicador com + de 40%
     valores nulos
```

Listing 4.1: Declaração das variáveis de parâmetros no início do script de modelagem.

Ao fim da redução de dados com os parâmetros do código 4.1, a base agora contém 10.653 registros e 551 colunas (portanto 549 indicadores). A proporção de valores nulos na base de dados diminuiu de aproximadamente **66%** para aproximadamente **20%** (vide a Figura 4.8) - uma melhora significativa.

Country Code	Country Name	NaN values
INX	Not classified	95488
MAF	St. Martin (French part)	90013
SXM	Sint Maarten (Dutch part)	87659
IMN	Isle of Man	87572
MNP	Northern Mariana Islands	87519
CHI	Channel Islands	87251
CUW	Curacao	85609
GIB	Gibraltar	85536
ASM	American Samoa	85355
MCO	Monaco	85070
VGB	British Virgin Islands	84634
LIE	Liechtenstein	84264
TCA	Turks and Caicos Islands	84198
VIR	Virgin Islands (U.S.)	83824
GUM	Guam	83667
FRO	Faroe Islands	83512
CYM	Cayman Islands	83076
GRL	Greenland	82438
SMR	San Marino	81964
AND	Andorra	81523

Figura 4.5: Lista de países e territórios com maior ausência de valores na base original (visualizado pela interface do Spyder).

4.4.2 Inferência de valores vazios

A inferência de valores vazios tem como objetivo eliminar a ocorrência de valores nulos em toda a base, preenchendo esses valores seguindo alguma estratégia. Existem diversos métodos para a inferência de vazios, explicados na seção 2.3.4.

Além das estratégias triviais, como atribuir o valor 0 ou uma constante qualquer a todos os valores nulos, rejeitamos também usar medidas estatísticas como a média, moda e mediana dos atributos, já que não podemos generalizar os valores observados de um indicador em todos os países do mundo para inferir o que ocorre em um país específico.

Faz mais sentido, aplicando um pouco de senso comum, que um valor ausente de um certo indicador em um certo país, região ou território siga a tendência dos valores observados mais "próximos", isto é: valores válidos de anos próximos do mesmo indicador e mesmo país, ou valores de indicadores correlacionados, ou valores de países com valores semelhantes de indicadores. Tendo isso em mente, adotamos a estratégia preditiva de

Name	NaN values	Percentage
Net official flows from UN agencies, UNIDIR (current US$)	17024	100
Net official flows from UN agencies, UNCTAD (current US$)	17004	99.88
Disaster risk reduction progress score (1-5 scale; 5=best)	16941	99.51
Female genital mutilation prevalence (%)	16933	99.47
Present value of external debt (% of exports of goods, services and income)	16931	99.45
Present value of external debt (% of GNI)	16929	99.44
Children in employment, self-employed, female (% of female children in employment, a…	16926	99.42
Children in employment, self-employed, male (% of male children in employment, ages …	16926	99.42
Children in employment, self-employed (% of children in employment, ages 7-14)	16926	99.42
Proportion of women subjected to physical and/or sexual violence in the last 12 mont…	16918	99.38
Public private partnerships investment in ICT (current US$)	16914	99.35
Multidimensional poverty headcount ratio (UNDP) (% of population)	16914	99.35
Adequacy of unemployment benefits and ALMP (% of total welfare of beneficiary househ…	16911	99.34
Benefit incidence of unemployment benefits and ALMP to poorest quintile (% of total …	16911	99.34
Present value of external debt (current US$)	16910	99.33
Women making their own informed decisions regarding sexual relations, contraceptive …	16910	99.33
Net official flows from UN agencies, UNCDF (current US$)	16904	99.3
Net official flows from UN agencies, UNWTO (current US$)	16901	99.28
Average working hours of children, working only, female, ages 7-14 (hours per week)	16900	99.27
Annualized average growth rate in per capita real survey mean consumption or income,…	16899	99.27
Annualized average growth rate in per capita real survey mean consumption or income,…	16899	99.27
Net official flows from UN agencies, UNEP (current US$)	16898	99.26
Average working hours of children, working only, male, ages 7-14 (hours per week)	16898	99.26
Average working hours of children, working only, ages 7-14 (hours per week)	16898	99.26

Figura 4.6: Lista de indicadores com mais valores nulos na base original (visualizado pela interface do Spyder).

utilizar o algoritmo de Aprendizado de Máquina dos **k vizinhos mais próximos** ou *K-Nearest Neighbors* (KNN) para inferir os valores ausentes, cuja implementação no Scikit-Learn é chamada `KNNImputer` [28].

Dessa forma, estamos aplicando um modelo de Aprendizado de Máquina de forma a auxiliar um outro modelo (esse último atingindo objetivo final da mineração) a ser mais eficiente. Nesta aplicação de regressão com KNN, os dados de treinamento são todos os valores presentes da base de dados. Assim, para cada valor de indicador a ser inferido, tal indicador será o rótulo desejado e os valores ausentes serão aproximados de acordo com os registros mais "próximos"(semelhantes).

Ao final do treinamento e estimação do KNN, com o parâmetro listado no código 4.2, obtém-se uma base de dados sem nenhum valor ausente restante. Esse parâmetro é capaz de influenciar as aproximações obtidas e, portanto, influenciar as decisões do modelo final

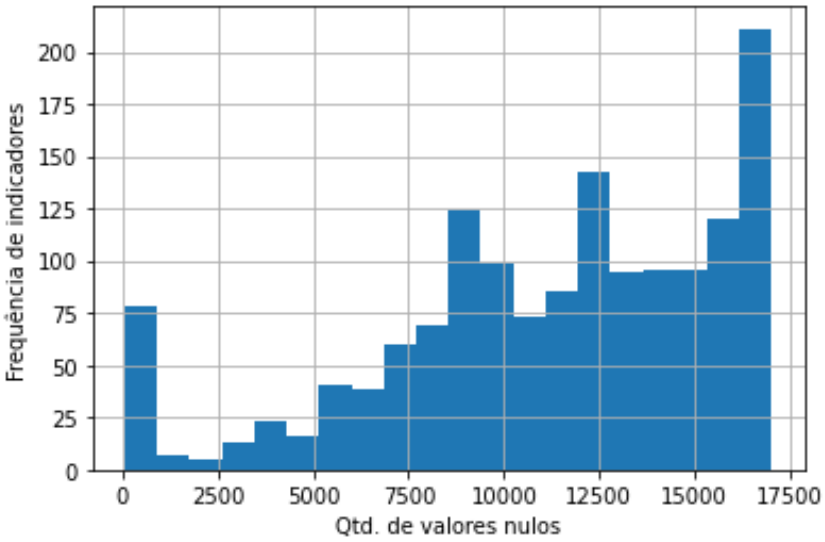

Figura 4.7: Histograma mostrando a distribuição de valores nulos aferidos por todos os indicadores.

na seção 4.7.

```
KNN_IMPUTER_NEIGHBOURS = 5
```

Listing 4.2: Parâmetro da quantidade de vizinhos que o KNN considera.

4.5 Conjuntos de testes e treinamento

Após o processo de limpeza dos dados, é feita a separação entre conjunto de teste e conjunto de treinamento.

Diversas estratégias foram cogitadas, como, por exemplo, realizar uma separação temporal - anos anteriores a um ano x seriam parte do conjunto de treinamento e os anos restantes, mais recentes, seriam o conjunto de treinamento. Foi analisado também utilizar uma separação heurística por países - países específicos em cada continente seriam o conjunto de teste.

No entanto, tal separação geográfica exigiria uma aplicação de senso comum, sujeito aos vieses dos autores, que poderia influenciar o resultado final sem uma análise mais aprofundada e heurística de seu impacto. Sendo assim, foi implementada uma separação entre teste e treinamento totalmente aleatória, com proporção correspondente ao parâmetro declarado no código 4.3 - onde o valor declarado é a proporção do tamanho do conjunto de teste em relação ao todo. A separação aleatória é implementada com o método `train_test_split`[29] do Scikit-Learn.

```
In [5]: total_nan = wdi.isna().sum().sum()

In [6]: total_nan
Out[6]: 1242059

In [7]: total_values = total_countries * total_indicators * total_years

In [8]: total_values
Out[8]: 6234048

In [9]: total_nan / total_values
Out[9]: 0.1992379590275853
```

Figura 4.8: Cálculo da quantidade de valores nulos no *Dataframe* após as três etapas de redução de dados (visualizado pelo console do Spyder).

```
1 TEST_SET_RATIO = 0.25
```

Listing 4.3: Parâmetro do tamanho do conjunto de treinamento.

Como explicado na seção 4.2.2, é realizada nessa etapa a separação entre variável-alvo, chamada `NY.GDP.MKTP.KD.ZG` e denotada como `y` no script, e o conjunto restante dos atributos, denotados como `X` - na verdade, essa fragmentação é realizada antes da separação entre conjuntos de teste e treinamento, pois o `train_test_split` requer ambos `X` e `y` como parâmetros. Logo, existem agora 4 subconjuntos distintos do dataframe WDI: `X_train`, `X_test`, `Y_train` e `Y_test`.

4.6 Seleção de Atributos

Em seguida, será feita a seleção de atributos, realizada de forma algorítmica pelo Scikit-Learn. Dentre as muitas metodologias possíveis , como as de limiares de variância ou a eliminação recursiva de atributos [30], foi escolhido o algoritmo `SelectKBest`, considerado um método *univariável*. Ele consiste em manter apenas um número pré-determinado (pelo usuário) dos melhores atributos no dataframe.

O critério para saber o que é "melhor"consiste em cálculos estatísticos sobre as variáveis de interesse e a variável-alvo, afim de medir a correlação entre elas. O cálculo específico a ser usado também é determinado pelo usuário; neste caso, foi escolhido o `r_regression`, que calcula o coeficiente de determinação (R^2) entre as variáveis para problemas de regressão, que lidam com o domínio dos números reais.

Escolhido o critério de filtragem dos atributos, o algoritmo será executado usando como argumentos os conjuntos `X_train` (características) e `Y_train` (rótulo desejado). A quantidade de atributos a serem preservados são determinados pelo parâmetro declarado

no código 4.4. A partir do resultado da seleção, esse filtro precisa ser aplicado não apenas em `X_train` mas também no `Y_train`

```
FEATURES_TO_SELECT = 32
```

Listing 4.4: Parâmetro da quantidade de melhores atributos a serem selecionados.

Saber quais os atributos foram escolhidos pelo modelo `SelectKBest` é um dos objetivos centrais do estudo, por isso essa informação será detalhada na seção 5.1.

4.7 Modelo de *Random Forest*

Com o término da seleção de atributos pelo `SelectKBest`, é dado o início a aplicação do modelo de *Random Forest*. Como explicado na seção 2.5.3, o *Random Forest* é um algoritmo de Aprendizado de Máquina que consiste em utilizar vários subconjuntos de dados de treinamento para construir uma série de árvores de decisão proporcionado pela biblioteca Scikit-Learn como `RandomForestRegressor` [31].

Uma característica relevante das árvores de decisão e, por extensão, das florestas aleatórias, é que os dados não precisam passar por normalização, pela sua característica dos nós de decisão, que são as etapas por onde a predição é realizada. A normalização consiste em aplicar uma escala de grandeza comum para todos os valores analisados, de forma proporcional, para atenuar a possível influência dos dados que tem uma grandeza e amplitude maior sobre o modelo como um todo. Caso fosse necessário, ao usar algum outro algoritmo de regressão, a normalização seria realizada junto com as outras etapas de pré-processamento, logo antes da inferência de valores vazios (seção 4.4.2).

O modelo é treinado utilizando o subconjunto da seleção aplicada aos atributos de características (`X_train_selected`) e a variável-alvo (`Y_train`) no código 4.5.

```
random_forest = RandomForestRegressor(random_state=0)
random_forest.fit(X_train_selected, y_train)
```

Listing 4.5: Treinamento do modelo de Random Forest

O único parâmetro explicitamente declarado para o algoritmo é o `random_state`, que garante uma execução aleatória inédita cada vez que o *script* for executado. No entanto, o modelo possui diversos parâmetros que, por não estarem explicitados, são definidos de acordo com o padrão do Scikit-Learn. Esses parâmetros são mostrados na Figura 4.9. Entre eles, destacam-se de relevantes para a análise o `n_estimators`, que determina a quantidade de árvores de decisão usadas na floresta e o `max_features`, que determina a quantidade máxima de características usadas em cada nó de decisão.

```
In [22]: random_forest.get_params()
Out[22]:
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': None,
 'max_features': 1.0,
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 0,
 'verbose': 0,
 'warm_start': False}
```

Figura 4.9: Parâmetros do modelo Random Forest.

Em seguida, o desempenho do modelo é avaliado no conjunto de testes pelo método score do `RandomForestRegressor`, mostrado no código 4.6, e que será detalhado na seção 5.2.

```
score = random_forest.score(X_test_selected, y_test)
```

Listing 4.6: Avaliação do desempenho do modelo de *Random Forest*

Capítulo 5

Resultados dos modelos

5.1 Resultado da seleção de atributos

O algoritmo `SelectKBest`, mencionado na seção 4.6 e executado com o parâmetro informado `FEATURES_TO_SELECT`, foi executado sobre o conjunto de características de treinamento (`X_train`) e a variável de interesse respectiva, também de treinamento (`Y_train`). Os indicadores restantes no dataframe `X_train`, após o procedimento, estão listados nas tabelas 5.1 e 5.2. Esse mesmo filtro de indicadores foi então aplicado ao dataframe do conjunto de teste `X_test`.

Com a obtenção dos indicadores mais correlacionados à análise do crescimento do PIB, está provada a última afirmação da **Hipótese 1**, que dispunha sobre a capacidade do modelo de selecionar os atributos conforme sua relevância ou "score"comparado com o indicador desejado.

Os padrões de correlação detectados pelo seletor de atributos revela a influência de alguns indicadores que, à primeira vista, considerando o senso comum, não seriam considerados numa análise manual de indicadores que influenciam o crescimento do PIB, como o `NY.ADJ.DNGY.GN.ZS` (*Poupança ajustada: esgotamento de energia (% da Renda Nacional Bruta)*, corroborando com a **Hipótese 2** da seção 1.5.1.

No entanto, várias das características selecionadas corroboram de certa forma com o senso comum do que se imagina quando se pensa no que mais influencia tal indicador, como dados de população, crescimento populacional, poupança bruta, poupança ajustada e crescimento do consumo das famílias.

Series Code	Indicator Name
NE.CON.GOVT.KD.ZG	General government final consumption expenditure (annual % growth)
NE.CON.PRVT.KD.ZG	Household and NPISHs Final consumption expenditure (annual % growth)
NE.CON.PRVT.PC.KD.ZG	Household final consumption expenditure per capita growth (annual %)
NE.CON.TOTL.KD.ZG	Final consumption expenditure (annual % growth)
NE.EXP.GNFS.KD.ZG	Exports of goods and services (annual % growth)
NE.GDI.FTOT.KD.ZG	Gross fixed capital formation (annual % growth)
NE.GDI.FTOT.ZS	Gross fixed capital formation (% of GDP)
NE.GDI.TOTL.KD.ZG	Gross capital formation (annual % growth)
NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
NE.IMP.GNFS.KD.ZG	Imports of goods and services (annual % growth)
NV.AGR.TOTL.KD.ZG	Agriculture, forestry, and fishing, value added (annual % growth)
NV.IND.MANF.KD.ZG	Manufacturing, value added (annual % growth)
NV.IND.TOTL.KD.ZG	Industry (including construction), value added (annual % growth)
NV.IND.TOTL.ZS	Industry (including construction), value added (% of GDP)
NV.SRV.TOTL.KD.ZG	Services, value added (annual % growth)
NY.ADJ.DNGY.GN.ZS	Adjusted savings: energy depletion (% of GNI)

Tabela 5.1: Indicadores escolhidos na seleção de atributos (parte 1)

Series Code	Indicator Name
NY.ADJ.DRES.GN.ZS	Adjusted savings: natural resources depletion (% of GNI)
NY.ADJ.ICTR.GN.ZS	Adjusted savings: gross savings (% of GNI)
NY.ADJ.NNAT.GN.ZS	Adjusted savings: net national savings (% of GNI)
NY.GDP.PCAP.KD.ZG	GDP per capita growth (annual %)
NY.GDP.PETR.RT.ZS	Oil rents (% of GDP)
NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)
NY.GDS.TOTL.ZS	Gross domestic savings (% of GDP)
NY.GNS.ICTR.GN.ZS	Gross savings (% of GNI)
NY.GNS.ICTR.ZS	Gross savings (% of GDP)
SL.FAM.WORK.FE.ZS	Contributing family workers, female (% of female employment) (modeled ILO estimate)
SP.POP.1519.FE.5Y	Population ages 15-19, female (% of female population)
SP.POP.1519.MA.5Y	Population ages 15-19, male (% of male population)
SP.POP.2024.FE.5Y	Population ages 20-24, female (% of female population)
SP.POP.2024.MA.5Y	Population ages 20-24, male (% of male population)
SP.POP.GROW	Population growth (annual %)
SP.URB.GROW	Urban population growth (annual %)

Tabela 5.2: Indicadores escolhidos na seleção de atributos (parte 2)

5.2 Desempenho da predição do indicador de crescimento do PIB

Para avaliar o desempenho do modelo, foi realizado o cálculo do coeficiente de determinação (R^2), que irá indicar a proporção de variância na variável desejado. Um valor próximo de 1 indicada que o modelo consegue explica a maior parte da variabilidade dos dados, enquanto valores próximos de 0 indicam que o modelo não está capturando bem a relação entre os atributos e a variável-alvo.

O *Random Forest* possuem uma grande capacidade de lidar bem com grandes conjuntos de dados e diminiu o risco de overfitting, ou seja, ele garante precisão para novos dados a serem inseridos após o treinamento. Além disso, pelo resultado do coeficiente de determinação é possível atribuir quais indicadores selecionados possuem maior importância na previsão do crescimento do PIB

```
In [23]: score
Out[23]: 0.9672221842107698
```

Figura 5.1: Valor do *score* do modelo executado sobre o conjunto de teste, mostrando acurácia de aproximadamente 97%.

A Figura 5.1 apresenta uma avaliação quantitativa do desempenho do modelo utilizado no estudo, exibindo um *score* de aproximadamente 97% no conjunto de teste.

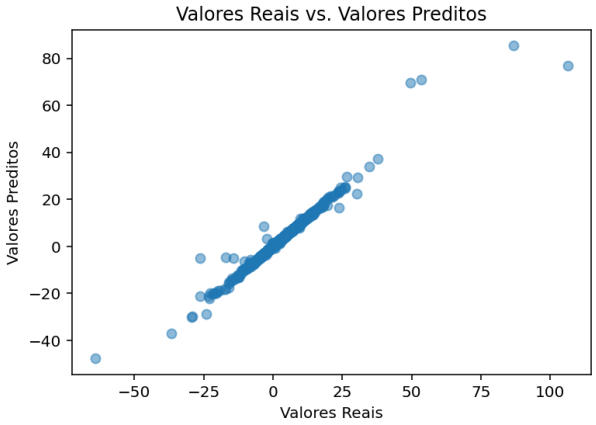

Figura 5.2: Análise de Desempenho do modelo: Valores Reais vs Valores Preditos.

A Figura 5.2 fornece um gráfico de dispersão, com o eixo `X` representando os valores da variável-alvo e o eixo `Y` os valores previstos para o modelo. A distribuição dos pontos

em relação à linha diagonal (45 graus) indicam que as previsões estão bem alinhadas com os valores reais, validando a alta acurácia observada.

Ambas as figuras corroboram com a primeira afirmação da **Hipótese 1** apresentada no Capítulo 1, reforçando a eficácia e a precisão do modelo.

Em seguida, foi analisado os resíduos do modelo. Os resíduos são gerado a partir da diferença entre os valores reais e o valores preditos, este gráfico se diferencia do de dispersão pois ajuda a identificar padrões nos erros que podem indicar problemas no conjunto de variáveis.

Na Figura 5.3 é observado que os resíduos estão distribuídos em sua maioria sobre a linha horizontal de zero. Isso indica mais uma vez, que o modelo não possui padrões de error sistemático e que está capturando bem a relação entre as variáveis de entrada e a variável-alvo.

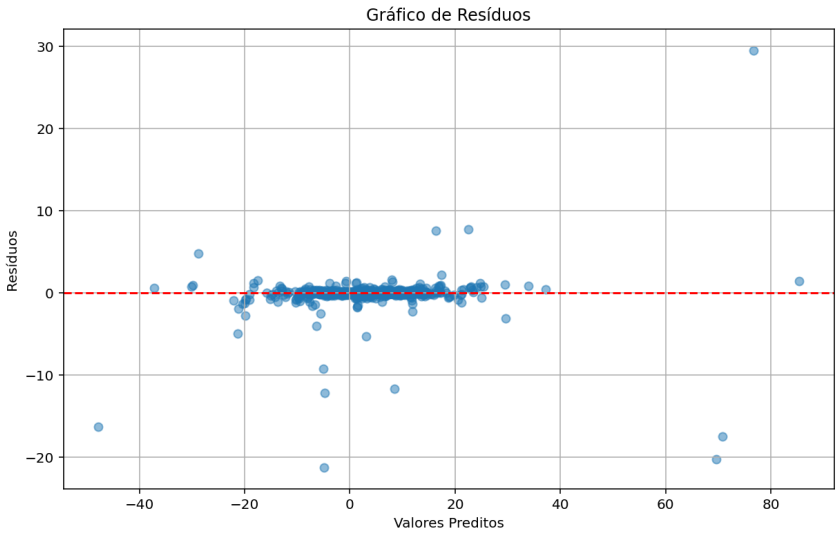

Figura 5.3: Análise de Desempenho do modelo: Valores Reais vs Valores Preditos.

Capítulo 6

Conclusões

Esse estudo foi um exemplo do uso da linguagem *Python* e ferramentas *Pandas*, *Scikit-Learn*, *Spyder* e *Matplotlib* para a mineração de dados. A exemplo de alguns dos trabalhos relacionados, constatou-se que é possível aplicar um algoritmo de mineração de dados em uma base de dados obtida da *web* diretamente em código, sem necessidade de ambientes de persistência de dados ou da estruturação de dados em formato relacional, por exemplo.

As diversas classes e métodos presentes no *Scikit-Learn* proveram funcionalidades para a mineração de dados com muita flexibilidade, que permitiram realizar todos os objetivos específicos e testar todas as hipóteses do estudo, listadas no Capítulo 1.

Entre as diversas afirmações presentes na **Hipótese 1**, algumas são subjetivas, mas foram, na nossa análise, corroboradas pela grande quantidade de gráficos e imagens mostrando conhecimento sobre a base WDI ao longo do Capítulo 4.

6.1 Limitações

Uma limitação técnica mas também conceitual sobre esse estudo é que ele é incapaz de analisar a progressão temporal dos indicadores para cada país e/ou região ao longo dos anos - todos os registros são analisados de forma isolada dos demais, sem o emprego de séries temporais.

O motivo é que, caso essa tal abordagem fosse escolhida, seria necessário um modelo de regressão separado para cada um dos países, causando um grande obstáculo com respeito à complexidade e tempo de execução dos algoritmos.

A falta da noção de continuidade temporal potencialmente atrapalhou o processo de inferência de valores vazios, pois é sensato pensar que um valor qualquer, uma vez ausente, deveria acompanhar, na média, os valores válidos dos anos mais próximos.

Um obstáculo técnico importante encontrado no projeto foi a de como utilizar a ferramenta *Matplotlib* para a construção dos gráficos usados no projeto. Em muitos casos, em

particular após a instanciação do modelo da Floresta Aleatória, há muitas informações disponíveis sobre o modelo em forma de atributos de classe, mas nenhuma forma predefinida de exibir a configuração da árvore em forma de gráfico, sendo necessário construí-lo de forma manual.

Outra limitação importante, mostrada na seção 5.1, é a presença de indicadores trivialmente dependentes da classe desejada na base de treinamento filtrada. Esses indicadores podem enviesar o modelo a ter uma acurácia muito alta, em detrimento da detecção de padrões não-óbvios sobre quais fatores mais influenciam o indicador desejado.

6.2 Trabalhos Futuros

Para trabalhos futuros que trabalhem com a mesma base com o objetivo de mineração de dados, são propostas sugestões de caminhos, tarefas e metodologias que ficaram de fora do escopo deste projeto.

A primeira sugestão é aplicar um modelo que utiliza informações geográficas sobre os países (por exemplo, a região a qual pertencem e sua área) como atributos válidos categóricos, em busca do descobrimento de padrões dentro das regiões e em países de regiões separadas.

Ainda no assunto das regiões geográficas, seria interessante executar um algoritmo de agrupamento, dentro do ramo dos modelos não-supervisionados, para agrupar países baseados nas suas semelhanças de valores aferidos nos indicadores, e comparar esse agrupamento com as fronteiras geográficas dos países, de forma a testar o "determinismo geográfico"de um país sobre os seus indicadores de crescimento do PIB ou de outros indicadores.

Uma proposta relacionada aos resultados da seção 5.1 é, ainda na etapa de pré-processamento, executar a remoção de indicadores diretamente dependentes do PIB (como aqueles que têm como sufixo o *"(% of GDP)"*). Tais indicadores, por definição, são dependentes do valor do Produto Interno Bruto (PIB) e sua taxa de crescimento. Remover da base de dados indicadores trivialmente correlacionados antes da execução da seleção de atributos potencialmente resultaria numa lista mais interessante, rica e com maior conhecimento agregado ao estudo.

Uma outra proposta de trabalho futuro é utilizar, para a predição dos indicadores, um outro tipo de algoritmo de regressão, o *Histogram-Based Gradient Boosting Regressor*, em complemento ou substituição ao algoritmo usado neste trabalho, o de Florestas Aleatórias. O Scikit-Learn possui um artigo em sua página dedicado à comparação entre eles [32].

Por último, fica também uma sugestão de uso dos chamados *hiperparâmetros*, técnica iterativa empregada em Aprendizado de Máquina para descobrir, testar e reforçar valores

de parâmetros em busca de achar as configurações ideais para o modelo, assim como de Validação Cruzada, uma metodologia iterativa de separação de conjuntos de teste e treinamento que visa reduzir a chance de *overfitting* e *underfitting* do modelo treinado.

Referências

[1] World Bank Group. Disponível em: `https://www.worldbank.org/en/home`. Acesso em: 13/07/2024. 1

[2] World Bank Group: *World Development Indicators*. Disponível em: `https://datatopics.worldbank.org/world-development-indicators/`. Acesso em: 13/07/2024. 2, 38

[3] Castro, Leandro N. e Daniel G. Ferrari: *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. Saraiva, 1ª edição, 2016. 5, 6, 8, 11, 14, 15, 17, 18, 19, 22, 26, 27, 28, 29, 30

[4] Laudon, Kenneth C. e Jane P. Laudon: *Sistemas de informação gerenciais*. Pearson Education do Brasil, 11ª edição, 2014. 6, 7

[5] Elmasri, Ramez e Shamkant B. Navathe: *Sistemas de banco de dados*. Pearson Education do Brasil, 6ª edição, 2011. 6, 7, 8, 9, 12

[6] Silberschatz, Abraham, Henry F. Korth e S. Sudarshan: *Database system concepts*. McGraw-Hill, 7ª edição, 2020. 8

[7] *What are the differences between data, a dataset, and a database?* Disponível em: `https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database`. Acesso em: 19/06/2024. 8

[8] Cavique, Luís: *Big data e data science*. Boletim da APDIO - Associação Portuguesa de Investigação Operacional, 51:11–14, 2014. 8, 26

[9] Inmon, William H., Derek Strauss e Genia Neushloss: *DW 2.0: the Architecture for the Next Generation of Data Warehousing*. Elsevier, 2008. 9

[10] Kimball, Ralph, Margy Ross, Warren Thornthwaite, Joy Mundy e Bob Becker: *The data warehouse lifecycle toolkit*. Wiley, 2ª edição, 2008. 10

[11] Santana, Matheus S. e Ytalo A. S. Carvalho: *Mineração de dados aplicados aos dados públicos do banco mundial*. Monografia (Graduação) de Bacharelado em Ciência da Computação. Universidade de Brasília (UnB), Brasília, Brasil, 2017. 10, 11, 33

[12] Larose, Daniel T. e Chantal D. Larose: *Discovering knowledge in data: an introduction to data mining*. Wiley, 2ª edição, 2014. 11, 12, 13, 16, 17, 23, 24, 30

[13] Han, Jiawei, Micheline Kamber e Jian Pei: *Data mining: concepts and techniques*. Elsevier, 3ª edição, 2012. 14, 17, 18, 19, 23, 24, 25, 26, 27, 28

[14] Russell, Stuart e Peter Norvig: *Inteligência Artificial - tradução da 3ª edição*. Elsevier, 2013. 19, 20, 21

[15] Geron, Aurélien: *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books, 2019. 21, 22, 23, 24, 25, 26

[16] Samuel, Arthur L.: *Some studies in machine learning using the game of checkers*. IBM Journal of Research and Development, 3(3):210–229, 1959. 21

[17] Mitchell, Tom M.: *Machine Learning*. McGraw-Hill, 1997. 21

[18] Gilgoldm, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons. *File:File:Decision Tree.jpg*. Disponível em: `https://commons.wikimedia.org/wiki/File:Decision_Tree.jpg`. Acesso em: 27/08/2024. 24

[19] *datasets/titanic.csv*. Disponível em: `https://github.com/datasciencedojo/datasets/blob/master/titanic.csv`. Acesso em: 27/08/2024. 24

[20] Traskas, Georgios: *Influence of feature selection and pca on a small dataset*. Disponível em: `https://gtraskas.github.io/post/titanic_prediction/`. Acesso em: 19/06/2024. 27

[21] Hirano, Felipe Ken e Iuri B. Beserra: *Análise dos votos dos senadores por meio de mineração de dados*. Monografia (Graduação) de Computação - Licenciatura. Universidade de Brasília (UnB), Brasília, Brasil, 2018. 31, 32

[22] Alves, Hudson S.: *Mineração de dados do portal if.data do bacen relativos às instituições financeiras que operam no brasil*. Monografia (Graduação) de Computação - Licenciatura. Universidade de Brasília (UnB), Brasília, Brasil, 2021. 32, 33

[23] Porto, Klark G. S.: *Descoberta de conhecimento através da análise e mineração em dados do enem*. Monografia (Graduação) de Engenharia da Computação. Universidade de Brasília (UnB), Brasília, Brasil, 2019. 34

[24] Python. Disponível em: `https://www.python.org/about/`. Acesso em: 28/08/2024. 36

[25] Scikit-Learn. Disponível em: `https://scikit-learn.org/stable/index.html`. Acesso em: 28/08/2024. 36

[26] Spyder. Disponível em: `https://www.spyder-ide.org/`. Acesso em: 28/08/2024. 36

[27] World Bank Group: *Base de dados World Development Indicators*. Versão 28/06/2024. Documento eletrônico disponível em: `https://datacatalogfiles.worldbank.org/ddh-published/0037712/DR0045575/WDI_CSV_2024_06_28.zip?versionId=2024-07-01T13:30:38.4396512Z`. Acesso em: 13/07/2024. 37

[28] Scikit-Learn: Nearest neighbors imputation. Disponível em: https://scikit-learn.org/stable/modules/impute.html#knnimpute. Acesso em: 07/08/2024. 45

[29] Scikit-Learn: Train Test Split. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. Acesso em: 07/08/2024. 46

[30] Scikit-Learn: Feature selection. Disponível em: https://scikit-learn.org/stable/modules/feature_selection.html. Acesso em: 07/08/2024. 47

[31] Scikit-Learn: Random Forest Regressor. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Acesso em: 08/08/2024. 48

[32] Scikit-Learn: Comparing Random Forests and Histogram Gradient Boosting models. Disponível em: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_hist_grad_boosting_comparison.html#sphx-glr-auto-examples-ensemble-plot-forest-hist-grad-boosting-comparison-py. Acesso em: 08/08/2024. 56

Apêndice A

Script para obter e tratar a base original

```
""" Módulo com as funções para extrair
a base de dados de sua pasta original.
"""
import pandas as pd

EXTRACTS_PATH = '../Data/WDI_CSV_2024_06_28/'
RAW_FILENAME = 'WDICSV.csv'

WORKSPACE_PATH = './dataframes/'
MAIN_FILENAME = 'WDItratado.csv'

def get_wdi_dataframe():
    """
    Lê os dados extraídos em csv e o transforma em um dataframe Pandas,
    além de realizar o tratamento com melt+pivot para
    mover os anos para representação em linhas
    e os atributos (indicadores) para representação em colunas.
    """
    return pd \
        .read_csv(
            EXTRACTS_PATH + RAW_FILENAME,
            usecols= lambda col: col!='Indicator Name'
            ) \
        .melt(
            id_vars=['Country Name', 'Country Code', 'Indicator Code'],
            var_name='Year',
            value_name = 'Value'
            ) \
        .pivot(
```

```
30              index=['Country Name', 'Country Code', 'Year'],
31              columns='Indicator Code',
32              values='Value'
33              ) \
34          .reset_index()
35
36
37  data = get_wdi_dataframe()
38  data.to_csv(WORKSPACE_PATH + MAIN_FILENAME, index=False)
```

gera_dataframes.py

Apêndice B

Script da construção do modelo com *Scikit-Learn*

```
""" O projeto em si. Usa o framework Pandas para mineração de dados dos
    indicadores do WorldBank.
"""
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.feature_selection import SelectKBest, r_regression
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.impute import KNNImputer

### Parâmetros

DATAFRAMES_PATH = './dataframes/'
TABLES_PATH = './material_overleaf/tabelas/'

MAIN_DF_PATH = DATAFRAMES_PATH + 'WDItratado.csv'
RAW_DF_PATH = DATAFRAMES_PATH + 'WDICSV.csv'
COUNTRIES_PATH = DATAFRAMES_PATH + 'WDICountry.csv'
INDICATORS_PATH = DATAFRAMES_PATH + 'WDISeries.csv'


YEARS_TO_DROP = 16                  # 1/4 do total de anos
COUNTRIES_TO_DROP = 28              # aprox. 10% dos 266 países e regiõ
    es
INDICATORS_NOT_NAN_THRESHOLD = 0.6  # exclui todo indicador com + de 40%
     valores nulos
KNN_IMPUTER_NEIGHBOURS = 10
TEST_SET_RATIO = 0.25
```

```
FEATURES_TO_SELECT = 32

### Extração dos dados

wdi = pd.read_csv(MAIN_DF_PATH)
raw_wdi = pd.read_csv(RAW_DF_PATH)
countries = pd.read_csv(COUNTRIES_PATH, index_col='Country Code')
indicators = pd.read_csv(INDICATORS_PATH, index_col='Series Code')
indicators['Indicator Name'] = indicators['Indicator Name'].map(lambda x
    : x if len(x) <= 49 else x[:49] + '...')
indicators.sort_values(['Topic']).to_csv(
    TABLES_PATH + 'indicadoresfull.csv',
    columns = ['Indicator Name']
)


### Variáveis para análise sobre o dataset inicial

total_indicators = len(wdi.columns) - 3
total_countries = len(wdi.groupby('Country Code'))
total_years = len(wdi.groupby('Year'))
total_nan = wdi.isna().sum().sum()
total_values = total_countries * total_indicators * total_years

# Dataframe que mostra a qtd. de valores vazios para cada indicador
nan_per_indicator = wdi.isna().sum()[3:] \
    .to_frame().rename(columns={0: 'NaN values'})
nan_per_indicator.insert(0, 'Name', indicators['Indicator Name'])
nan_per_indicator.insert(2, 'Percentage', (nan_per_indicator['NaN values
    '] / len(wdi) * 100).round(2))

# Série que mostra a qtd. de valores vazios por ano, somando todos os pa
    íses e indicadores
nan_per_year = raw_wdi.isna().sum()[4:]

# Série que mostra a qtd. de valores vazios para cada país, somando
    todos os anos
nan_per_country = wdi.groupby(['Country Code', 'Country Name']) \
    .count() \
    .drop(columns='Year') \
    .sum(axis=1) \
    .apply(lambda x: total_indicators * total_years - x) \
    .to_frame('NaN values') \
    .reset_index()

```

```
68
69 ### Criação de gráficos e tabelas para análise sobre o dataset inicial
70
71 nan_per_indicator['NaN values'].plot.hist(
72     xlabel='Qtd. de valores nulos',
73     ylabel='Frequência de indicadores',
74     bins=20,
75     grid=True)
76 nan_per_year.plot(
77     xlabel='Ano',
78     ylabel='Valores nulos',
79     ylim=(0, 400000),
80     grid=True)
81
82
83 ### Pré-processamento
84
85 emptiest_indicators = nan_per_indicator.nlargest(50, 'NaN values')
86 emptiest_years = nan_per_year.nlargest(YEARS_TO_DROP)
87 emptiest_countries = nan_per_country.nlargest(COUNTRIES_TO_DROP, 'NaN
       values')
88
89
90 # Exlui registros que possuem a variável "crescimento do PIB" (o alvo do
       modelo) vazia
91 [gdp_growth_code] = indicators.query("`Indicator Name` == 'GDP growth (
       annual %)'").index
92 wdi = wdi.dropna(subset=[gdp_growth_code])
93
94 # Remove os anos que possuem mais valores vazios, conforme parâmetro
95 wdi = wdi[~wdi['Year'].isin(emptiest_years.index.astype(int))]
96
97 # Remove os países que possuem mais valores vazios, conforme parâmetro
98 wdi = wdi[~wdi['Country Code'].isin(emptiest_countries['Country Code'])]
99
100 # Mantém apenas indicadores que possuem uma porcentagem de valores não-
       nulos, conforme parâmetro
101 wdi = wdi.dropna(axis=1, thresh=INDICATORS_NOT_NAN_THRESHOLD*len(wdi))
102
103
104
105
106 ### Processamento dos conjuntos de teste e treinamento
107
108 # Separa as variáveis de entrada (X) e variável alvo (y)
```

```
109 wdi = wdi.set_index(['Country Name', 'Country Code', 'Year'])
110 X = wdi.drop(columns=[gdp_growth_code])
111 y = wdi[gdp_growth_code]
112
113 # Normaliza o conjunto de entrada
114 # scaler = StandardScaler()
115 # X_scaled = scaler.fit_transform(X, y)
116
117 # Preenche os valores vazios no conjunto de entrada
118
119 imputer = KNNImputer(n_neighbors=KNN_IMPUTER_NEIGHBOURS, weights='
        uniform')
120 X_imputed = imputer.fit_transform(X)
121 X_imputed = pd.DataFrame(X_imputed, columns=X.columns, index=X.index)
122
123
124
125 # Separa em conjuntos de teste e treinamento
126 X_train, X_test, y_train, y_test = train_test_split(
127     X_imputed, y, test_size=TEST_SET_RATIO, random_state=0)
128
129 # Filtra os melhores indicadores, conforme parâmetro
130 feature_selector = SelectKBest(r_regression, k=FEATURES_TO_SELECT)
131 feature_selector.fit(X_train, y_train)
132 X_train_selected = pd.DataFrame(
133     feature_selector.transform(X_train),
134     columns = X_train.columns[feature_selector.get_support()],
135     index = X_train.index
136 )
137 X_test_selected = pd.DataFrame(
138     feature_selector.transform(X_test),
139     columns = X_test.columns[feature_selector.get_support()],
140     index = X_test.index
141 )
142
143 selected_indicators = indicators[indicators.index.isin(X_train_selected.
        columns)]
144 selected_indicators.to_csv(
145     TABLES_PATH +'selecaoIndicadores.csv',
146     columns = ['Indicator Name']
147 )
148
149
150
151 ### Aplicação dos modelos
```

```
152
153 random_forest = RandomForestRegressor(random_state=0)
154 random_forest.fit(X_train_selected, y_train)
155
156 score = random_forest.score(X_test_selected, y_test)
157
158
159
160 ### Criação de gráficos para análise sobre o resultado
161 ## Calcula a previsão sobre os dados de teste
162
163 y_pred = random_forest.predict(X_test_selected)
164
165 # Cria um gráfico de disperção
166 plt.scatter(y_test, y_pred, alpha=0.5)
167 plt.xlabel('Valores Reais')
168 plt.ylabel('Valores Preditos')
169 plt.title('Valores Reais vs. Valores Preditos')
170 plt.show()
171
172 ## Calcula a diferença entre os valores reais x valores preditos (
    residuos)
173
174 residuals = y_test - y_pred
175
176 # Cria um gráfico de resíduos
177 plt.figure(figsize=(10, 6))
178 plt.scatter(y_pred, residuals, alpha=0.5)
179 plt.axhline(y=0, color='r', linestyle='--')
180 plt.xlabel('Valores Preditos')
181 plt.ylabel('Resíduos')
182 plt.title('Gráfico de Resíduos')
183 plt.grid(True)
184 plt.show()
```

main.py

Apêndice C

Lista completa de indicadores socioeconômicos da base de dados WDI (em inglês)

Series Code	Indicator Name
BM.KLT.DINV.CD.WD	Foreign direct investment, net outflows (BoP, cur...
BX.PEF.TOTL.CD.WD	Portfolio equity, net inflows (BoP, current US$)
BM.KLT.DINV.WD.GD.ZS	Foreign direct investment, net outflows (% of GDP...
BX.KLT.DINV.WD.GD.ZS	Foreign direct investment, net inflows (% of GDP)
BN.KAC.EOMS.CD	Net errors and omissions (BoP, current US$)
BN.FIN.TOTL.CD	Net financial account (BoP, current US$)
BN.KLT.PTXL.CD	Portfolio Investment, net (BoP, current US$)
BN.RES.INCL.CD	Reserves and related items (BoP, current US$)
BN.TRF.KOGT.CD	Net capital account (BoP, current US$)
BN.KLT.DINV.CD	Foreign direct investment, net (BoP, current US$)
BX.KLT.DINV.CD.WD	Foreign direct investment, net inflows (BoP, curr...
BN.GSR.MRCH.CD	Net trade in goods (BoP, current US$)
BN.GSR.GNFS.CD	Net trade in goods and services (BoP, current US$...
BN.CAB.XOKA.GD.ZS	Current account balance (% of GDP)
BN.CAB.XOKA.CD	Current account balance (BoP, current US$)
BM.GSR.FCTY.CD	Primary income payments (BoP, current US$)
BM.GSR.GNFS.CD	Imports of goods and services (BoP, current US$)
BM.GSR.INSF.ZS	Insurance and financial services (% of service im...
BM.GSR.MRCH.CD	Goods imports (BoP, current US$)

BM.GSR.NFSV.CD	Service imports (BoP, current US$)
BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)
BM.GSR.TOTL.CD	Imports of goods, services and primary income (Bo...
BM.GSR.TRAN.ZS	Transport services (% of service imports, BoP)
BM.GSR.TRVL.ZS	Travel services (% of service imports, BoP)
BM.GSR.CMCP.ZS	Communications, computer, etc. (% of service impo...
BN.GSR.FCTY.CD	Net primary income (BoP, current US$)
BM.GSR.ROYL.CD	Charges for the use of intellectual property, pay...
BX.GSR.FCTY.CD	Primary income receipts (BoP, current US$)
BX.GSR.CMCP.ZS	Communications, computer, etc. (% of service expo...
BX.GSR.INSF.ZS	Insurance and financial services (% of service ex...
BX.GSR.MRCH.CD	Goods exports (BoP, current US$)
BX.GSR.NFSV.CD	Service exports (BoP, current US$)
BX.GSR.ROYL.CD	Charges for the use of intellectual property, rec...
BX.GSR.TOTL.CD	Exports of goods, services and primary income (Bo...
BX.GSR.TRAN.ZS	Transport services (% of service exports, BoP)
BX.GSR.TRVL.ZS	Travel services (% of service exports, BoP)
BX.GSR.GNFS.CD	Exports of goods and services (BoP, current US$)
BM.TRF.PWKR.CD.DT	Personal remittances, paid (current US$)
BM.TRF.PRVT.CD	Secondary income, other sectors, payments (BoP, c...
BX.TRF.PWKR.DT.GD.ZS	Personal remittances, received (% of GDP)
BX.TRF.CURR.CD	Secondary income receipts (BoP, current US$)
BX.TRF.PWKR.CD	Personal transfers, receipts (BoP, current US$)
BN.TRF.CURR.CD	Net secondary income (BoP, current US$)
BX.TRF.PWKR.CD.DT	Personal remittances, received (current US$)
BX.GRT.EXTA.CD.WD	Grants, excluding technical cooperation (BoP, cur...
FI.RES.TOTL.MO	Total reserves in months of imports
FI.RES.XGLD.CD	Total reserves minus gold (current US$)
BX.GRT.TECH.CD.WD	Technical cooperation grants (BoP, current US$)
FI.RES.TOTL.DT.ZS	Total reserves (% of total external debt)
FI.RES.TOTL.CD	Total reserves (includes gold, current US$)
DT.DOD.DSTC.CD	External debt stocks, short-term (DOD, current US...
DT.DOD.MIBR.CD	PPG, IBRD (DOD, current US$)
DT.DOD.MIDA.CD	PPG, IDA (DOD, current US$)
DT.DOD.MWBG.CD	IBRD loans and IDA credits (DOD, current US$)
DT.DOD.PVLX.CD	Present value of external debt (current US$)

DT.DOD.DPNG.CD	External debt stocks, private nonguaranteed (PNG)...
DT.DOD.DIMF.CD	Use of IMF credit (DOD, current US$)
DT.DOD.DECT.CD	External debt stocks, total (DOD, current US$)
DT.DOD.DPPG.CD	External debt stocks, public and publicly guarant...
DT.DOD.DLXF.CD	External debt stocks, long-term (DOD, current US$...
DT.DOD.DSTC.IR.ZS	Short-term debt (% of total reserves)
DT.TDS.DECT.EX.ZS	Total debt service (% of exports of goods, servic...
DT.TDS.DECT.GN.ZS	Total debt service (% of GNI)
DT.TDS.DPPG.GN.ZS	Public and publicly guaranteed debt service (% of...
DT.TDS.DPPG.XP.ZS	Public and publicly guaranteed debt service (% of...
DT.TDS.MLAT.PG.ZS	Multilateral debt service (% of public and public...
DT.DOD.DSTC.XP.ZS	Short-term debt (% of exports of goods, services ...
DT.DOD.DSTC.ZS	Short-term debt (% of total external debt)
DT.DOD.DECT.GN.ZS	External debt stocks (% of GNI)
DT.DOD.PVLX.GN.ZS	Present value of external debt (% of GNI)
DT.DOD.PVLX.EX.ZS	Present value of external debt (% of exports of g...
DT.TDS.DIMF.CD	IMF repurchases and charges (TDS, current US$)
DT.TDS.DPPF.XP.ZS	Debt service to exports (%)
DT.TDS.DECT.CD	Debt service on external debt, total (TDS, curren...
DT.TDS.MLAT.CD	Multilateral debt service (TDS, current US$)
DT.TDS.DPPG.CD	Debt service on external debt, public and publicl...
DT.NFL.PBND.CD	PPG, bonds (NFL, current US$)
DT.NFL.RDBC.CD	Net financial flows, RDB concessional (NFL, curre...
DT.NFL.PRVT.CD	PPG, private creditors (NFL, current US$)
DT.NFL.PROP.CD	PPG, other private creditors (NFL, current US$)
DT.NFL.PNGC.CD	PNG, commercial banks and other creditors (NFL, c...
DT.NFL.PNGB.CD	PNG, bonds (NFL, current US$)
DT.NFL.PCBO.CD	Commercial banks and other lending (PPG + PNG) (N...
DT.NFL.OFFT.CD	PPG, official creditors (NFL, current US$)
DT.NFL.DPNG.CD	Net flows on external debt, private nonguaranteed...
DT.NFL.MOTH.CD	Net financial flows, others (NFL, current US$)
DT.NFL.MLAT.CD	Net financial flows, multilateral (NFL, current U...
DT.NFL.MIDA.CD	Net financial flows, IDA (NFL, current US$)
DT.NFL.BOND.CD	Portfolio investment, bonds (PPG + PNG) (NFL, cur...
DT.NFL.IMFN.CD	Net financial flows, IMF nonconcessional (NFL, cu...
DT.NFL.IMFC.CD	Net financial flows, IMF concessional (NFL, curre...

DT.NFL.RDBN.CD	Net financial flows, RDB nonconcessional (NFL, cu...
DT.NFL.BLAT.CD	Net financial flows, bilateral (NFL, current US$)
DT.NFL.NIFC.CD	IFC, private nonguaranteed (NFL, current US$)
DT.NFL.MIBR.CD	Net financial flows, IBRD (NFL, current US$)
DT.NFL.PCBK.CD	PPG, commercial banks (NFL, current US$)
NY.ADJ.DMIN.GN.ZS	Adjusted savings: mineral depletion (% of GNI)
NY.ADJ.SVNX.GN.ZS	Adjusted net savings, excluding particulate emiss...
NY.ADJ.SVNX.CD	Adjusted net savings, excluding particulate emiss...
NY.ADJ.SVNG.GN.ZS	Adjusted net savings, including particulate emiss...
NY.ADJ.SVNG.CD	Adjusted net savings, including particulate emiss...
NY.ADJ.NNTY.PC.KD.ZG	Adjusted net national income per capita (annual %...
NY.ADJ.NNTY.PC.KD	Adjusted net national income per capita (constant...
NY.ADJ.NNTY.PC.CD	Adjusted net national income per capita (current ...
NY.ADJ.NNTY.KD.ZG	Adjusted net national income (annual % growth)
NY.ADJ.NNTY.KD	Adjusted net national income (constant 2015 US$)
NY.ADJ.DNGY.GN.ZS	Adjusted savings: energy depletion (% of GNI)
NY.ADJ.NNAT.GN.ZS	Adjusted savings: net national savings (% of GNI)
NY.ADJ.NNAT.CD	Adjusted savings: net national savings (current U...
NY.ADJ.ICTR.GN.ZS	Adjusted savings: gross savings (% of GNI)
NY.ADJ.NNTY.CD	Adjusted net national income (current US$)
NY.ADJ.DPEM.GN.ZS	Adjusted savings: particulate emission damage (% ...
NY.ADJ.DMIN.CD	Adjusted savings: mineral depletion (current US$)
NY.ADJ.DRES.GN.ZS	Adjusted savings: natural resources depletion (% ...
NY.ADJ.DKAP.CD	Adjusted savings: consumption of fixed capital (c...
NY.ADJ.DFOR.GN.ZS	Adjusted savings: net forest depletion (% of GNI)
NY.ADJ.DFOR.CD	Adjusted savings: net forest depletion (current U...
NY.ADJ.DKAP.GN.ZS	Adjusted savings: consumption of fixed capital (%...
NY.ADJ.DCO2.CD	Adjusted savings: carbon dioxide damage (current ...
NY.ADJ.AEDU.GN.ZS	Adjusted savings: education expenditure (% of GNI...
NY.ADJ.AEDU.CD	Adjusted savings: education expenditure (current ...
NY.ADJ.DPEM.CD	Adjusted savings: particulate emission damage (cu...
NY.ADJ.DCO2.GN.ZS	Adjusted savings: carbon dioxide damage (% of GNI...
NY.ADJ.DNGY.CD	Adjusted savings: energy depletion (current US$)
NY.GNP.ATLS.CD	GNI, Atlas method (current US$)
NY.GNP.PCAP.CD	GNI per capita, Atlas method (current US$)
NV.SRV.TOTL.KD.ZG	Services, value added (annual % growth)

NY.GDP.MKTP.KD.ZG	GDP growth (annual %)
NY.GDP.PCAP.KD.ZG	GDP per capita growth (annual %)
NY.GNP.MKTP.KD.ZG	GNI growth (annual %)
NY.GNP.PCAP.KD.ZG	GNI per capita growth (annual %)
NE.CON.GOVT.KD.ZG	General government final consumption expenditure ...
NE.EXP.GNFS.KD.ZG	Exports of goods and services (annual % growth)
NV.IND.TOTL.KD.ZG	Industry (including construction), value added (a...
NE.GDI.TOTL.KD.ZG	Gross capital formation (annual % growth)
NE.GDI.FTOT.KD.ZG	Gross fixed capital formation (annual % growth)
NE.CON.TOTL.KD.ZG	Final consumption expenditure (annual % growth)
NE.CON.PRVT.PC.KD.ZG	Household final consumption expenditure per capit...
NV.IND.MANF.KD.ZG	Manufacturing, value added (annual % growth)
NV.AGR.TOTL.KD.ZG	Agriculture, forestry, and fishing, value added (...
NE.CON.PRVT.KD.ZG	Household and NPISHs Final consumption expenditur...
NE.IMP.GNFS.KD.ZG	Imports of goods and services (annual % growth)
NY.TRF.NCTR.KN	Net secondary income (Net current transfers from ...
NY.GSR.NFCY.KN	Net primary income (Net income from abroad) (cons...
NY.GNP.MKTP.KN	GNI (constant LCU)
NY.GDP.FCST.KN	Gross value added at basic prices (GVA) (constant...
NY.TAX.NIND.KN	Taxes less subsidies on products (constant LCU)
NY.GNP.PCAP.KN	GNI per capita (constant LCU)
NY.GDP.PCAP.KN	GDP per capita (constant LCU)
NY.GDP.MKTP.KN	GDP (constant LCU)
NE.GDI.STKB.KN	Changes in inventories (constant LCU)
NE.EXP.GNFS.KN	Exports of goods and services (constant LCU)
NE.CON.PRVT.KN	Household and NPISHs Final consumption expenditur...
NE.GDI.TOTL.KN	Gross capital formation (constant LCU)
NE.DAB.TOTL.KN	Gross national expenditure (constant LCU)
NE.CON.TOTL.KN	Final consumption expenditure (constant LCU)
NE.RSB.GNFS.KN	External balance on goods and services (constant ...
NE.CON.GOVT.KN	General government final consumption expenditure ...
NY.GDP.DISC.KN	Discrepancy in expenditure estimate of GDP (const...
NE.GDI.FTOT.KN	Gross fixed capital formation (constant LCU)
NE.IMP.GNFS.KN	Imports of goods and services (constant LCU)
NY.GDY.TOTL.KN	Gross domestic income (constant LCU)
NY.TTF.GNFS.KN	Terms of trade adjustment (constant LCU)

NY.EXP.CAPM.KN	Exports as a capacity to import (constant LCU)
NV.AGR.TOTL.KN	Agriculture, forestry, and fishing, value added (...
NV.SRV.TOTL.KN	Services, value added (constant LCU)
NV.IND.MANF.KN	Manufacturing, value added (constant LCU)
NV.IND.TOTL.KN	Industry (including construction), value added (c...
NV.FSM.TOTL.KN	Financial intermediary services indirectly Measur...
NY.GSR.NFCY.CN	Net primary income (Net income from abroad) (curr...
NY.GNP.PCAP.CN	GNI per capita (current LCU)
NY.GDS.TOTL.CN	Gross domestic savings (current LCU)
NY.TRF.NCTR.CN	Net secondary income (Net current transfers from ...
NY.GDP.FCST.CN	Gross value added at basic prices (GVA) (current ...
NY.GNP.MKTP.CN	GNI (current LCU)
NY.GDP.MKTP.CN	GDP (current LCU)
NY.GDP.MKTP.CN.AD	GDP: linked series (current LCU)
NY.GDP.PCAP.CN	GDP per capita (current LCU)
NY.TAX.NIND.CN	Taxes less subsidies on products (current LCU)
NY.GNS.ICTR.CN	Gross savings (current LCU)
NE.GDI.FPRV.CN	Gross fixed capital formation, private sector (cu...
NE.CON.TOTL.CN	Final consumption expenditure (current LCU)
NY.GDP.DISC.CN	Discrepancy in expenditure estimate of GDP (curre...
NY.GNP.MKTP.CN.AD	GNI: linked series (current LCU)
NE.EXP.GNFS.CN	Exports of goods and services (current LCU)
NE.GDI.STKB.CN	Changes in inventories (current LCU)
NE.DAB.TOTL.CN	Gross national expenditure (current LCU)
NE.GDI.TOTL.CN	Gross capital formation (current LCU)
NE.GDI.FTOT.CN	Gross fixed capital formation (current LCU)
NE.IMP.GNFS.CN	Imports of goods and services (current LCU)
NE.CON.PRVT.CN.AD	Households and NPISHs final consumption expenditu...
NE.CON.PRVT.CN	Household and NPISHs Final consumption expenditur...
NE.RSB.GNFS.CN	External balance on goods and services (current L...
NE.CON.GOVT.CN	General government final consumption expenditure ...
NV.IND.TOTL.CN	Industry (including construction), value added (c...
NV.IND.MANF.CN	Manufacturing, value added (current LCU)
NV.FSM.TOTL.CN	Financial intermediary services indirectly Measur...
NV.SRV.TOTL.CN	Services, value added (current LCU)
NV.AGR.TOTL.CN	Agriculture, forestry, and fishing, value added (...

NY.GNS.ICTR.GN.ZS	Gross savings (% of GNI)
NV.MNF.TXTL.ZS.UN	Textiles and clothing (% of value added in manufa...
NV.IND.TOTL.ZS	Industry (including construction), value added (%...
NV.MNF.OTHR.ZS.UN	Other manufacturing (% of value added in manufact...
NV.MNF.CHEM.ZS.UN	Chemicals (% of value added in manufacturing)
NV.MNF.FBTO.ZS.UN	Food, beverages and tobacco (% of value added in ...
NV.MNF.MTRN.ZS.UN	Machinery and transport equipment (% of value add...
NE.CON.GOVT.ZS	General government final consumption expenditure ...
NV.MNF.TECH.ZS.UN	Medium and high-tech manufacturing value added (%...
NE.DAB.DEFL.ZS	Gross national expenditure deflator (base year va...
NE.CON.TOTL.ZS	Final consumption expenditure (% of GDP)
NV.IND.MANF.ZS	Manufacturing, value added (% of GDP)
NY.GDS.TOTL.ZS	Gross domestic savings (% of GDP)
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
NE.GDI.FTOT.ZS	Gross fixed capital formation (% of GDP)
NE.RSB.GNFS.ZS	External balance on goods and services (% of GDP)
NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
NE.TRD.GNFS.ZS	Trade (% of GDP)
NV.SRV.TOTL.ZS	Services, value added (% of GDP)
NY.GNS.ICTR.ZS	Gross savings (% of GDP)
NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (...
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NE.CON.PRVT.ZS	Households and NPISHs final consumption expenditu...
NE.DAB.TOTL.ZS	Gross national expenditure (% of GDP)
NE.GDI.FPRV.ZS	Gross fixed capital formation, private sector (% ...
NY.GDP.FCST.KD	Gross value added at basic prices (GVA) (constant...
NY.GDP.MKTP.KD	GDP (constant 2015 US$)
NY.GDP.PCAP.KD	GDP per capita (constant 2015 US$)
NY.GNP.MKTP.KD	GNI (constant 2015 US$)
NY.GNP.PCAP.KD	GNI per capita (constant 2015 US$)
NE.EXP.GNFS.KD	Exports of goods and services (constant 2015 US$)
NE.CON.GOVT.KD	General government final consumption expenditure ...
NE.CON.PRVT.KD	Households and NPISHs Final consumption expenditu...
NE.DAB.TOTL.KD	Gross national expenditure (constant 2015 US$)
NE.CON.PRVT.PC.KD	Households and NPISHs final consumption expenditu...
NE.CON.TOTL.KD	Final consumption expenditure (constant 2015 US$)

NE.GDI.FTOT.KD	Gross fixed capital formation (constant 2015 US$)
NE.IMP.GNFS.KD	Imports of goods and services (constant 2015 US$)
NE.GDI.TOTL.KD	Gross capital formation (constant 2015 US$)
NV.SRV.EMPL.KD	Services, value added per worker (constant 2015 U...
NV.IND.TOTL.KD	Industry (including construction), value added (c...
NV.IND.MANF.KD	Manufacturing, value added (constant 2015 US$)
NV.IND.EMPL.KD	Industry (including construction), value added pe...
NV.AGR.TOTL.KD	Agriculture, forestry, and fishing, value added (...
NV.SRV.TOTL.KD	Services, value added (constant 2015 US$)
NV.AGR.EMPL.KD	Agriculture, forestry, and fishing, value added p...
NY.TRF.NCTR.CD	Net secondary income (Net current transfers from ...
NY.GSR.NFCY.CD	Net primary income (Net income from abroad) (curr...
NY.TAX.NIND.CD	Taxes less subsidies on products (current US$)
NY.GNS.ICTR.CD	Gross savings (current US$)
NY.GDP.FCST.CD	Gross value added at basic prices (GVA) (current ...
NY.GDP.MKTP.CD	GDP (current US$)
NY.GNP.MKTP.CD	GNI (current US$)
NY.GDP.PCAP.CD	GDP per capita (current US$)
NY.GDS.TOTL.CD	Gross domestic savings (current US$)
NE.GDI.TOTL.CD	Gross capital formation (current US$)
NE.RSB.GNFS.CD	External balance on goods and services (current U...
NE.GDI.STKB.CD	Changes in inventories (current US$)
NE.GDI.FTOT.CD	Gross fixed capital formation (current US$)
NE.EXP.GNFS.CD	Exports of goods and services (current US$)
NE.CON.TOTL.CD	Final consumption expenditure (current US$)
NE.DAB.TOTL.CD	Gross national expenditure (current US$)
NE.CON.PRVT.CD	Household and NPISHs Final consumption expenditur...
NE.IMP.GNFS.CD	Imports of goods and services (current US$)
NE.CON.GOVT.CD	General government final consumption expenditure ...
NV.SRV.TOTL.CD	Services, value added (current US$)
NV.IND.TOTL.CD	Industry (including construction), value added (c...
NV.IND.MANF.CD	Manufacturing, value added (current US$)
NV.AGR.TOTL.CD	Agriculture, forestry, and fishing, value added (...
DT.ODA.ODAT.GN.ZS	Net ODA received (% of GNI)
DT.NFL.UNDP.CD	Net official flows from UN agencies, UNDP (curren...
DT.NFL.UNCV.CD	Net official flows from UN agencies, UNCOVID (cur...

DT.NFL.UNCTAD.CD	Net official flows from UN agencies, UNCTAD (curr...
DT.NFL.UNCR.CD	Net official flows from UN agencies, UNHCR (curre...
DT.ODA.ODAT.MP.ZS	Net ODA received (% of imports of goods, services...
DT.NFL.UNCF.CD	Net official flows from UN agencies, UNICEF (curr...
DT.NFL.UNCD.CD	Net official flows from UN agencies, UNCDF (curre...
DT.NFL.UNAI.CD	Net official flows from UN agencies, UNAIDS (curr...
DT.NFL.SPRP.CD	Net official flows from UN agencies, SPRP (curren...
DT.NFL.SDGF.CD	Net official flows from UN agencies, SDGFUND (cur...
DT.ODA.ODAT.XP.ZS	Net ODA received (% of central government expense...
DT.NFL.ILOG.CD	Net official flows from UN agencies, ILO (current...
DT.NFL.IFAD.CD	Net official flows from UN agencies, IFAD (curren...
DT.NFL.IAEA.CD	Net official flows from UN agencies, IAEA (curren...
DT.NFL.FAOG.CD	Net official flows from UN agencies, FAO (current...
DT.NFL.UNEC.CD	Net official flows from UN agencies, UNECE (curre...
DT.NFL.UNEP.CD	Net official flows from UN agencies, UNEP (curren...
DT.NFL.UNFP.CD	Net official flows from UN agencies, UNFPA (curre...
DT.NFL.UNID.CD	Net official flows from UN agencies, UNIDIR (curr...
DT.ODA.ODAT.GI.ZS	Net ODA received (% of gross capital formation)
DT.ODA.ODAT.CD	Net official development assistance received (cur...
DT.ODA.OATL.KD	Net official aid received (constant 2021 US$)
DT.ODA.OATL.CD	Net official aid received (current US$)
DT.ODA.ALLD.KD	Net official development assistance and official ...
DT.ODA.ALLD.CD	Net official development assistance and official ...
DT.NFL.WITC.CD	Net official flows from UN agencies, WTO-ITC (cur...
DT.ODA.ODAT.KD	Net official development assistance received (con...
DT.NFL.WHOL.CD	Net official flows from UN agencies, WHO (current...
DT.NFL.UNWT.CD	Net official flows from UN agencies, UNWTO (curre...
DT.NFL.UNWN.CD	Net official flows from UN agencies, UNWOMEN (cur...
DT.NFL.UNTA.CD	Net official flows from UN agencies, UNTA (curren...
DT.NFL.UNRW.CD	Net official flows from UN agencies, UNRWA (curre...
DT.NFL.UNPB.CD	Net official flows from UN agencies, UNPBF (curre...
DT.NFL.UNIDO.CD	Net official flows from UN agencies, UNIDO (curre...
DT.NFL.CERF.CD	Net official flows from UN agencies, CERF (curren...
DT.NFL.WFPG.CD	Net official flows from UN agencies, WFP (current...
DC.ODA.TOTL.KD	Net ODA provided, total (constant 2021 US$)
DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)

DC.ODA.TOTL.CD	Net ODA provided, total (current US$)
DC.DAC.HUNL.CD	Net bilateral aid flows from DAC donors, Hungary ...
DC.DAC.GRCL.CD	Net bilateral aid flows from DAC donors, Greece (...
DC.ODA.TOTL.GN.ZS	Net ODA provided, total (% of GNI)
DC.DAC.FRAL.CD	Net bilateral aid flows from DAC donors, France (...
DC.DAC.FINL.CD	Net bilateral aid flows from DAC donors, Finland ...
DC.DAC.ESTL.CD	Net bilateral aid flows from DAC donors, Estonia ...
DC.DAC.ESPL.CD	Net bilateral aid flows from DAC donors, Spain (c...
DC.DAC.IRLL.CD	Net bilateral aid flows from DAC donors, Ireland ...
DC.DAC.DNKL.CD	Net bilateral aid flows from DAC donors, Denmark ...
DC.DAC.CZEL.CD	Net bilateral aid flows from DAC donors, Czech Re...
DC.DAC.CHEL.CD	Net bilateral aid flows from DAC donors, Switzerl...
DC.DAC.CECL.CD	Net bilateral aid flows from DAC donors, European...
DC.DAC.CANL.CD	Net bilateral aid flows from DAC donors, Canada (...
DC.DAC.BELL.CD	Net bilateral aid flows from DAC donors, Belgium ...
DC.DAC.AUTL.CD	Net bilateral aid flows from DAC donors, Austria ...
DC.DAC.AUSL.CD	Net bilateral aid flows from DAC donors, Australi...
DC.DAC.DEUL.CD	Net bilateral aid flows from DAC donors, Germany ...
DC.DAC.ISLL.CD	Net bilateral aid flows from DAC donors, Iceland ...
DC.DAC.GBRL.CD	Net bilateral aid flows from DAC donors, United K...
DC.DAC.JPNL.CD	Net bilateral aid flows from DAC donors, Japan (c...
DC.ODA.TLDC.GN.ZS	Net ODA provided to the least developed countries...
DC.ODA.TLDC.CD	Net ODA provided, to the least developed countrie...
DC.DAC.USAL.CD	Net bilateral aid flows from DAC donors, United S...
DC.DAC.ITAL.CD	Net bilateral aid flows from DAC donors, Italy (c...
DC.DAC.SWEL.CD	Net bilateral aid flows from DAC donors, Sweden (...
DC.DAC.SVNL.CD	Net bilateral aid flows from DAC donors, Slovenia...
DC.DAC.SVKL.CD	Net bilateral aid flows from DAC donors, Slovak R...
DC.DAC.PRTL.CD	Net bilateral aid flows from DAC donors, Portugal...
DC.DAC.POLL.CD	Net bilateral aid flows from DAC donors, Poland (...
DC.DAC.TOTL.CD	Net bilateral aid flows from DAC donors, Total (c...
DC.DAC.NORL.CD	Net bilateral aid flows from DAC donors, Norway (...
DC.DAC.NLDL.CD	Net bilateral aid flows from DAC donors, Netherla...
DC.DAC.LUXL.CD	Net bilateral aid flows from DAC donors, Luxembou...
DC.DAC.LTUL.CD	Net bilateral aid flows from DAC donors, Lithuani...
DC.DAC.NZLL.CD	Net bilateral aid flows from DAC donors, New Zeal...

DC.DAC.KORL.CD	Net bilateral aid flows from DAC donors, Korea, R...
NY.GDP.MKTP.PP.KD	GDP, PPP (constant 2021 international $)
NY.GDP.MKTP.PP.CD	GDP, PPP (current international $)
NY.GNP.MKTP.PP.CD	GNI, PPP (current international $)
NY.GNP.MKTP.PP.KD	GNI, PPP (constant 2021 international $)
NE.CON.PRVT.PP.CD	Households and NPISHs Final consumption expenditu...
NE.CON.PRVT.PP.KD	Households and NPISHs Final consumption expenditu...
NY.GNP.PCAP.PP.KD	GNI per capita, PPP (constant 2021 international ...
PA.NUS.PRVT.PP	PPP conversion factor, private consumption (LCU p...
PA.NUS.PPPC.RF	Price level ratio of PPP conversion factor (GDP) ...
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international $)
PA.NUS.PPP	PPP conversion factor, GDP (LCU per international...
NY.GNP.PCAP.PP.CD	GNI per capita, PPP (current international $)
NY.GDP.PCAP.PP.KD	GDP per capita, PPP (constant 2021 international ...
SE.PRM.REPT.FE.ZS	Repeaters, primary, female (% of female enrollmen...
SE.SEC.PROG.ZS	Progression to secondary school (%)
SE.SEC.PROG.MA.ZS	Progression to secondary school, male (%)
SE.SEC.PROG.FE.ZS	Progression to secondary school, female (%)
SE.PRM.PRSL.ZS	Persistence to last grade of primary, total (% of...
SE.PRM.PRSL.MA.ZS	Persistence to last grade of primary, male (% of ...
SE.PRM.NINT.MA.ZS	Net intake rate in grade 1, male (% of official s...
SE.PRM.PRSL.FE.ZS	Persistence to last grade of primary, female (% o...
SE.PRM.REPT.ZS	Repeaters, primary, total (% of total enrollment)
SE.PRM.NINT.FE.ZS	Net intake rate in grade 1, female (% of official...
SE.PRM.GINT.ZS	Gross intake ratio in first grade of primary educ...
SE.PRM.GINT.MA.ZS	Gross intake ratio in first grade of primary educ...
SE.PRM.NINT.ZS	Net intake rate in grade 1 (% of official school-...
SE.PRM.GINT.FE.ZS	Gross intake ratio in first grade of primary educ...
SE.PRM.OENR.ZS	Over-age students, primary (% of enrollment)
SE.PRM.PRS5.FE.ZS	Persistence to grade 5, female (% of cohort)
SE.PRM.PRS5.MA.ZS	Persistence to grade 5, male (% of cohort)
SE.PRM.PRS5.ZS	Persistence to grade 5, total (% of cohort)
SE.PRM.REPT.MA.ZS	Repeaters, primary, male (% of male enrollment)
SE.PRM.OENR.MA.ZS	Over-age students, primary, male (% of male enrol...
SE.PRM.OENR.FE.ZS	Over-age students, primary, female (% of female e...
SE.SEC.TCAQ.LO.MA.ZS	Trained teachers in lower secondary education, ma...

SE.PRM.AGES	Primary school starting age (years)
SE.PRE.TCAQ.ZS	Trained teachers in preprimary education (% of to...
SE.PRE.TCAQ.MA.ZS	Trained teachers in preprimary education, male (%...
SE.PRE.TCAQ.FE.ZS	Trained teachers in preprimary education, female ...
SE.PRE.ENRL.TC.ZS	Pupil-teacher ratio, preprimary
SE.PRE.DURS	Preprimary education, duration (years)
SE.PRM.ENRL.TC.ZS	Pupil-teacher ratio, primary
SE.SEC.TCAQ.LO.ZS	Trained teachers in lower secondary education (% ...
SE.SEC.TCAQ.UP.FE.ZS	Trained teachers in upper secondary education, fe...
SE.SEC.TCAQ.UP.MA.ZS	Trained teachers in upper secondary education, ma...
SE.XPD.TOTL.GD.ZS	Government expenditure on education, total (% of ...
SE.XPD.TOTL.GB.ZS	Government expenditure on education, total (% of ...
SE.XPD.TERT.ZS	Expenditure on tertiary education (% of governmen...
SE.XPD.TERT.PC.ZS	Government expenditure per student, tertiary (% o...
SE.XPD.SECO.ZS	Expenditure on secondary education (% of governme...
SE.XPD.SECO.PC.ZS	Government expenditure per student, secondary (% ...
SE.XPD.PRIM.ZS	Expenditure on primary education (% of government...
SE.XPD.PRIM.PC.ZS	Government expenditure per student, primary (% of...
SE.XPD.CTOT.ZS	Current education expenditure, total (% of total ...
SE.XPD.CTER.ZS	Current education expenditure, tertiary (% of tot...
SE.XPD.CSEC.ZS	Current education expenditure, secondary (% of to...
SE.XPD.CPRM.ZS	Current education expenditure, primary (% of tota...
SE.TER.TCHR.FE.ZS	Tertiary education, academic staff (% female)
SE.TER.ENRL.TC.ZS	Pupil-teacher ratio, tertiary
SE.SEC.TCHR.FE.ZS	Secondary education, teachers (% female)
SE.SEC.TCHR.FE	Secondary education, teachers, female
SE.SEC.TCHR	Secondary education, teachers
SE.SEC.TCAQ.ZS	Trained teachers in secondary education (% of tot...
SE.SEC.TCAQ.UP.ZS	Trained teachers in upper secondary education (% ...
SE.SEC.TCAQ.MA.ZS	Trained teachers in secondary education, male (% ...
SE.PRM.TCAQ.FE.ZS	Trained teachers in primary education, female (% ...
SE.PRM.TCHR.FE.ZS	Primary education, teachers (% female)
SE.SEC.AGES	Lower secondary school starting age (years)
SE.SEC.TCAQ.LO.FE.ZS	Trained teachers in lower secondary education, fe...
SE.PRM.TCHR	Primary education, teachers
SE.PRM.TCAQ.ZS	Trained teachers in primary education (% of total...

SE.SEC.TCAQ.FE.ZS	Trained teachers in secondary education, female (...
SE.PRM.TCAQ.MA.ZS	Trained teachers in primary education, male (% of...
SE.SEC.ENRL.LO.TC.ZS	Pupil-teacher ratio, lower secondary
SE.SEC.ENRL.TC.ZS	Pupil-teacher ratio, secondary
SE.SEC.ENRL.UP.TC.ZS	Pupil-teacher ratio, upper secondary
SE.SEC.CUAT.UP.FE.ZS	Educational attainment, at least completed upper ...
SE.SEC.CMPT.LO.MA.ZS	Lower secondary completion rate, male (% of relev...
SE.SEC.CMPT.LO.ZS	Lower secondary completion rate, total (% of rele...
SE.SEC.CUAT.LO.FE.ZS	Educational attainment, at least completed lower ...
SE.SEC.CUAT.LO.MA.ZS	Educational attainment, at least completed lower ...
SE.SEC.CUAT.UP.ZS	Educational attainment, at least completed upper ...
SE.SEC.CUAT.UP.MA.ZS	Educational attainment, at least completed upper ...
SE.SEC.CUAT.LO.ZS	Educational attainment, at least completed lower ...
SE.SEC.CUAT.PO.FE.ZS	Educational attainment, at least completed post-s...
SE.SEC.CUAT.PO.MA.ZS	Educational attainment, at least completed post-s...
SE.TER.CUAT.BA.FE.ZS	Educational attainment, at least Bachelor's or eq...
SE.TER.CUAT.BA.ZS	Educational attainment, at least Bachelor's or eq...
SE.SEC.CMPT.LO.FE.ZS	Lower secondary completion rate, female (% of rel...
SE.TER.CUAT.DO.MA.ZS	Educational attainment, Doctoral or equivalent, p...
SE.TER.CUAT.DO.ZS	Educational attainment, Doctoral or equivalent, p...
SE.TER.CUAT.MS.FE.ZS	Educational attainment, at least Master's or equi...
SE.TER.CUAT.MS.MA.ZS	Educational attainment, at least Master's or equi...
SE.TER.CUAT.MS.ZS	Educational attainment, at least Master's or equi...
SE.TER.CUAT.ST.FE.ZS	Educational attainment, at least completed short-...
SE.TER.CUAT.ST.MA.ZS	Educational attainment, at least completed short-...
SE.TER.CUAT.ST.ZS	Educational attainment, at least completed short-...
SE.SEC.CUAT.PO.ZS	Educational attainment, at least completed post-s...
SE.TER.CUAT.BA.MA.ZS	Educational attainment, at least Bachelor's or eq...
SE.TER.CUAT.DO.FE.ZS	Educational attainment, Doctoral or equivalent, p...
SE.ADT.LITR.FE.ZS	Literacy rate, adult female (% of females ages 15...
SE.PRM.CUAT.FE.ZS	Educational attainment, at least completed primar...
SE.PRM.CMPT.ZS	Primary completion rate, total (% of relevant age...
SE.PRM.CMPT.MA.ZS	Primary completion rate, male (% of relevant age ...
SE.PRM.CMPT.FE.ZS	Primary completion rate, female (% of relevant ag...
SE.SEC.DURS	Secondary education, duration (years)
SE.ADT.1524.LT.FE.ZS	Literacy rate, youth female (% of females ages 15...

SE.LPV.PRIM	Learning poverty: Share of Children at the End-of...
SE.LPV.PRIM.FE	Learning poverty: Share of Female Children at the...
SE.LPV.PRIM.LD	Pupils below minimum reading proficiency at end o...
SE.LPV.PRIM.LD.FE	Female pupils below minimum reading proficiency a...
SE.LPV.PRIM.LD.MA	Male pupils below minimum reading proficiency at ...
SE.PRM.DURS	Primary education, duration (years)
SE.ADT.1524.LT.FM.ZS	Literacy rate, youth (ages 15-24), gender parity ...
SE.ADT.1524.LT.MA.ZS	Literacy rate, youth male (% of males ages 15-24)
SE.LPV.PRIM.SD.MA	Male primary school age children out-of-school (%...
SE.LPV.PRIM.SD.FE	Female primary school age children out-of-school ...
SE.LPV.PRIM.SD	Primary school age children out-of-school (%)
SE.PRM.CUAT.MA.ZS	Educational attainment, at least completed primar...
SE.LPV.PRIM.MA	Learning poverty: Share of Male Children at the E...
SE.PRM.CUAT.ZS	Educational attainment, at least completed primar...
SE.ADT.1524.LT.ZS	Literacy rate, youth total (% of people ages 15-2...
SE.ADT.LITR.MA.ZS	Literacy rate, adult male (% of males ages 15 and...
SE.ADT.LITR.ZS	Literacy rate, adult total (% of people ages 15 a...
SE.COM.DURS	Compulsory education, duration (years)
SE.SEC.ENRL.FE.ZS	Secondary education, pupils (% female)
SE.SEC.ENRL.GC	Secondary education, general pupils
SE.SEC.ENRL.GC.FE.ZS	Secondary education, general pupils (% female)
SE.ENR.PRIM.FM.ZS	School enrollment, primary (gross), gender parity...
SE.SEC.ENRL.VO	Secondary education, vocational pupils
SE.SEC.ENRL.VO.FE.ZS	Secondary education, vocational pupils (% female)
SE.ENR.TERT.FM.ZS	School enrollment, tertiary (gross), gender parit...
SE.SEC.NENR	School enrollment, secondary (% net)
SE.TER.ENRR	School enrollment, tertiary (% gross)
SE.ENR.PRSC.FM.ZS	School enrollment, primary and secondary (gross),...
SE.ENR.SECO.FM.ZS	School enrollment, secondary (gross), gender pari...
SE.SEC.ENRR	School enrollment, secondary (% gross)
SE.SEC.ENRR.FE	School enrollment, secondary, female (% gross)
SE.SEC.ENRR.MA	School enrollment, secondary, male (% gross)
SE.TER.ENRR.MA	School enrollment, tertiary, male (% gross)
SE.TER.ENRR.FE	School enrollment, tertiary, female (% gross)
SE.PRM.ENRL	Primary education, pupils
SE.SEC.UNER.LO.ZS	Adolescents out of school (% of lower secondary s...

SE.SEC.UNER.LO.FE.ZS	Adolescents out of school, female (% of female lo...
SE.PRM.ENRL.FE.ZS	Primary education, pupils (% female)
SE.PRM.TENR	Adjusted net enrollment rate, primary (% of prima...
SE.PRM.ENRR	School enrollment, primary (% gross)
SE.PRM.ENRR.FE	School enrollment, primary, female (% gross)
SE.PRM.ENRR.MA	School enrollment, primary, male (% gross)
SE.PRM.TENR.FE	Adjusted net enrollment rate, primary, female (% ...
SE.PRM.TENR.MA	Adjusted net enrollment rate, primary, male (% of...
SE.PRM.UNER	Children out of school, primary
SE.PRM.NENR	School enrollment, primary (% net)
SE.PRM.NENR.FE	School enrollment, primary, female (% net)
SE.SEC.UNER.LO.MA.ZS	Adolescents out of school, male (% of male lower ...
SE.PRM.NENR.MA	School enrollment, primary, male (% net)
SE.PRM.UNER.FE.ZS	Children out of school, female (% of female prima...
SE.PRM.UNER.MA	Children out of school, primary, male
SE.PRM.UNER.MA.ZS	Children out of school, male (% of male primary s...
SE.PRM.UNER.ZS	Children out of school (% of primary school age)
SE.PRM.PRIV.ZS	School enrollment, primary, private (% of total p...
SE.PRE.ENRR.MA	School enrollment, preprimary, male (% gross)
SE.PRE.ENRR.FE	School enrollment, preprimary, female (% gross)
SE.PRE.ENRR	School enrollment, preprimary (% gross)
SE.SEC.PRIV.ZS	School enrollment, secondary, private (% of total...
SE.SEC.NENR.MA	School enrollment, secondary, male (% net)
SE.SEC.NENR.FE	School enrollment, secondary, female (% net)
SE.PRM.UNER.FE	Children out of school, primary, female
SE.SEC.ENRL	Secondary education, pupils
SG.LAW.INDX	Women Business and the Law Index Score (scale 1-1...
AG.AGR.TRAC.NO	Agricultural machinery, tractors
AG.CON.FERT.ZS	Fertilizer consumption (kilograms per hectare of ...
AG.YLD.CREL.KG	Cereal yield (kg per hectare)
AG.PRD.LVSK.XD	Livestock production index (2014-2016 = 100)
ER.FSH.PROD.MT	Total fisheries production (metric tons)
ER.FSH.AQUA.MT	Aquaculture production (metric tons)
AG.PRD.FOOD.XD	Food production index (2014-2016 = 100)
ER.FSH.CAPT.MT	Capture fisheries production (metric tons)
AG.PRD.CREL.MT	Cereal production (metric tons)

AG.LND.TRAC.ZS	Agricultural machinery, tractors per 100 sq. km o...
AG.LND.CREL.HA	Land under cereal production (hectares)
AG.CON.FERT.PT.ZS	Fertilizer consumption (% of fertilizer productio...
AG.PRD.CROP.XD	Crop production index (2014-2016 = 100)
ER.LND.PTLD.ZS	Terrestrial protected areas (% of total land area...
EN.FSH.THRD.NO	Fish species, threatened
ER.PTD.TOTL.ZS	Terrestrial and marine protected areas (% of tota...
ER.MRN.PTMR.ZS	Marine protected areas (% of territorial waters)
EN.BIR.THRD.NO	Bird species, threatened
EN.HPT.THRD.NO	Plant species (higher), threatened
EN.MAM.THRD.NO	Mammal species, threatened
SP.RUR.TOTL.ZG	Rural population growth (annual %)
SP.RUR.TOTL.ZS	Rural population (% of total population)
SP.URB.GROW	Urban population growth (annual %)
SP.URB.TOTL.IN.ZS	Urban population (% of total population)
SP.RUR.TOTL	Rural population
EN.POP.SLUM.UR.ZS	Population living in slums (% of urban population...
SP.URB.TOTL	Urban population
EN.URB.LCTY	Population in largest city
EN.URB.LCTY.UR.ZS	Population in the largest city (% of urban popula...
EN.URB.MCTY	Population in urban agglomerations of more than 1...
EN.URB.MCTY.TL.ZS	Population in urban agglomerations of more than 1...
EN.POP.DNST	Population density (people per sq. km of land are...
EN.CO2.TRAN.ZS	CO2 emissions from transport (% of total fuel com...
EN.ATM.HFCG.KT.CE	HFC gas emissions (thousand metric tons of CO2 eq...
EN.ATM.GHGT.ZG	Total greenhouse gas emissions (% change from 199...
EN.ATM.GHGT.KT.CE	Total greenhouse gas emissions (kt of CO2 equival...
EN.ATM.GHGO.ZG	Other greenhouse gas emissions (% change from 199...
EN.ATM.GHGO.KT.CE	Other greenhouse gas emissions, HFC, PFC and SF6 ...
EN.ATM.CO2E.SF.ZS	CO2 emissions from solid fuel consumption (% of t...
EN.ATM.CO2E.SF.KT	CO2 emissions from solid fuel consumption (kt)
EN.ATM.CO2E.PP.GD.KD	CO2 emissions (kg per 2021 PPP $ of GDP)
EN.ATM.CO2E.LF.ZS	CO2 emissions from liquid fuel consumption (% of ...
EN.ATM.CO2E.PC	CO2 emissions (metric tons per capita)
EN.ATM.METH.AG.KT.CE	Agricultural methane emissions (thousand metric t...
EN.ATM.CO2E.LF.KT	CO2 emissions from liquid fuel consumption (kt)

EN.ATM.CO2E.KT	CO2 emissions (kt)
EN.ATM.CO2E.KD.GD	CO2 emissions (kg per 2015 US$ of GDP)
EN.ATM.CO2E.GF.ZS	CO2 emissions from gaseous fuel consumption (% of...
EN.ATM.CO2E.GF.KT	CO2 emissions from gaseous fuel consumption (kt)
EN.ATM.CO2E.EG.ZS	CO2 intensity (kg per kg of oil equivalent energy...
EN.ATM.CO2E.PP.GD	CO2 emissions (kg per PPP $ of GDP)
EN.ATM.METH.AG.ZS	Agricultural methane emissions (% of total)
EN.ATM.PM25.MC.T2.ZS	PM2.5 pollution, population exposed to levels exc...
EN.ATM.METH.EG.ZS	Energy related methane emissions (% of total)
EN.ATM.METH.EG.KT.CE	Methane emissions in energy sector (thousand metr...
EN.CLC.GHGR.MT.CE	GHG net emissions/removals by LUCF (Mt of CO2 equ...
EN.ATM.SF6G.KT.CE	SF6 gas emissions (thousand metric tons of CO2 eq...
EN.ATM.PM25.MC.ZS	PM2.5 air pollution, population exposed to levels...
EN.ATM.PM25.MC.T3.ZS	PM2.5 pollution, population exposed to levels exc...
EN.ATM.PM25.MC.M3	PM2.5 air pollution, mean annual exposure (microg...
EN.ATM.PFCG.KT.CE	PFC gas emissions (thousand metric tons of CO2 eq...
EN.CO2.BLDG.ZS	CO2 emissions from residential buildings and comm...
EN.CO2.ETOT.ZS	CO2 emissions from electricity and heat productio...
EN.ATM.PM25.MC.T1.ZS	PM2.5 pollution, population exposed to levels exc...
EN.CO2.OTHX.ZS	CO2 emissions from other sectors, excluding resid...
EN.ATM.NOXE.ZG	Nitrous oxide emissions (% change from 1990)
EN.ATM.METH.KT.CE	Methane emissions (kt of CO2 equivalent)
EN.ATM.NOXE.KT.CE	Nitrous oxide emissions (thousand metric tons of ...
EN.ATM.NOXE.EG.ZS	Nitrous oxide emissions in energy sector (% of to...
EN.ATM.NOXE.EG.KT.CE	Nitrous oxide emissions in energy sector (thousan...
EN.ATM.NOXE.AG.ZS	Agricultural nitrous oxide emissions (% of total)
EN.ATM.NOXE.AG.KT.CE	Agricultural nitrous oxide emissions (thousand me...
EN.CO2.MANF.ZS	CO2 emissions from manufacturing industries and c...
EN.ATM.METH.ZG	Methane emissions (% change from 1990)
EG.ELC.NUCL.ZS	Electricity production from nuclear sources (% of...
EG.ELC.LOSS.ZS	Electric power transmission and distribution loss...
EG.ELC.HYRO.ZS	Electricity production from hydroelectric sources...
EG.ELC.FOSL.ZS	Electricity production from oil, gas and coal sou...
EG.ELC.COAL.ZS	Electricity production from coal sources (% of to...
EG.EGY.PRIM.PP.KD	Energy intensity level of primary energy (MJ/$201...
EG.ELC.ACCS.UR.ZS	Access to electricity, urban (% of urban populati...

EG.ELC.ACCS.RU.ZS	Access to electricity, rural (% of rural populati...
EG.CFT.ACCS.ZS	Access to clean fuels and technologies for cookin...
EG.ELC.PETR.ZS	Electricity production from oil sources (% of tot...
EG.CFT.ACCS.UR.ZS	Access to clean fuels and technologies for cookin...
EG.CFT.ACCS.RU.ZS	Access to clean fuels and technologies for cookin...
EG.ELC.ACCS.ZS	Access to electricity (% of population)
EG.ELC.RNEW.ZS	Renewable electricity output (% of total electric...
EG.ELC.NGAS.ZS	Electricity production from natural gas sources (...
EG.ELC.RNWX.ZS	Electricity production from renewable sources, ex...
EG.ELC.RNWX.KH	Electricity production from renewable sources, ex...
EG.USE.PCAP.KG.OE	Energy use (kg of oil equivalent per capita)
EG.USE.ELEC.KH.PC	Electric power consumption (kWh per capita)
EG.USE.COMM.GD.PP.KD	Energy use (kg of oil equivalent) per $1,000 GDP ...
EG.USE.COMM.FO.ZS	Fossil fuel energy consumption (% of total)
EG.USE.CRNW.ZS	Combustible renewables and waste (% of total ener...
EG.IMP.CONS.ZS	Energy imports, net (% of energy use)
EG.GDP.PUSE.KO.PP.KD	GDP per unit of energy use (constant 2021 PPP $ p...
EG.GDP.PUSE.KO.PP	GDP per unit of energy use (PPP $ per kg of oil e...
EG.USE.COMM.CL.ZS	Alternative and nuclear energy (% of total energy...
EG.FEC.RNEW.ZS	Renewable energy consumption (% of total final en...
ER.H2O.INTR.K3	Renewable internal freshwater resources, total (b...
ER.H2O.FWTL.ZS	Annual freshwater withdrawals, total (% of intern...
ER.H2O.FWTL.K3	Annual freshwater withdrawals, total (billion cub...
ER.GDP.FWTL.M3.KD	Water productivity, total (constant 2015 US$ GDP ...
ER.H2O.FWST.ZS	Level of water stress: freshwater withdrawal as a...
ER.H2O.FWDM.ZS	Annual freshwater withdrawals, domestic (% of tot...
ER.H2O.FWIN.ZS	Annual freshwater withdrawals, industry (% of tot...
ER.H2O.INTR.PC	Renewable internal freshwater resources per capit...
ER.H2O.FWAG.ZS	Annual freshwater withdrawals, agriculture (% of ...
AG.SRF.TOTL.K2	Surface area (sq. km)
EN.CLC.MDAT.ZS	Droughts, floods, extreme temperatures (% of popu...
AG.LND.AGRI.ZS	Agricultural land (% of land area)
AG.LND.ARBL.HA	Arable land (hectares)
AG.LND.ARBL.HA.PC	Arable land (hectares per person)
AG.LND.ARBL.ZS	Arable land (% of land area)
EN.CLC.DRSK.XQ	Disaster risk reduction progress score (1-5 scale...

AG.LND.CROP.ZS	Permanent cropland (% of land area)
AG.LND.EL5M.RU.K2	Rural land area where elevation is below 5 meters...
AG.LND.EL5M.RU.ZS	Rural land area where elevation is below 5 meters...
AG.LND.EL5M.UR.K2	Urban land area where elevation is below 5 meters...
AG.LND.EL5M.ZS	Land area where elevation is below 5 meters (% of...
AG.LND.FRST.K2	Forest area (sq. km)
AG.LND.FRST.ZS	Forest area (% of land area)
AG.LND.IRIG.AG.ZS	Agricultural irrigated land (% of total agricultu...
AG.LND.PRCP.MM	Average precipitation in depth (mm per year)
AG.LND.TOTL.RU.K2	Rural land area (sq. km)
AG.LND.TOTL.UR.K2	Urban land area (sq. km)
AG.LND.AGRI.K2	Agricultural land (sq. km)
EN.POP.EL5M.UR.ZS	Urban population living in areas where elevation ...
EN.POP.EL5M.RU.ZS	Rural population living in areas where elevation ...
AG.LND.EL5M.UR.ZS	Urban land area where elevation is below 5 meters...
AG.LND.TOTL.K2	Land area (sq. km)
EN.POP.EL5M.ZS	Population living in areas where elevation is bel...
NY.GDP.NGAS.RT.ZS	Natural gas rents (% of GDP)
NY.GDP.FRST.RT.ZS	Forest rents (% of GDP)
NY.GDP.MINR.RT.ZS	Mineral rents (% of GDP)
NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)
NY.GDP.PETR.RT.ZS	Oil rents (% of GDP)
NY.GDP.COAL.RT.ZS	Coal rents (% of GDP)
FX.OWN.TOTL.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.PL.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.OL.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.MA.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.FE.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.SO.ZS	Account ownership at a financial institution or w...
FX.OWN.TOTL.60.ZS	Account ownership at a financial institution or w...
FB.CBK.DPTR.P3	Depositors with commercial banks (per 1,000 adult...
FB.CBK.BRWR.P3	Borrowers from commercial banks (per 1,000 adults...
FB.CBK.BRCH.P5	Commercial bank branches (per 100,000 adults)
FX.OWN.TOTL.40.ZS	Account ownership at a financial institution or w...
FB.ATM.TOTL.P5	Automated teller machines (ATMs) (per 100,000 adu...
SI.RMT.COST.IB.ZS	Average transaction cost of sending remittances t...

SI.RMT.COST.OB.ZS	Average transaction cost of sending remittances f...
FX.OWN.TOTL.YG.ZS	Account ownership at a financial institution or w...
FS.AST.PRVT.GD.ZS	Domestic credit to private sector (% of GDP)
FS.AST.CGOV.GD.ZS	Claims on central government, etc. (% GDP)
FM.AST.PRVT.ZG.M3	Claims on private sector (annual growth as % of b...
FM.AST.PRVT.GD.ZS	Monetary Sector credit to private sector (% GDP)
FM.AST.NFRG.CN	Net foreign assets (current LCU)
FM.AST.DOMS.CN	Net domestic credit (current LCU)
FM.AST.DOMO.ZG.M3	Claims on other sectors of the domestic economy (...
FM.AST.CGOV.ZG.M3	Claims on central government (annual growth as % ...
FD.RES.LIQU.AS.ZS	Bank liquid reserves to bank assets ratio (%)
FD.AST.PRVT.GD.ZS	Domestic credit to private sector by banks (% of ...
FB.BNK.CAPA.ZS	Bank capital to assets ratio (%)
FB.AST.NPER.ZS	Bank nonperforming loans to total gross loans (%)
FS.AST.DOMS.GD.ZS	Domestic credit provided by financial sector (% o...
FS.AST.DOMO.GD.ZS	Claims on other sectors of the domestic economy (...
CM.MKT.LCAP.CD	Market capitalization of listed domestic companie...
CM.MKT.TRAD.GD.ZS	Stocks traded, total value (% of GDP)
CM.MKT.TRNR	Stocks traded, turnover ratio of domestic shares ...
CM.MKT.LCAP.GD.ZS	Market capitalization of listed domestic companie...
CM.MKT.INDX.ZG	S&P Global Equity Indices (annual % change)
CM.MKT.TRAD.CD	Stocks traded, total value (current US$)
CM.MKT.LDOM.NO	Listed domestic companies, total
NY.GDP.DEFL.KD.ZG	Inflation, GDP deflator (annual %)
NY.GDP.DEFL.KD.ZG.AD	Inflation, GDP deflator: linked series (annual %)
NY.GDP.DEFL.ZS	GDP deflator (base year varies by country)
NY.GDP.DEFL.ZS.AD	GDP deflator: linked series (base year varies by ...
FP.WPI.TOTL	Wholesale price index (2010 = 100)
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
FP.CPI.TOTL	Consumer price index (2010 = 100)
PA.NUS.ATLS	DEC alternative conversion factor (LCU per US$)
PA.NUS.FCRF	Official exchange rate (LCU per US$, period avera...
PX.REX.REER	Real effective exchange rate index (2010 = 100)
FR.INR.RISK	Risk premium on lending (lending rate minus treas...
FR.INR.LEND	Lending interest rate (%)
FR.INR.DPST	Deposit interest rate (%)

FR.INR.RINR	Real interest rate (%)
FR.INR.LNDP	Interest rate spread (lending rate minus deposit ...
FM.LBL.BMNY.CN	Broad money (current LCU)
FM.LBL.BMNY.GD.ZS	Broad money (% of GDP)
FM.LBL.BMNY.IR.ZS	Broad money to total reserves ratio
FM.LBL.BMNY.ZG	Broad money growth (annual %)
SP.M15.2024.FE.ZS	Women who were first married by age 15 (% of wome...
SP.M18.2024.FE.ZS	Women who were first married by age 18 (% of wome...
SG.VAW.GOES.ZS	Women who believe a husband is justified in beati...
SG.VAW.REFU.ZS	Women who believe a husband is justified in beati...
SG.VAW.REAS.ZS	Women who believe a husband is justified in beati...
SG.VAW.NEGL.ZS	Women who believe a husband is justified in beati...
SG.VAW.BURN.ZS	Women who believe a husband is justified in beati...
SG.VAW.ARGU.ZS	Women who believe a husband is justified in beati...
SG.VAW.1549.ZS	Proportion of women subjected to physical and/or ...
SG.TIM.UWRK.FE	Proportion of time spent on unpaid domestic and c...
SG.TIM.UWRK.MA	Proportion of time spent on unpaid domestic and c...
SG.DMK.ALLD.FN.ZS	Women participating in the three decisions (own h...
SG.DMK.SRCR.FN.ZS	Women making their own informed decisions regardi...
SG.GEN.PARL.ZS	Proportion of seats held by women in national par...
SN.ITK.MSFI.ZS	Prevalence of moderate or severe food insecurity ...
SN.ITK.SVFI.ZS	Prevalence of severe food insecurity in the popul...
SH.STA.BASS.ZS	People using at least basic sanitation services (...
SH.STA.SMSS.ZS	People using safely managed sanitation services (...
SH.TBS.CURE.ZS	Tuberculosis treatment success rate (% of new cas...
SH.TBS.DTEC.ZS	Tuberculosis case detection rate (%, all forms)
SH.STA.ORTH	Diarrhea treatment (% of children under 5 who rec...
SH.STA.ORCF.ZS	Diarrhea treatment (% of children under 5 receivi...
SH.IMM.MEAS	Immunization, measles (% of children ages 12-23 m...
SH.MLR.NETS.ZS	Use of insecticide-treated bed nets (% of under-5...
SH.MLR.TRET.ZS	Children with fever receiving antimalarial drugs ...
SH.STA.BASS.UR.ZS	People using at least basic sanitation services, ...
SH.CON.1524.MA.ZS	Condom use, population ages 15-24, male (% of mal...
SH.CON.1524.FE.ZS	Condom use, population ages 15-24, female (% of f...
SH.IMM.IDPT	Immunization, DPT (% of children ages 12-23 month...
SH.STA.SMSS.UR.ZS	People using safely managed sanitation services, ...

SH.H2O.SMDW.RU.ZS	People using safely managed drinking water servic...
SH.STA.SMSS.RU.ZS	People using safely managed sanitation services, ...
SH.H2O.SMDW.ZS	People using safely managed drinking water servic...
SH.STA.HYGN.ZS	People with basic handwashing facilities includin...
SH.H2O.BASW.ZS	People using at least basic drinking water servic...
SH.H2O.BASW.UR.ZS	People using at least basic drinking water servic...
SH.H2O.SMDW.UR.ZS	People using safely managed drinking water servic...
SH.STA.HYGN.UR.ZS	People with basic handwashing facilities includin...
SH.H2O.BASW.RU.ZS	People using at least basic drinking water servic...
SH.STA.BASS.RU.ZS	People using at least basic sanitation services, ...
SH.STA.ARIC.ZS	ARI treatment (% of children under 5 taken to a h...
SH.IMM.HEPB	Immunization, HepB3 (% of one-year-old children)
SH.STA.HYGN.RU.ZS	People with basic handwashing facilities includin...
SH.XPD.CHEX.GD.ZS	Current health expenditure (% of GDP)
SH.XPD.CHEX.PC.CD	Current health expenditure per capita (current US...
SH.MED.BEDS.ZS	Hospital beds (per 1,000 people)
SH.XPD.OOPC.PP.CD	Out-of-pocket expenditure per capita, PPP (curren...
SH.XPD.CHEX.PP.CD	Current health expenditure per capita, PPP (curre...
SH.XPD.PVTD.PC.CD	Domestic private health expenditure per capita (c...
SH.XPD.EHEX.PC.CD	External health expenditure per capita (current U...
SH.XPD.EHEX.PP.CD	External health expenditure per capita, PPP (curr...
SH.XPD.GHED.CH.ZS	Domestic general government health expenditure (%...
SH.XPD.GHED.GD.ZS	Domestic general government health expenditure (%...
SH.XPD.GHED.GE.ZS	Domestic general government health expenditure (%...
SH.XPD.GHED.PC.CD	Domestic general government health expenditure pe...
SH.XPD.GHED.PP.CD	Domestic general government health expenditure pe...
SH.XPD.OOPC.PC.CD	Out-of-pocket expenditure per capita (current US$...
SH.XPD.PVTD.CH.ZS	Domestic private health expenditure (% of current...
SH.XPD.PVTD.PP.CD	Domestic private health expenditure per capita, P...
SH.XPD.EHEX.CH.ZS	External health expenditure (% of current health ...
SH.XPD.OOPC.CH.ZS	Out-of-pocket expenditure (% of current health ex...
SH.MED.CMHW.P3	Community health workers (per 1,000 people)
SH.SGR.PROC.P5	Number of surgical procedures (per 100,000 popula...
SH.MED.NUMW.P3	Nurses and midwives (per 1,000 people)
SH.MED.PHYS.ZS	Physicians (per 1,000 people)
SH.MED.SAOP.P5	Specialist surgical workforce (per 100,000 popula...

SH.DYN.1519	Probability of dying among adolescents ages 15-19...
SH.STA.POIS.P5.MA	Mortality rate attributed to unintentional poison...
SH.DYN.MORT	Mortality rate, under-5 (per 1,000 live births)
SH.DYN.MORT.FE	Mortality rate, under-5, female (per 1,000 live b...
SH.DYN.MORT.MA	Mortality rate, under-5, male (per 1,000 live bir...
SH.DYN.NCOM.FE.ZS	Mortality from CVD, cancer, diabetes or CRD betwe...
SH.DYN.NCOM.MA.ZS	Mortality from CVD, cancer, diabetes or CRD betwe...
SH.DYN.NCOM.ZS	Mortality from CVD, cancer, diabetes or CRD betwe...
SH.DYN.NMRT	Mortality rate, neonatal (per 1,000 live births)
SH.STA.AIRP.P5	Mortality rate attributed to household and ambien...
SH.STA.AIRP.MA.P5	Mortality rate attributed to household and ambien...
SH.STA.POIS.P5.FE	Mortality rate attributed to unintentional poison...
SH.DYN.1014	Probability of dying among adolescents ages 10-14...
SP.DYN.TO65.MA.ZS	Survival to age 65, male (% of cohort)
SP.DYN.TO65.FE.ZS	Survival to age 65, female (% of cohort)
SP.DYN.LE00.MA.IN	Life expectancy at birth, male (years)
SP.DYN.LE00.IN	Life expectancy at birth, total (years)
SP.DYN.LE00.FE.IN	Life expectancy at birth, female (years)
SP.DYN.IMRT.MA.IN	Mortality rate, infant, male (per 1,000 live birt...
SP.DYN.IMRT.IN	Mortality rate, infant (per 1,000 live births)
SP.DYN.IMRT.FE.IN	Mortality rate, infant, female (per 1,000 live bi...
SP.DYN.AMRT.MA	Mortality rate, adult, male (per 1,000 male adult...
SP.DYN.AMRT.FE	Mortality rate, adult, female (per 1,000 female a...
SH.STA.AIRP.FE.P5	Mortality rate attributed to household and ambien...
SH.DYN.0509	Probability of dying among children ages 5-9 year...
SH.DYN.2024	Probability of dying among youth ages 20-24 years...
SH.DTH.MORT	Number of under-five deaths
SH.DTH.IMRT	Number of infant deaths
SH.DTH.2024	Number of deaths ages 20-24 years
SH.STA.POIS.P5	Mortality rate attributed to unintentional poison...
SH.DTH.NMRT	Number of neonatal deaths
SH.DTH.1519	Number of deaths ages 15-19 years
SH.DTH.1014	Number of deaths ages 10-14 years
SH.STA.WASH.P5	Mortality rate attributed to unsafe water, unsafe...
SH.STA.TRAF.P5	Mortality caused by road traffic injury (per 100,...
SH.STA.SUIC.FE.P5	Suicide mortality rate, female (per 100,000 femal...

SH.DTH.0509	Number of deaths ages 5-9 years
SH.STA.SUIC.MA.P5	Suicide mortality rate, male (per 100,000 male po...
SH.STA.SUIC.P5	Suicide mortality rate (per 100,000 population)
SH.STA.MALN.FE.ZS	Prevalence of underweight, weight for age, female...
SH.STA.WAST.FE.ZS	Prevalence of wasting, weight for height, female ...
SH.STA.STNT.ZS	Prevalence of stunting, height for age (% of chil...
SH.STA.STNT.ME.ZS	Prevalence of stunting, height for age (modeled e...
SN.ITK.VITA.ZS	Vitamin A supplementation coverage rate (% of chi...
SH.STA.STNT.MA.ZS	Prevalence of stunting, height for age, male (% o...
SN.ITK.SALT.ZS	Consumption of iodized salt (% of households)
SH.STA.WAST.MA.ZS	Prevalence of wasting, weight for height, male (%...
SH.STA.WAST.ZS	Prevalence of wasting, weight for height (% of ch...
SH.SVR.WAST.ZS	Prevalence of severe wasting, weight for height (...
SH.SVR.WAST.MA.ZS	Prevalence of severe wasting, weight for height, ...
SH.PRG.ANEM	Prevalence of anemia among pregnant women (%)
SH.STA.OWGH.ZS	Prevalence of overweight, weight for height (% of...
SH.STA.OWGH.ME.ZS	Prevalence of overweight (modeled estimate, % of ...
SH.STA.BFED.ZS	Exclusive breastfeeding (% of children under 6 mo...
SH.STA.BRTW.ZS	Low-birthweight babies (% of births)
SH.STA.OWGH.MA.ZS	Prevalence of overweight, weight for height, male...
SH.STA.OWGH.FE.ZS	Prevalence of overweight, weight for height, fema...
SH.STA.MALN.ZS	Prevalence of underweight, weight for age (% of c...
SH.STA.MALN.MA.ZS	Prevalence of underweight, weight for age, male (...
SH.ANM.ALLW.ZS	Prevalence of anemia among women of reproductive ...
SH.ANM.CHLD.ZS	Prevalence of anemia among children (% of childre...
SH.ANM.NPRG.ZS	Prevalence of anemia among non-pregnant women (% ...
SH.SVR.WAST.FE.ZS	Prevalence of severe wasting, weight for height, ...
SH.STA.STNT.FE.ZS	Prevalence of stunting, height for age, female (%...
SN.ITK.DEFC.ZS	Prevalence of undernourishment (% of population)
SP.REG.BRTH.UR.ZS	Completeness of birth registration, urban (%)
SP.DYN.CDRT.IN	Death rate, crude (per 1,000 people)
SP.REG.BRTH.RU.ZS	Completeness of birth registration, rural (%)
SP.REG.BRTH.MA.ZS	Completeness of birth registration, male (%)
SP.REG.BRTH.FE.ZS	Completeness of birth registration, female (%)
SP.HOU.FEMA.ZS	Female headed households (% of households with a ...
SP.DYN.CBRT.IN	Birth rate, crude (per 1,000 people)

SP.POP.GROW	Population growth (annual %)
SP.POP.DPND.YG	Age dependency ratio, young (% of working-age pop...
SP.POP.DPND.OL	Age dependency ratio, old (% of working-age popul...
SP.POP.DPND	Age dependency ratio (% of working-age population...
SP.REG.BRTH.ZS	Completeness of birth registration (%)
SP.REG.DTHS.ZS	Completeness of death registration with cause-of-...
SP.POP.1519.FE.5Y	Population ages 15-19, female (% of female popula...
SP.POP.4044.MA.5Y	Population ages 40-44, male (% of male population...
SP.POP.4044.FE.5Y	Population ages 40-44, female (% of female popula...
SP.POP.2529.FE.5Y	Population ages 25-29, female (% of female popula...
SP.POP.2024.MA.5Y	Population ages 20-24, male (% of male population...
SP.POP.3034.FE.5Y	Population ages 30-34, female (% of female popula...
SP.POP.3034.MA.5Y	Population ages 30-34, male (% of male population...
SP.POP.2024.FE.5Y	Population ages 20-24, female (% of female popula...
SP.POP.3539.FE.5Y	Population ages 35-39, female (% of female popula...
SP.POP.0004.FE.5Y	Population ages 00-04, female (% of female popula...
SP.POP.1564.TO.ZS	Population ages 15-64 (% of total population)
SP.POP.0004.MA.5Y	Population ages 00-04, male (% of male population...
SP.POP.1564.TO	Population ages 15-64, total
SP.POP.1564.MA.ZS	Population ages 15-64, male (% of male population...
SP.POP.1564.MA.IN	Population ages 15-64, male
SP.POP.1564.FE.ZS	Population ages 15-64, female (% of female popula...
SP.POP.0014.FE.IN	Population ages 0-14, female
SP.POP.0014.FE.ZS	Population ages 0-14, female (% of female populat...
SP.POP.0014.MA.IN	Population ages 0-14, male
SP.POP.1564.FE.IN	Population ages 15-64, female
SP.POP.0014.TO	Population ages 0-14, total
SP.POP.0014.TO.ZS	Population ages 0-14 (% of total population)
SP.POP.0509.FE.5Y	Population ages 05-09, female (% of female popula...
SP.POP.3539.MA.5Y	Population ages 35-39, male (% of male population...
SP.POP.1014.FE.5Y	Population ages 10-14, female (% of female popula...
SP.POP.1014.MA.5Y	Population ages 10-14, male (% of male population...
SP.POP.1519.MA.5Y	Population ages 15-19, male (% of male population...
SP.POP.0014.MA.ZS	Population ages 0-14, male (% of male population)
SP.POP.0509.MA.5Y	Population ages 05-09, male (% of male population...
SP.POP.65UP.TO	Population ages 65 and above, total

SP.POP.5559.FE.5Y	Population ages 55-59, female (% of female popula...
SP.POP.5559.MA.5Y	Population ages 55-59, male (% of male population...
SP.POP.6064.FE.5Y	Population ages 60-64, female (% of female popula...
SP.POP.6064.MA.5Y	Population ages 60-64, male (% of male population...
SP.POP.6569.FE.5Y	Population ages 65-69, female (% of female popula...
SP.POP.6569.MA.5Y	Population ages 65-69, male (% of male population...
SP.POP.65UP.FE.IN	Population ages 65 and above, female
SP.POP.65UP.FE.ZS	Population ages 65 and above, female (% of female...
SP.POP.65UP.MA.IN	Population ages 65 and above, male
SP.POP.65UP.MA.ZS	Population ages 65 and above, male (% of male pop...
SP.POP.4549.MA.5Y	Population ages 45-49, male (% of male population...
SP.POP.65UP.TO.ZS	Population ages 65 and above (% of total populati...
SP.POP.7074.FE.5Y	Population ages 70-74, female (% of female popula...
SP.POP.7074.MA.5Y	Population ages 70-74, male (% of male population...
SP.POP.7579.FE.5Y	Population ages 75-79, female (% of female popula...
SP.POP.7579.MA.5Y	Population ages 75-79, male (% of male population...
SP.POP.80UP.FE.5Y	Population ages 80 and above, female (% of female...
SP.POP.80UP.MA.5Y	Population ages 80 and above, male (% of male pop...
SP.POP.BRTH.MF	Sex ratio at birth (male births per female births...
SP.POP.2529.MA.5Y	Population ages 25-29, male (% of male population...
SP.POP.TOTL	Population, total
SP.POP.TOTL.FE.IN	Population, female
SP.POP.TOTL.FE.ZS	Population, female (% of total population)
SP.POP.TOTL.MA.IN	Population, male
SP.POP.TOTL.MA.ZS	Population, male (% of total population)
SP.POP.4549.FE.5Y	Population ages 45-49, female (% of female popula...
SP.POP.5054.MA.5Y	Population ages 50-54, male (% of male population...
SP.POP.5054.FE.5Y	Population ages 50-54, female (% of female popula...
SH.STA.ANVC.ZS	Pregnant women receiving prenatal care (%)
SH.MMR.DTHS	Number of maternal deaths
SH.MMR.RISK.ZS	Lifetime risk of maternal death (%)
SH.MMR.RISK	Lifetime risk of maternal death (1 in: rate varie...
SH.STA.BRTC.ZS	Births attended by skilled health staff (% of tot...
SH.STA.MMRT	Maternal mortality ratio (modeled estimate, per 1...
SH.FPL.SATM.ZS	Demand for family planning satisfied by modern me...
SP.DYN.TFRT.IN	Fertility rate, total (births per woman)

SH.STA.MMRT.NE	Maternal mortality ratio (national estimate, per ...
SP.UWT.TFRT	Unmet need for contraception (% of married women ...
SP.DYN.CONM.ZS	Contraceptive prevalence, any modern method (% of...
SP.DYN.CONU.ZS	Contraceptive prevalence, any method (% of marrie...
SH.VAC.TTNS.ZS	Newborns protected against tetanus (%)
SP.ADO.TFRT	Adolescent fertility rate (births per 1,000 women...
SP.MTR.1519.ZS	Teenage mothers (% of women ages 15-19 who have h...
SP.DYN.WFRT	Wanted fertility rate (births per woman)
SH.HIV.INCD	Adults (ages 15-49) newly infected with HIV
SH.HIV.ARTC.ZS	Antiretroviral therapy coverage (% of people livi...
SH.HIV.1524.MA.ZS	Prevalence of HIV, male (% ages 15-24)
SH.HIV.1524.FE.ZS	Prevalence of HIV, female (% ages 15-24)
SH.STA.ODFC.RU.ZS	People practicing open defecation, rural (% of ru...
SH.HIV.INCD.14	Children (ages 0-14) newly infected with HIV
SH.DYN.AIDS.ZS	Prevalence of HIV, total (% of population ages 15...
SH.DYN.AIDS.FE.ZS	Women's share of population ages 15+ living with ...
SH.DTH.NCOM.ZS	Cause of death, by non-communicable diseases (% o...
SH.DTH.INJR.ZS	Cause of death, by injury (% of total)
SH.DTH.COMM.ZS	Cause of death, by communicable diseases and mate...
SH.ALC.PCAP.MA.LI	Total alcohol consumption per capita, male (liter...
SH.HIV.0014	Children (0-14) living with HIV
SH.HIV.INCD.TL	Adults (ages 15+) and children (ages 0-14) newly ...
SH.MLR.INCD.P3	Incidence of malaria (per 1,000 population at ris...
SH.HIV.INCD.YG	Young people (ages 15-24) newly infected with HIV
SH.HIV.INCD.YG.P3	Incidence of HIV, ages 15-24 (per 1,000 uninfecte...
SH.HIV.PMTC.ZS	Antiretroviral therapy coverage for PMTCT (% of p...
SH.PRV.SMOK	Prevalence of current tobacco use (% of adults)
SH.PRV.SMOK.FE	Prevalence of current tobacco use, females (% of ...
SH.PRV.SMOK.MA	Prevalence of current tobacco use, males (% of ma...
SH.SGR.CRSK.ZS	Risk of catastrophic expenditure for surgical car...
SH.ALC.PCAP.LI	Total alcohol consumption per capita (liters of p...
SH.TBS.INCD	Incidence of tuberculosis (per 100,000 people)
SH.SGR.IRSK.ZS	Risk of impoverishing expenditure for surgical ca...
SH.STA.ODFC.ZS	People practicing open defecation (% of populatio...
SH.STA.ODFC.UR.ZS	People practicing open defecation, urban (% of ur...
SH.STA.DIAB.ZS	Diabetes prevalence (% of population ages 20 to 7...

SH.STA.FGMS.ZS	Female genital mutilation prevalence (%)
SH.HIV.INCD.TL.P3	Incidence of HIV, all (per 1,000 uninfected popul...
SH.ALC.PCAP.FE.LI	Total alcohol consumption per capita, female (lit...
SH.HIV.INCD.ZS	Incidence of HIV, ages 15-49 (per 1,000 uninfecte...
SH.UHC.FBP2.ZS	Proportion of population pushed further below the...
SH.UHC.NOP1.ZS	Proportion of population pushed below the $2.15 (...
SH.UHC.NOP2.ZS	Proportion of population pushed below the $3.65 (...
SH.UHC.NOPR.ZS	Proportion of population pushed below the 60% med...
SH.UHC.OOPC.10.ZS	Proportion of population spending more than 10% o...
SH.UHC.OOPC.25.ZS	Proportion of population spending more than 25% o...
SH.UHC.SRVS.CV.XD	UHC service coverage index
SH.UHC.TOT1.ZS	Proportion of population pushed or further pushed...
SH.UHC.TOT2.ZS	Proportion of population pushed or further pushed...
SH.UHC.TOTR.ZS	Proportion of population pushed or further pushed...
SH.UHC.FBP1.ZS	Proportion of population pushed further below the...
SH.UHC.FBPR.ZS	Proportion of population pushed further below the...
IT.NET.USER.ZS	Individuals using the Internet (% of population)
IT.NET.SECR.P6	Secure Internet servers (per 1 million people)
IT.NET.SECR	Secure Internet servers
IT.NET.BBND.P2	Fixed broadband subscriptions (per 100 people)
IT.NET.BBND	Fixed broadband subscriptions
IT.MLT.MAIN.P2	Fixed telephone subscriptions (per 100 people)
BX.GSR.CCIS.CD	ICT service exports (BoP, current US$)
IT.CEL.SETS	Mobile cellular subscriptions
IT.MLT.MAIN	Fixed telephone subscriptions
TX.VAL.ICTG.ZS.UN	ICT goods exports (% of total goods exports)
BX.GSR.CCIS.ZS	ICT service exports (% of service exports, BoP)
TM.VAL.ICTG.ZS.UN	ICT goods imports (% total goods imports)
IT.CEL.SETS.P2	Mobile cellular subscriptions (per 100 people)
SP.POP.SCIE.RD.P6	Researchers in R&D (per million people)
SP.POP.TECH.RD.P6	Technicians in R&D (per million people)
GB.XPD.RSDV.GD.ZS	Research and development expenditure (% of GDP)
TX.VAL.TECH.MF.ZS	High-technology exports (% of manufactured export...
TX.VAL.TECH.CD	High-technology exports (current US$)
IP.IDS.NRCT	Industrial design applications, nonresident, by c...
IP.JRN.ARTC.SC	Scientific and technical journal articles

IP.PAT.NRES	Patent applications, nonresidents
IP.PAT.RESD	Patent applications, residents
IP.TMK.NRCT	Trademark applications, nonresident, by count
IP.TMK.RSCT	Trademark applications, resident, by count
TX.MNF.TECH.ZS.UN	Medium and high-tech exports (% manufactured expo...
IP.IDS.RSCT	Industrial design applications, resident, by coun...
EP.PMP.DESL.CD	Pump price for diesel fuel (US$ per liter)
EP.PMP.SGAS.CD	Pump price for gasoline (US$ per liter)
IS.SHP.GOOD.TU	Container port traffic (TEU: 20 foot equivalent u...
IS.AIR.GOOD.MT.K1	Air transport, freight (million ton-km)
IS.SHP.GCNW.XQ	Liner shipping connectivity index (maximum value ...
IS.RRS.TOTL.KM	Rail lines (total route-km)
IS.RRS.PASG.KM	Railways, passengers carried (million passenger-k...
IS.RRS.GOOD.MT.K6	Railways, goods transported (million ton-km)
IS.AIR.PSGR	Air transport, passengers carried
IS.AIR.DPRT	Air transport, registered carrier departures worl...
SI.POV.GINI	Gini index
SI.DST.03RD.20	Income share held by third 20%
SI.DST.FRST.20	Income share held by lowest 20%
SI.DST.FRST.10	Income share held by lowest 10%
SI.DST.50MD	Proportion of people living below 50 percent of m...
SI.DST.10TH.10	Income share held by highest 10%
SI.DST.05TH.20	Income share held by highest 20%
SI.DST.04TH.20	Income share held by fourth 20%
SI.DST.02ND.20	Income share held by second 20%
SI.POV.UMIC.GP	Poverty gap at $6.85 a day (2017 PPP) (%)
SI.POV.UMIC	Poverty headcount ratio at $6.85 a day (2017 PPP)...
SI.POV.SOPO	Poverty headcount ratio at societal poverty line ...
SI.POV.NAHC	Poverty headcount ratio at national poverty lines...
SI.POV.MPWB	Multidimensional poverty headcount ratio (World B...
SI.POV.MPUN	Multidimensional poverty headcount ratio (UNDP) (...
SI.POV.LMIC.GP	Poverty gap at $3.65 a day (2017 PPP) (%)
SI.POV.LMIC	Poverty headcount ratio at $3.65 a day (2017 PPP)...
SI.POV.GAPS	Poverty gap at $2.15 a day (2017 PPP) (%)
SI.POV.DDAY	Poverty headcount ratio at $2.15 a day (2017 PPP)...
SI.SPR.PC40	Survey mean consumption or income per capita, bot...

SI.SPR.PC40.ZG	Annualized average growth rate in per capita real...
SI.SPR.PCAP	Survey mean consumption or income per capita, tot...
SI.SPR.PCAP.ZG	Annualized average growth rate in per capita real...
IC.WRH.PROC	Procedures to build a warehouse (number)
IC.CRD.PUBL.ZS	Public credit registry coverage (% of adults)
IC.WRH.DURS	Time required to build a warehouse (days)
IC.CUS.DURS.EX	Average time to clear exports through customs (da...
IC.REG.COST.PC.MA.ZS	Cost of business start-up procedures, male (% of ...
IC.FRM.FREG.ZS	Firms formally registered when operations started...
IC.FRM.INFM.ZS	Firms that do not report all sales for tax purpos...
IC.FRM.METG.ZS	Firms visited or required meetings with tax offic...
IC.FRM.OUTG.ZS	Value lost due to electrical outages (% of sales ...
IC.BUS.NDNS.ZS	New business density (new registrations per 1,000...
IC.FRM.RSDV.ZS	Firms that spend on R&D (% of firms)
IC.BUS.EASE.XQ	Ease of doing business rank (1=most business-frie...
IC.FRM.THEV.ZS	Firms experiencing losses due to theft and vandal...
IC.BUS.DISC.XQ	Business extent of disclosure index (0=less discl...
IC.FRM.TRNG.ZS	Firms offering formal training (% of firms)
IC.GOV.DURS.ZS	Time spent dealing with the requirements of gover...
IC.BUS.DFRN.XQ	Ease of doing business score (0 = lowest performa...
IC.ISV.DURS	Time to resolve insolvency (years)
IC.LGL.CRED.XQ	Strength of legal rights index (0=weak to 12=stro...
IC.LGL.DURS	Time required to enforce a contract (days)
IC.PRP.DURS	Time required to register property (days)
IC.PRP.PROC	Procedures to register property (number)
IC.REG.COST.PC.FE.ZS	Cost of business start-up procedures, female (% o...
IC.REG.COST.PC.ZS	Cost of business start-up procedures (% of GNI pe...
IC.REG.DURS	Time required to start a business (days)
IC.REG.DURS.FE	Time required to start a business, female (days)
IC.REG.DURS.MA	Time required to start a business, male (days)
IC.REG.PROC	Start-up procedures to register a business (numbe...
IC.FRM.FEMO.ZS	Firms with female participation in ownership (% o...
IC.FRM.FEMM.ZS	Firms with female top manager (% of firms)
IC.FRM.DURS	Time required to obtain an operating license (day...
IC.FRM.BKWC.ZS	Firms using banks to finance working capital (% o...
IC.ELC.OUTG	Power outages in firms in a typical month (number...

IC.TAX.PAYM	Tax payments (number)
IC.TAX.OTHR.CP.ZS	Other taxes payable by businesses (% of commercia...
IC.TAX.TOTL.CP.ZS	Total tax and contribution rate (% of profit)
IC.ELC.TIME	Time required to get electricity (days)
IC.ELC.DURS	Time to obtain an electrical connection (days)
IC.REG.PROC.FE	Start-up procedures to register a business, femal...
IC.FRM.BNKS.ZS	Firms using banks to finance investment (% of fir...
IC.FRM.BRIB.ZS	Bribery incidence (% of firms experiencing at lea...
IC.ELC.OUTG.ZS	Firms experiencing electrical outages (% of firms...
IC.FRM.CMPU.ZS	Firms competing against unregistered firms (% of ...
IC.CRD.INFO.XQ	Depth of credit information index (0=low to 8=hig...
IC.BUS.NREG	New businesses registered (number)
IC.TAX.METG	Number of visits or required meetings with tax of...
IC.FRM.CORR.ZS	Informal payments to public officials (% of firms...
IC.FRM.CRIM.ZS	Losses due to theft and vandalism (% of annual sa...
IC.TAX.LABR.CP.ZS	Labor tax and contributions (% of commercial prof...
IC.TAX.GIFT.ZS	Firms expected to give gifts in meetings with tax...
IC.TAX.DURS	Time to prepare and pay taxes (hours)
IC.REG.PROC.MA	Start-up procedures to register a business, male ...
IC.CRD.PRVT.ZS	Private credit bureau coverage (% of adults)
IC.TAX.PRFT.CP.ZS	Profit tax (% of commercial profits)
TX.VAL.MRCH.WR.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.TRVL.ZS.WT	Travel services (% of commercial service exports)
TX.VAL.TRAN.ZS.WT	Transport services (% of commercial service expor...
TX.VAL.SERV.CD.WT	Commercial service exports (current US$)
TX.VAL.OTHR.ZS.WT	Computer, communications and other services (% of...
TX.VAL.MRCH.WL.CD	Merchandise exports by the reporting economy (cur...
TX.VAL.MMTL.ZS.UN	Ores and metals exports (% of merchandise exports...
TX.VAL.MRCH.R6.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.MANF.ZS.UN	Manufactures exports (% of merchandise exports)
TX.VAL.INSF.ZS.WT	Insurance and financial services (% of commercial...
TX.VAL.MRCH.HI.ZS	Merchandise exports to high-income economies (% o...
TX.VAL.FUEL.ZS.UN	Fuel exports (% of merchandise exports)
TX.VAL.FOOD.ZS.UN	Food exports (% of merchandise exports)
TX.VAL.AGRI.ZS.UN	Agricultural raw materials exports (% of merchand...
TX.VAL.MRCH.OR.ZS	Merchandise exports to low- and middle-income eco...

TX.VAL.MRCH.R1.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.MRCH.R2.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.MRCH.CD.WT	Merchandise exports (current US$)
TX.VAL.MRCH.R4.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.MRCH.R3.ZS	Merchandise exports to low- and middle-income eco...
TX.VAL.MRCH.AL.ZS	Merchandise exports to economies in the Arab Worl...
TX.VAL.MRCH.RS.ZS	Merchandise exports by the reporting economy, res...
TX.VAL.MRCH.R5.ZS	Merchandise exports to low- and middle-income eco...
TM.VAL.MRCH.R3.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.MMTL.ZS.UN	Ores and metals imports (% of merchandise imports...
TM.VAL.MRCH.R2.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.MANF.ZS.UN	Manufactures imports (% of merchandise imports)
TM.VAL.SERV.CD.WT	Commercial service imports (current US$)
TM.VAL.TRAN.ZS.WT	Transport services (% of commercial service impor...
TM.VAL.MRCH.HI.ZS	Merchandise imports from high-income economies (%...
TM.VAL.FUEL.ZS.UN	Fuel imports (% of merchandise imports)
TM.VAL.MRCH.R1.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.FOOD.ZS.UN	Food imports (% of merchandise imports)
TM.VAL.AGRI.ZS.UN	Agricultural raw materials imports (% of merchand...
TM.VAL.MRCH.OR.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.TRVL.ZS.WT	Travel services (% of commercial service imports)
TM.VAL.INSF.ZS.WT	Insurance and financial services (% of commercial...
TM.VAL.MRCH.AL.ZS	Merchandise imports from economies in the Arab Wo...
TM.VAL.MRCH.CD.WT	Merchandise imports (current US$)
TM.VAL.MRCH.WR.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.OTHR.ZS.WT	Computer, communications and other services (% of...
TM.VAL.MRCH.WL.CD	Merchandise imports by the reporting economy (cur...
TM.VAL.MRCH.RS.ZS	Merchandise imports by the reporting economy, res...
TM.VAL.MRCH.R6.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.MRCH.R5.ZS	Merchandise imports from low- and middle-income e...
TM.VAL.MRCH.R4.ZS	Merchandise imports from low- and middle-income e...
IE.PPN.TRAN.CD	Public private partnerships investment in transpo...
IE.PPN.WATR.CD	Public private partnerships investment in water a...
IE.PPI.TRAN.CD	Investment in transport with private participatio...
IE.PPI.ICTI.CD	Investment in ICT with private participation (cur...
IE.PPI.ENGY.CD	Investment in energy with private participation (...

IE.PPN.ENGY.CD	Public private partnerships investment in energy ...
IE.PPN.ICTI.CD	Public private partnerships investment in ICT (cu...
IE.PPI.WATR.CD	Investment in water and sanitation with private p...
TM.TAX.MRCH.SM.FN.ZS	Tariff rate, most favored nation, simple mean, al...
TM.TAX.MRCH.SR.ZS	Share of tariff lines with specific rates, all pr...
TM.TAX.MRCH.SM.AR.ZS	Tariff rate, applied, simple mean, all products (...
TM.TAX.MRCH.BC.ZS	Binding coverage, all products (%)
TM.TAX.MRCH.BR.ZS	Bound rate, simple mean, all products (%)
TM.TAX.MANF.WM.AR.ZS	Tariff rate, applied, weighted mean, manufactured...
TM.TAX.MANF.SR.ZS	Share of tariff lines with specific rates, manufa...
TM.TAX.MANF.SM.FN.ZS	Tariff rate, most favored nation, simple mean, ma...
TM.TAX.MANF.SM.AR.ZS	Tariff rate, applied, simple mean, manufactured p...
TM.TAX.MANF.IP.ZS	Share of tariff lines with international peaks, m...
TM.TAX.MANF.BR.ZS	Bound rate, simple mean, manufactured products (%...
TM.TAX.MANF.BC.ZS	Binding coverage, manufactured products (%)
TM.TAX.TCOM.IP.ZS	Share of tariff lines with international peaks, p...
TM.TAX.MRCH.IP.ZS	Share of tariff lines with international peaks, a...
TM.TAX.MRCH.WM.AR.ZS	Tariff rate, applied, weighted mean, all products...
TM.TAX.MANF.WM.FN.ZS	Tariff rate, most favored nation, weighted mean, ...
TM.TAX.TCOM.BC.ZS	Binding coverage, primary products (%)
TM.TAX.TCOM.WM.FN.ZS	Tariff rate, most favored nation, weighted mean, ...
TM.TAX.TCOM.WM.AR.ZS	Tariff rate, applied, weighted mean, primary prod...
TM.TAX.TCOM.SR.ZS	Share of tariff lines with specific rates, primar...
TM.TAX.TCOM.SM.FN.ZS	Tariff rate, most favored nation, simple mean, pr...
TM.TAX.TCOM.SM.AR.ZS	Tariff rate, applied, simple mean, primary produc...
TM.TAX.TCOM.BR.ZS	Bound rate, simple mean, primary products (%)
TM.TAX.MRCH.WM.FN.ZS	Tariff rate, most favored nation, weighted mean, ...
TG.VAL.TOTL.GD.ZS	Merchandise trade (% of GDP)
IC.EXP.TMDC	Time to export, documentary compliance (hours)
IC.EXP.TMBC	Time to export, border compliance (hours)
IC.EXP.CSDC.CD	Cost to export, documentary compliance (US$)
LP.LPI.INFR.XQ	Logistics performance index: Quality of trade and...
LP.EXP.DURS.MD	Lead time to export, median case (days)
IC.IMP.TMDC	Time to import, documentary compliance (hours)
LP.IMP.DURS.MD	Lead time to import, median case (days)
LP.LPI.TIME.XQ	Logistics performance index: Frequency with which...

LP.LPI.TRAC.XQ	Logistics performance index: Ability to track and...
IC.EXP.CSBC.CD	Cost to export, border compliance (US$)
IC.IMP.TMBC	Time to import, border compliance (hours)
LP.LPI.ITRN.XQ	Logistics performance index: Ease of arranging co...
LP.LPI.OVRL.XQ	Logistics performance index: Overall (1=low to 5=...
LP.LPI.LOGS.XQ	Logistics performance index: Competence and quali...
IC.IMP.CSDC.CD	Cost to import, documentary compliance (US$)
IC.IMP.CSBC.CD	Cost to import, border compliance (US$)
TX.VAL.MRCH.XD.WD	Export value index (2000 = 100)
TM.VAL.MRCH.XD.WD	Import value index (2000 = 100)
TT.PRI.MRCH.XD.WD	Net barter terms of trade index (2000 = 100)
TX.QTY.MRCH.XD.WD	Export volume index (2000 = 100)
TM.QTY.MRCH.XD.WD	Import volume index (2000 = 100)
TX.UVI.MRCH.XD.WD	Export unit value index (2015 = 100)
TM.UVI.MRCH.XD.WD	Import unit value index (2015 = 100)
ST.INT.XPND.MP.ZS	International tourism, expenditures (% of total i...
ST.INT.TVLR.CD	International tourism, receipts for travel items ...
ST.INT.RCPT.XP.ZS	International tourism, receipts (% of total expor...
ST.INT.DPRT	International tourism, number of departures
ST.INT.ARVL	International tourism, number of arrivals
ST.INT.TRNR.CD	International tourism, receipts for passenger tra...
ST.INT.XPND.CD	International tourism, expenditures (current US$)
ST.INT.TVLX.CD	International tourism, expenditures for travel it...
ST.INT.RCPT.CD	International tourism, receipts (current US$)
ST.INT.TRNX.CD	International tourism, expenditures for passenger...
VC.BTL.DETH	Battle-related deaths (number of people)
VC.IDP.NWCV	Internally displaced persons, new displacement as...
VC.IDP.NWDS	Internally displaced persons, new displacement as...
VC.IDP.TOCV	Internally displaced persons, total displaced by ...
VC.IHR.PSRC.FE.P5	Intentional homicides, female (per 100,000 female...
VC.IHR.PSRC.P5	Intentional homicides (per 100,000 people)
VC.IHR.PSRC.MA.P5	Intentional homicides, male (per 100,000 male)
MS.MIL.XPRT.KD	Arms exports (SIPRI trend indicator values)
MS.MIL.XPND.ZS	Military expenditure (% of general government exp...
MS.MIL.XPND.GD.ZS	Military expenditure (% of GDP)
MS.MIL.XPND.CN	Military expenditure (current LCU)

MS.MIL.XPND.CD	Military expenditure (current USD)
MS.MIL.TOTL.TF.ZS	Armed forces personnel (% of total labor force)
MS.MIL.TOTL.P1	Armed forces personnel, total
MS.MIL.MPRT.KD	Arms imports (SIPRI trend indicator values)
GF.XPD.BUDG.ZS	Primary government expenditures as a proportion o...
GC.AST.TOTL.GD.ZS	Net acquisition of financial assets (% of GDP)
GC.DOD.TOTL.CN	Central government debt, total (current LCU)
GC.DOD.TOTL.GD.ZS	Central government debt, total (% of GDP)
GC.LBL.TOTL.CN	Net incurrence of liabilities, total (current LCU...
GC.LBL.TOTL.GD.ZS	Net incurrence of liabilities, total (% of GDP)
GC.NFN.TOTL.GD.ZS	Net investment in nonfinancial assets (% of GDP)
GC.NFN.TOTL.CN	Net investment in nonfinancial assets (current LC...
GC.NLD.TOTL.GD.ZS	Net lending (+) / net borrowing (-) (% of GDP)
GC.AST.TOTL.CN	Net acquisition of financial assets (current LCU)
GC.NLD.TOTL.CN	Net lending (+) / net borrowing (-) (current LCU)
GC.XPN.TRFT.ZS	Subsidies and other transfers (% of expense)
GC.XPN.TRFT.CN	Subsidies and other transfers (current LCU)
GC.XPN.TOTL.GD.ZS	Expense (% of GDP)
GC.XPN.TOTL.CN	Expense (current LCU)
GC.XPN.OTHR.CN	Other expense (current LCU)
GC.XPN.INTP.ZS	Interest payments (% of expense)
GC.XPN.INTP.RV.ZS	Interest payments (% of revenue)
GC.XPN.INTP.CN	Interest payments (current LCU)
GC.XPN.GSRV.ZS	Goods and services expense (% of expense)
GC.XPN.GSRV.CN	Goods and services expense (current LCU)
GC.XPN.COMP.ZS	Compensation of employees (% of expense)
GC.XPN.COMP.CN	Compensation of employees (current LCU)
GC.XPN.OTHR.ZS	Other expense (% of expense)
GC.TAX.IMPT.CN	Customs and other import duties (current LCU)
GC.TAX.YPKG.RV.ZS	Taxes on income, profits and capital gains (% of ...
GC.TAX.YPKG.ZS	Taxes on income, profits and capital gains (% of ...
GC.TAX.TOTL.CN	Tax revenue (current LCU)
GC.TAX.OTHR.RV.ZS	Other taxes (% of revenue)
GC.TAX.OTHR.CN	Other taxes (current LCU)
GC.TAX.INTT.RV.ZS	Taxes on international trade (% of revenue)
GC.TAX.INTT.CN	Taxes on international trade (current LCU)

GC.TAX.IMPT.ZS	Customs and other import duties (% of tax revenue...
GC.TAX.GSRV.VA.ZS	Taxes on goods and services (% value added of ind...
GC.TAX.GSRV.RV.ZS	Taxes on goods and services (% of revenue)
GC.TAX.GSRV.CN	Taxes on goods and services (current LCU)
GC.TAX.EXPT.ZS	Taxes on exports (% of tax revenue)
GC.TAX.EXPT.CN	Taxes on exports (current LCU)
GC.TAX.TOTL.GD.ZS	Tax revenue (% of GDP)
GC.REV.XGRT.CN	Revenue, excluding grants (current LCU)
GC.REV.SOCL.ZS	Social contributions (% of revenue)
GC.REV.SOCL.CN	Social contributions (current LCU)
GC.REV.GOTR.ZS	Grants and other revenue (% of revenue)
GC.REV.GOTR.CN	Grants and other revenue (current LCU)
GC.REV.XGRT.GD.ZS	Revenue, excluding grants (% of GDP)
GC.TAX.YPKG.CN	Taxes on income, profits and capital gains (curre...
RQ.STD.ERR	Regulatory Quality: Standard Error
CC.STD.ERR	Control of Corruption: Standard Error
CC.PER.RNK.UPPER	Control of Corruption: Percentile Rank, Upper Bou...
CC.PER.RNK	Control of Corruption: Percentile Rank
RQ.PER.RNK.UPPER	Regulatory Quality: Percentile Rank, Upper Bound ...
CC.NO.SRC	Control of Corruption: Number of Sources
PV.PER.RNK.LOWER	Political Stability and Absence of Violence/Terro...
PV.PER.RNK	Political Stability and Absence of Violence/Terro...
CC.PER.RNK.LOWER	Control of Corruption: Percentile Rank, Lower Bou...
RQ.PER.RNK.LOWER	Regulatory Quality: Percentile Rank, Lower Bound ...
RQ.NO.SRC	Regulatory Quality: Number of Sources
RQ.PER.RNK	Regulatory Quality: Percentile Rank
RQ.EST	Regulatory Quality: Estimate
RL.STD.ERR	Rule of Law: Standard Error
PV.STD.ERR	Political Stability and Absence of Violence/Terro...
RL.EST	Rule of Law: Estimate
RL.NO.SRC	Rule of Law: Number of Sources
RL.PER.RNK	Rule of Law: Percentile Rank
PV.EST	Political Stability and Absence of Violence/Terro...
RL.PER.RNK.LOWER	Rule of Law: Percentile Rank, Lower Bound of 90% ...
PV.PER.RNK.UPPER	Political Stability and Absence of Violence/Terro...
PV.NO.SRC	Political Stability and Absence of Violence/Terro...

RL.PER.RNK.UPPER	Rule of Law: Percentile Rank, Upper Bound of 90% ...
GE.EST	Government Effectiveness: Estimate
IQ.CPA.PROP.XQ	CPIA property rights and rule-based governance ra...
IQ.CPA.PROT.XQ	CPIA social protection rating (1=low to 6=high)
IQ.CPA.PUBS.XQ	CPIA public sector management and institutions cl...
IQ.CPA.REVN.XQ	CPIA efficiency of revenue mobilization rating (1...
IQ.CPA.SOCI.XQ	CPIA policies for social inclusion/equity cluster...
IQ.CPA.STRC.XQ	CPIA structural policies cluster average (1=low t...
IQ.SPI.PIL5	Statistical performance indicators (SPI): Pillar ...
IQ.SPI.PIL4	Statistical performance indicators (SPI): Pillar ...
IQ.SPI.PIL3	Statistical performance indicators (SPI): Pillar ...
IQ.SPI.PIL2	Statistical performance indicators (SPI): Pillar ...
IQ.CPA.PRES.XQ	CPIA equity of public resource use rating (1=low ...
IQ.SPI.PIL1	Statistical performance indicators (SPI): Pillar ...
VA.NO.SRC	Voice and Accountability: Number of Sources
VA.PER.RNK	Voice and Accountability: Percentile Rank
VA.PER.RNK.LOWER	Voice and Accountability: Percentile Rank, Lower ...
VA.PER.RNK.UPPER	Voice and Accountability: Percentile Rank, Upper ...
VA.STD.ERR	Voice and Accountability: Standard Error
IQ.SPI.OVRL	Statistical performance indicators (SPI): Overall...
IQ.SCI.SRCE	Source data assessment of statistical capacity (s...
IQ.SCI.PRDC	Periodicity and timeliness assessment of statisti...
IQ.SCI.OVRL	Statistical Capacity Score (Overall Average) (sca...
IQ.SCI.MTHD	Methodology assessment of statistical capacity (s...
VA.EST	Voice and Accountability: Estimate
IQ.CPA.PADM.XQ	CPIA quality of public administration rating (1=l...
IQ.CPA.MACR.XQ	CPIA macroeconomic management rating (1=low to 6=...
IQ.CPA.IRAI.XQ	IDA resource allocation index (1=low to 6=high)
GE.NO.SRC	Government Effectiveness: Number of Sources
GE.PER.RNK	Government Effectiveness: Percentile Rank
GE.PER.RNK.LOWER	Government Effectiveness: Percentile Rank, Lower ...
GE.PER.RNK.UPPER	Government Effectiveness: Percentile Rank, Upper ...
GE.STD.ERR	Government Effectiveness: Standard Error
HD.HCI.OVRL	Human capital index (HCI) (scale 0-1)
HD.HCI.OVRL.FE	Human capital index (HCI), female (scale 0-1)
HD.HCI.OVRL.LB	Human capital index (HCI), lower bound (scale 0-1...

HD.HCI.OVRL.LB.FE	Human capital index (HCI), female, lower bound (s...
HD.HCI.OVRL.LB.MA	Human capital index (HCI), male, lower bound (sca...
HD.HCI.OVRL.MA	Human capital index (HCI), male (scale 0-1)
HD.HCI.OVRL.UB	Human capital index (HCI), upper bound (scale 0-1...
HD.HCI.OVRL.UB.FE	Human capital index (HCI), female, upper bound (s...
HD.HCI.OVRL.UB.MA	Human capital index (HCI), male, upper bound (sca...
IQ.CPA.BREG.XQ	CPIA business regulatory environment rating (1=lo...
IQ.CPA.DEBT.XQ	CPIA debt policy rating (1=low to 6=high)
IQ.CPA.ECON.XQ	CPIA economic management cluster average (1=low t...
IQ.CPA.ENVR.XQ	CPIA policy and institutions for environmental su...
IQ.CPA.FINQ.XQ	CPIA quality of budgetary and financial managemen...
IQ.CPA.FINS.XQ	CPIA financial sector rating (1=low to 6=high)
IQ.CPA.FISP.XQ	CPIA fiscal policy rating (1=low to 6=high)
IQ.CPA.GNDR.XQ	CPIA gender equality rating (1=low to 6=high)
IQ.CPA.HRES.XQ	CPIA building human resources rating (1=low to 6=...
CC.EST	Control of Corruption: Estimate
IQ.CPA.TRAD.XQ	CPIA trade rating (1=low to 6=high)
IQ.CPA.TRAN.XQ	CPIA transparency, accountability, and corruption...
SL.AGR.0714.MA.ZS	Child employment in agriculture, male (% of male ...
SL.SLF.0714.ZS	Children in employment, self-employed (% of child...
SL.SLF.0714.MA.ZS	Children in employment, self-employed, male (% of...
SL.SLF.0714.FE.ZS	Children in employment, self-employed, female (% ...
SL.MNF.0714.ZS	Child employment in manufacturing (% of economica...
SL.MNF.0714.MA.ZS	Child employment in manufacturing, male (% of mal...
SL.MNF.0714.FE.ZS	Child employment in manufacturing, female (% of f...
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (m...
SL.IND.EMPL.MA.ZS	Employment in industry, male (% of male employmen...
SL.IND.EMPL.FE.ZS	Employment in industry, female (% of female emplo...
SL.GDP.PCAP.EM.KD	GDP per person employed (constant 2021 PPP $)
SL.FAM.WORK.ZS	Contributing family workers, total (% of total em...
SL.FAM.WORK.MA.ZS	Contributing family workers, male (% of male empl...
SL.FAM.WORK.FE.ZS	Contributing family workers, female (% of female ...
SL.FAM.0714.ZS	Children in employment, unpaid family workers (% ...
SL.FAM.0714.MA.ZS	Children in employment, unpaid family workers, ma...
SL.FAM.0714.FE.ZS	Children in employment, unpaid family workers, fe...
SL.EMP.WORK.ZS	Wage and salaried workers, total (% of total empl...

SL.EMP.WORK.MA.ZS	Wage and salaried workers, male (% of male employ...
SL.EMP.WORK.FE.ZS	Wage and salaried workers, female (% of female em...
SL.EMP.VULN.ZS	Vulnerable employment, total (% of total employme...
SL.EMP.VULN.MA.ZS	Vulnerable employment, male (% of male employment...
SL.SRV.0714.FE.ZS	Child employment in services, female (% of female...
SL.SRV.0714.MA.ZS	Child employment in services, male (% of male eco...
SL.AGR.0714.FE.ZS	Child employment in agriculture, female (% of fem...
SL.WAG.0714.ZS	Children in employment, wage workers (% of childr...
SL.SRV.EMPL.MA.ZS	Employment in services, male (% of male employmen...
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (m...
SL.TLF.0714.FE.ZS	Children in employment, female (% of female child...
SL.TLF.0714.MA.ZS	Children in employment, male (% of male children ...
SL.TLF.0714.SW.FE.TM	Average working hours of children, study and work...
SL.TLF.0714.SW.FE.ZS	Children in employment, study and work, female (%...
SL.TLF.0714.SW.MA.TM	Average working hours of children, study and work...
SL.TLF.0714.SW.MA.ZS	Children in employment, study and work, male (% o...
SL.TLF.0714.SW.TM	Average working hours of children, study and work...
SL.TLF.0714.SW.ZS	Children in employment, study and work (% of chil...
SL.EMP.VULN.FE.ZS	Vulnerable employment, female (% of female employ...
SL.TLF.0714.WK.FE.TM	Average working hours of children, working only, ...
SL.TLF.0714.WK.MA.TM	Average working hours of children, working only, ...
SL.TLF.0714.WK.MA.ZS	Children in employment, work only, male (% of mal...
SL.TLF.0714.WK.TM	Average working hours of children, working only, ...
SL.TLF.0714.WK.ZS	Children in employment, work only (% of children ...
SL.TLF.0714.ZS	Children in employment, total (% of children ages...
SL.TLF.PART.FE.ZS	Part time employment, female (% of total female e...
SL.TLF.PART.MA.ZS	Part time employment, male (% of total male emplo...
SL.TLF.PART.ZS	Part time employment, total (% of total employmen...
SL.WAG.0714.FE.ZS	Children in employment, wage workers, female (% o...
SL.WAG.0714.MA.ZS	Children in employment, wage workers, male (% of ...
SL.TLF.0714.WK.FE.ZS	Children in employment, work only, female (% of f...
SL.EMP.TOTL.SP.ZS	Employment to population ratio, 15+, total (%) (m...
SL.SRV.EMPL.FE.ZS	Employment in services, female (% of female emplo...
SL.EMP.TOTL.SP.FE.ZS	Employment to population ratio, 15+, female (%) (...
SL.EMP.TOTL.SP.MA.NE.ZS	Employment to population ratio, 15+, male (%) (na...
SL.EMP.TOTL.SP.MA.ZS	Employment to population ratio, 15+, male (%) (mo...

SL.EMP.TOTL.SP.NE.ZS	Employment to population ratio, 15+, total (%) (n...
SL.EMP.SMGT.FE.ZS	Female share of employment in senior and middle m...
SL.EMP.TOTL.SP.FE.NE.ZS	Employment to population ratio, 15+, female (%) (...
SL.EMP.SELF.MA.ZS	Self-employed, male (% of male employment) (model...
SL.EMP.SELF.FE.ZS	Self-employed, female (% of female employment) (m...
SL.EMP.MPYR.ZS	Employers, total (% of total employment) (modeled...
SL.SRV.0714.ZS	Child employment in services (% of economically a...
SL.EMP.MPYR.MA.ZS	Employers, male (% of male employment) (modeled I...
SL.EMP.SELF.ZS	Self-employed, total (% of total employment) (mod...
SL.EMP.1524.SP.ZS	Employment to population ratio, ages 15-24, total...
SL.EMP.1524.SP.NE.ZS	Employment to population ratio, ages 15-24, total...
SL.EMP.1524.SP.MA.ZS	Employment to population ratio, ages 15-24, male ...
SL.EMP.1524.SP.MA.NE.ZS	Employment to population ratio, ages 15-24, male ...
SL.EMP.1524.SP.FE.ZS	Employment to population ratio, ages 15-24, femal...
SL.AGR.0714.ZS	Child employment in agriculture (% of economicall...
SL.AGR.EMPL.FE.ZS	Employment in agriculture, female (% of female em...
SL.AGR.EMPL.MA.ZS	Employment in agriculture, male (% of male employ...
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment)...
SL.EMP.MPYR.FE.ZS	Employers, female (% of female employment) (model...
SL.EMP.1524.SP.FE.NE.ZS	Employment to population ratio, ages 15-24, femal...
SL.TLF.BASC.MA.ZS	Labor force with basic education, male (% of male...
SL.TLF.BASC.ZS	Labor force with basic education (% of total work...
SL.TLF.CACT.FE.NE.ZS	Labor force participation rate, female (% of fema...
SL.TLF.CACT.FM.ZS	Ratio of female to male labor force participation...
SL.TLF.CACT.FM.NE.ZS	Ratio of female to male labor force participation...
SL.TLF.CACT.MA.NE.ZS	Labor force participation rate, male (% of male p...
SL.TLF.CACT.MA.ZS	Labor force participation rate, male (% of male p...
SL.TLF.CACT.NE.ZS	Labor force participation rate, total (% of total...
SL.TLF.CACT.ZS	Labor force participation rate, total (% of total...
SL.TLF.INTM.FE.ZS	Labor force with intermediate education, female (...
SL.TLF.BASC.FE.ZS	Labor force with basic education, female (% of fe...
SL.TLF.CACT.FE.ZS	Labor force participation rate, female (% of fema...
SL.TLF.ADVN.ZS	Labor force with advanced education (% of total w...
SL.TLF.TOTL.IN	Labor force, total
SL.TLF.ADVN.FE.ZS	Labor force with advanced education, female (% of...
SL.TLF.ACTI.ZS	Labor force participation rate, total (% of total...

SL.TLF.ACTI.MA.ZS	Labor force participation rate, male (% of male p...
SL.TLF.TOTL.FE.ZS	Labor force, female (% of total labor force)
SL.TLF.INTM.MA.ZS	Labor force with intermediate education, male (% ...
SL.TLF.ACTI.FE.ZS	Labor force participation rate, female (% of fema...
SL.TLF.ACTI.1524.ZS	Labor force participation rate for ages 15-24, to...
SL.TLF.ACTI.1524.NE.ZS	Labor force participation rate for ages 15-24, to...
SL.TLF.ACTI.1524.MA.ZS	Labor force participation rate for ages 15-24, ma...
SL.TLF.ACTI.1524.MA.NE.ZS	Labor force participation rate for ages 15-24, ma...
SL.TLF.ACTI.1524.FE.ZS	Labor force participation rate for ages 15-24, fe...
SL.TLF.ACTI.1524.FE.NE.ZS	Labor force participation rate for ages 15-24, fe...
SL.TLF.ADVN.MA.ZS	Labor force with advanced education, male (% of m...
SL.TLF.INTM.ZS	Labor force with intermediate education (% of tot...
SM.POP.REFG	Refugee population by country or territory of asy...
SM.POP.TOTL	International migrant stock, total
SM.POP.TOTL.ZS	International migrant stock (% of population)
SM.POP.NETM	Net migration
SM.POP.REFG.OR	Refugee population by country or territory of ori...
per_allsp.ben_q1_tot	Benefit incidence of social protection and labor ...
per_si_allsi.cov_q3_tot	Coverage of social insurance programs in 3rd quin...
per_si_allsi.cov_q5_tot	Coverage of social insurance programs in richest ...
per_si_allsi.cov_q1_tot	Coverage of social insurance programs in poorest ...
per_si_allsi.cov_pop_tot	Coverage of social insurance programs (% of popul...
per_si_allsi.ben_q1_tot	Benefit incidence of social insurance programs to...
per_si_allsi.adq_pop_tot	Adequacy of social insurance programs (% of total...
per_sa_allsa.cov_q5_tot	Coverage of social safety net programs in richest...
per_sa_allsa.cov_q4_tot	Coverage of social safety net programs in 4th qui...
per_sa_allsa.cov_q3_tot	Coverage of social safety net programs in 3rd qui...
per_sa_allsa.cov_q2_tot	Coverage of social safety net programs in 2nd qui...
per_sa_allsa.cov_q1_tot	Coverage of social safety net programs in poorest...
per_sa_allsa.cov_pop_tot	Coverage of social safety net programs (% of popu...
per_sa_allsa.ben_q1_tot	Benefit incidence of social safety net programs t...
per_sa_allsa.adq_pop_tot	Adequacy of social safety net programs (% of tota...
per_lm_alllm.cov_q5_tot	Coverage of unemployment benefits and ALMP in ric...
per_lm_alllm.cov_q4_tot	Coverage of unemployment benefits and ALMP in 4th...
per_lm_alllm.cov_q3_tot	Coverage of unemployment benefits and ALMP in 3rd...
per_lm_alllm.cov_q2_tot	Coverage of unemployment benefits and ALMP in 2nd...

per_lm_alllm.cov_q1_tot	Coverage of unemployment benefits and ALMP in poo...
per_lm_alllm.cov_pop_tot	Coverage of unemployment benefits and ALMP (% of ...
per_lm_alllm.ben_q1_tot	Benefit incidence of unemployment benefits and AL...
per_lm_alllm.adq_pop_tot	Adequacy of unemployment benefits and ALMP (% of ...
per_allsp.adq_pop_tot	Adequacy of social protection and labor programs ...
per_allsp.cov_pop_tot	Coverage of social protection and labor programs ...
per_si_allsi.cov_q4_tot	Coverage of social insurance programs in 4th quin...
per_si_allsi.cov_q2_tot	Coverage of social insurance programs in 2nd quin...
SL.UEM.1524.FE.ZS	Unemployment, youth female (% of female labor for...
SL.UEM.TOTL.ZS	Unemployment, total (% of total labor force) (mod...
SL.UEM.1524.MA.NE.ZS	Unemployment, youth male (% of male labor force a...
SL.UEM.1524.NE.ZS	Unemployment, youth total (% of total labor force...
SL.UEM.1524.ZS	Unemployment, youth total (% of total labor force...
SL.UEM.ADVN.FE.ZS	Unemployment with advanced education, female (% o...
SL.UEM.ADVN.MA.ZS	Unemployment with advanced education, male (% of ...
SL.UEM.ADVN.ZS	Unemployment with advanced education (% of total ...
SL.UEM.BASC.FE.ZS	Unemployment with basic education, female (% of f...
SL.UEM.BASC.MA.ZS	Unemployment with basic education, male (% of mal...
SL.UEM.BASC.ZS	Unemployment with basic education (% of total lab...
SL.UEM.INTM.FE.ZS	Unemployment with intermediate education, female ...
SL.UEM.INTM.MA.ZS	Unemployment with intermediate education, male (%...
SL.UEM.INTM.ZS	Unemployment with intermediate education (% of to...
SL.UEM.NEET.FE.ZS	Share of youth not in education, employment or tr...
SL.UEM.NEET.MA.ZS	Share of youth not in education, employment or tr...
SL.UEM.NEET.ZS	Share of youth not in education, employment or tr...
SL.UEM.TOTL.FE.NE.ZS	Unemployment, female (% of female labor force) (n...
SL.UEM.TOTL.FE.ZS	Unemployment, female (% of female labor force) (m...
SL.UEM.TOTL.MA.NE.ZS	Unemployment, male (% of male labor force) (natio...
SL.UEM.TOTL.MA.ZS	Unemployment, male (% of male labor force) (model...
SL.UEM.TOTL.NE.ZS	Unemployment, total (% of total labor force) (nat...
SL.UEM.1524.FE.NE.ZS	Unemployment, youth female (% of female labor for...
SL.UEM.1524.MA.ZS	Unemployment, youth male (% of male labor force a...
LP.LPI.CUST.XQ	Logistics performance index: Efficiency of custom...