IPBeja
Instituto Politécnico de Beja

Escola Superior de Tecnologia e Gestão
Licenciatura em Engenharia Informática

Sistemas de Informação

Exploração e Desigualdade Laboral Global através de Dados Abertos

Data Mining / Machine Learning

João Augusto Costa Branco Marado Torres

Instituto Politécnico de Beja
Escola Superior de Tecnologia e Gestão

Beja, dezembro de 2025

Instituto Politécnico de Beja

Escola Superior de Tecnologia e Gestão
Licenciatura em Engenharia Informática

Sistemas de Informação

Exploração e Desigualdade Laboral Global através de Dados Abertos

Data Mining / Machine Learning

João Augusto Costa Branco Marado Torres

Trabalho realizado no âmbito da unidade curricular de Sistemas de Informação

Orientação

Dr.ª Isabel Sofia Sousa Brito

Beja, dezembro de 2025

Júri

Responsável: Dr.ª Isabel Sofia Sousa Brito
Vogal: Dr. João Paulo Trindade
Vogal: Dr.ª Elsa da Piedade Chinita Soares Rodrigues

Conteúdo

Conteúdo **i**

Lista de Abreviaturas e Siglas **ii**

1 Introdução **1**

1.1 Objetivos 1

1.2 Abordagem e estrutura do trabalho 2

2 Enquadramento e Ferramentas **3**

2.1 Dados abertos e desigualdade laboral 3

2.2 *Data mining* e descoberta de conhecimento 3

2.3 Ferramentas e bibliotecas utilizadas 3

3 Metodologia **4**

3.1 Modelo de dados e Seleção 4

3.2 Pré-processamento 5

3.3 Transformação dos dados 5

3.4 Processo de *data mining* (KDD) 5

4 Experiências e Resultados **6**

4.1 Avaliação e métricas 6

5 Discussão **7**

5.1 Interpretação económica e social 7

5.2 Limitações dos modelos 7

6 Conclusão **8**

Licença **9**

Lista de Abreviaturas e Siglas

ETL *Extract, Transform, Load*

FAIR *Findable, Accessible, Interoperable, Reusable*

FLOSS *Free Libre and Open Source Software*

FMI Fundo Monetário Internacional

GUI *Graphical User Interface*

IED Investimento Estrangeiro Direto

ILO *International Labour Organization*

OLAP *Online Analytical Processing*

ONU Organização das Nações Unidas

PIB Produto Interno Bruto

POSIX *Portable Operating System Interface*

PPC Paridade do Poder de Compra

URI *Uniform Resource Identifier*

1 Introdução

A primeira fase deste projeto centrou-se na recolha, integração e modelação de dados abertos sobre trabalho, rendimento e condições económicas à escala global, Foi a construção de um *data warehouse* multidimensional orientado para análise *Online Analytical Processing* (OLAP). Esta infraestrutura constitui a base técnica necessária para análises sistemáticas e comparáveis entre países, regiões e períodos temporais, mitigando problemas comuns de fragmentação e inconsistência dos dados.

Uma vez assegurada essa base, torna-se possível avançar para uma segunda etapa analítica, orientada não apenas para a descrição dos dados, mas para a identificação de padrões e relações estruturais. É neste contexto que técnicas de *data mining* e *machine learning* assumem relevância, ao permitir explorar grandes volumes de dados multidimensionais de forma sistemática, indo além da análise univariada ou de correlações simples.

No domínio das ciências sociais, estas técnicas não devem ser entendidas como instrumentos de previsão determinista, mas como ferramentas exploratórias e analíticas que auxiliam a identificação de eventos recorrentes, assimetrias estruturais e relações complexas entre variáveis económicas e sociais, podendo elas contribuir para uma leitura empiríca das dinâmicas de desigualdade e exploração no capitalismo contemporâneo, se plicadas de forma crítica e acompanhadas de métricas de avaliação adequadas.

Apesar da disponibilidade crescente de indicadores socioeconómicos, a análise das desigualdades laborais globais enfrenta desafios importantes. A existência de cada vez mais variáveis relevantes, países e regiões com realidades diferentes, entre outras coisas, dificultam a identificação de padrões estruturais através de abordagens analíticas tradicionais.

O problema central desta fase do trabalho consiste, assim, em determinar de que forma técnicas de *data mining* e *machine learning* podem ser aplicadas a um *data warehouse* multidimensional para identificar padrões relevantes nas relações entre produtividade, *labour share*, dependência externa e desigualdades regionais. Coloca-se igualmente a questão de como avaliar empiricamente os resultados obtidos, garantindo que os modelos utilizados são interpretáveis, validados e metodologicamente justificados.

1.1 Objetivos

O objetivo geral desta segunda parte do projeto é aplicar técnicas de *data mining* e *machine learning* aos dados previamente integrados, de modo a identificar padrões estruturais e testar empiricamente hipóteses relacionadas com a desigualdade e a exploração do trabalho à escala global.

De forma mais específica, pretende-se:

- selecionar e preparar subconjuntos de dados adequados à aplicação de técnicas de *machine learning*;
- aplicar métodos de análise não supervisionada, como *clustering*, para identificar grupos de países ou períodos com características socioeconómicas semelhantes;
- aplicar modelos supervisionados de classificação e regressão para analisar a relação entre variáveis como produtividade, *labour share* e indicadores de dependência externa;
- avaliar o desempenho dos modelos através de métricas apropriadas, como *precision*, *recall*, *F1-score* e erro quadrático médio;
- interpretar criticamente os resultados.

1.2 Abordagem e estrutura do trabalho

Metodologicamente, o trabalho segue o processo de *Knowledge Discovery in Databases* (KDD), compreendendo as etapas de seleção, pré-processamento, transformação, aplicação de técnicas de *data mining* e interpretação/avaliação dos resultados. Esta abordagem permite estruturar de forma clara e reprodutível o percurso analítico desde os dados brutos até à extração de conhecimento.

Do ponto de vista técnico, a implementação das análises é realizada em *Python*, para variar com a escolha da etapa anterior, *R*, recorrendo a bibliotecas livres amplamente utilizadas na área da ciência de dados, como `Pandas`, `NumPy`, `Matplotlib`, `Seaborn` e `Scikit-learn`. Esta escolha visa garantir a reprodutibilidade do trabalho, a transparência metodológica e a coerência com a opção por *software* livre adotada ao longo de todo o projeto.

2 Enquadramento e Ferramentas

2.1 Dados abertos e desigualdade laboral

Isto se baseia no relatório anterior

2.2 *Data mining* e descoberta de conhecimento

Há uma diferença entre análise OLAP descritiva e descoberta e modelagem de padrões.

2.3 Ferramentas e bibliotecas utilizadas

3 Metodologia

Existem muitas tarefas que involvem *data mining* que eu consigo fazer usando os dados da minha *data warehouse*. Fazemos *clustering* para separar países ou regiões em como sendo do centro ou da periferia. Prevemos qual vai será o país com a maior descida do *labour share* nos próximos anos. Conseguimos identificar períodos de crise e austeridade. Identificamos *outliers*.

Este trabalho segue o processo KDD. FALTA SÓ EXPLICAR CADA ETAPA.

3.1 Modelo de dados e Seleção

O modelo de dados é o resultado da etapa de engenharia de dados deste projeto. Consiste num modelo multidimensional de esquema floco de neve para data warehouse.

Para esta etapa agora, vou decidir usar as seguintes informações desse modelo:

- Informações geográficas acerca de países e regiões;
- Indicadores macroeconómicos como o *labour share*, a "produtividade", os fluxos IDE, o índice Gini, salários;
- O tempo.

A razão de escolhermos esses indicadores específicos é...

Os dados do tempo começam desde o ano 1960, mas com bastante falta de dados nos países periféricos da altura. Então vou diminuir um pouco a janela em 10 anos, que coincide com:

- Uma época de choques petrolíferos — a 16 de outubro de 1973, delegados da Organização dos Países Exportadores de Petróleo (OPEP) decidem aumentar a 70% o preço do petróleo, e no dia seguinte diferenciam os fornecimentos com base na posição dos países consumidores e relação à guerra do Yom Kippur. A 5 de março de 1979, no Irão a exportação de petróleo é retomada, em quantidade reduzida à metade em relação ao nível normal antes da crise;
- Uma época de estagflação — inflação alta e um alto nível de desemprego ao mesmo tempo;
- Foi também nessa altura que a Europa e os EUA acabaram com a conversibilidade do ouro, acelerando as suas financeirizações.

Todos estes dados e informações são comparáveis e de relevância.

UNIT OF ANALYSIS country–year

3.2 Pré-processamento

Apesar de já termos diminuido a janela do tempo, ainda vão faltar dados, por diversas razões. Os dados podem não ser disponibilizados de forma consistente, nem que seja anualmente. Na maioria das vezes a ausência dos dados estão estruturalmente relacionados com posições periféricas.

Os valores em falta foram tratados utilizando interpolação linear específica por país, aplicada apenas a lacunas internas curtas. Não foi realizada qualquer extrapolação, e as lacunas longas foram deixadas em falta. Esta abordagem preserva a continuidade temporal, evitando simultaneamente a fabricação de ausência de dados estruturais.

Também é importante pensar sobre os *outliers*, e como tratar dos mesmos. Mas muito desse trabalho consegue ser feito da etapa 3.3 já asseguir.

3.3 Transformação dos dados

Alguns dados fará sentido serem normalizados. Um uso da normalização na primeira etapa do projeto foi na produtividade que variava demasiado entre regiões, e que por isso desviava as atenções do que realmente eu queria analisar. Ao normalizar a produtividade, ”acabou-se” a variadade.

Um exemplo de uma possível transformação está nos fluxos IED, onde tu encontras regiões com fluxos positivos de milhares de milhões, talvez até bilhões de IED, e outras regiões com fluxo negativo de alguns milhões de IED. As desparidades são enormes.

Talvez seja necessário ocorrer ao uso a discretização para transformar variáveis contínuas (com valores infinitos) em variáveis discretas (onde tu consegues contar a quantidade de valor que a variável consegue ter). É uma transformação importante para modelos de classificação por exemplo.

Lagged variables...

É nesta etapa onde começamos a usufruir das bibliotecas Python já mencionadas.

3.4 Processo de *data mining* (KDD)

4 Experiências e Resultados

4.1 Avaliação e métricas

5 Discussão

5.1 Interpretação económica e social

5.2 Limitações dos modelos

6 Conclusão

Licença