FACULDADE ENGENHARIA
Departamento de Informática

Elementos de Inteligência Artificial e Ciência de Dados, Professor João Neves

Trabalho Prático: ***Análise de Indicadores Municipais Portugueses***

João Silva, a51557, IACD

Sofia Machado, a52152, IACD

Universidade da Beira Interior, Maio 2025

Índice

Introdução

O presente projeto, desenvolvido no âmbito da unidade curricular de **Elementos de Inteligência Artificial e Ciência de Dados**, tem como objetivo analisar estatísticas municipais de Portugal, com especial foco em três indicadores socioeconómicos: **taxa de desemprego**, **ganho médio mensal por trabalhador** e **população residente**.

A partir da recolha, integração e análise destes dados, recorreu-se a técnicas de análise exploratória e de aprendizagem não supervisionada para identificar padrões e relações entre as variáveis em estudo. O trabalho foi organizado em várias fases, respeitando a metodologia proposta: **Recolha de Dados**, **Integração de Dados**, **Análise Exploratória de Dados**, **Limpeza e Processamento de Dados** e **Análise Descritiva**.

Recolha de dados

Os dados utilizados foram recolhidos a partir da base de dados estatísticos **PORDATA**, ao nível dos municípios portugueses. Para este estudo, foram selecionadas as seguintes variáveis:

- **Taxa de Desemprego**
- **Ganho Médio Mensal por Trabalhador**
- **População Residente**

A informação foi obtida através do download de três ficheiros CSV:

- `Desemprego.csv`
- `Ganho_medio_mensal.csv`
- `populacao_residente.csv`

Script utilizado:
Estas operações foram implementadas no script **Tratamento_de_dados.py**, recorrendo às funções personalizadas carregar_datasets(), remover_linhas_sem_regiao() e preparar_df().

Integração de dados

O processo de integração de dados seguiu as seguintes etapas:

- **Carregamento dos ficheiros CSV** contendo os dados estatísticos de cada indicador.
- **Remoção das linhas** cuja coluna **Região** apresentava valores em falta.
- **Conversão das colunas** correspondentes aos indicadores para o formato numérico, assegurando a coerência dos dados.
- **Integração dos diferentes conjuntos de dados** num único DataFrame, recorrendo à junção das tabelas com base nas colunas **Ano** e **Região**.

Este procedimento permitiu uniformizar os dados e criar uma estrutura adequada para a fase de análise exploratória subsequente.

Análise exploratória de dados

A fase de análise exploratória de dados teve como objetivo realizar uma caracterização inicial dos indicadores estatísticos recolhidos, permitindo descrever o seu comportamento, avaliar a distribuição das variáveis e detetar eventuais anomalias, como valores em falta, valores atípicos (outliers) e inconsistências. Adicionalmente, procurou-se identificar padrões ou relações preliminares entre os diferentes indicadores, de forma a fundamentar as decisões a tomar nas fases subsequentes de limpeza, processamento e modelação dos dados. As técnicas aplicadas incluíram:

- **Cálculo de estatísticas descritivas** para as variáveis selecionadas, com o objetivo de resumir as principais características dos dados, nomeadamente médias, medianas, desvios padrão, valores mínimos e máximos, entre outros parâmetros relevantes.

- **Elaboração de um mapa de calor (heatmap) das correlações** entre as variáveis, com o intuito de identificar a existência de associações lineares, positivas ou negativas, entre os diferentes indicadores socioeconómicos analisados.

- **Construção de diagramas de caixa (boxplots)** para cada variável, permitindo analisar a respetiva distribuição, detetar a presença de valores atípicos (outliers) e avaliar a dispersão e assimetria dos dados.

Principais conclusões da Análise Exploratória:

- Verificaram-se **correlações moderadas entre os indicadores**, evidenciando algumas associações lineares que justificam a sua análise conjunta em fases posteriores.
- Identificaram-se **assimetrias significativas na distribuição de algumas variáveis**, o que reforça a necessidade de proceder à normalização dos dados e de atentar à influência de outliers.
- A análise das **tendências temporais revelou oscilações moderadas ao longo dos anos**, sem tendências acentuadas, mas com variações suficientes para justificar o controlo de valores extremos e a normalização antes da aplicação de técnicas de aprendizagem automática e agrupamento.

Limpeza e processamento de dados

Com base nos resultados da análise exploratória, procedeu-se às seguintes operações de limpeza e pré-processamento:

- **Remoção de outliers** com base no método do intervalo interquartil (IQR), eliminando valores fora de 1,5 vezes o intervalo interquartil acima do terceiro quartil ou abaixo do primeiro quartil.
- **Normalização das variáveis** através da técnica **StandardScaler**, de modo a uniformizar a escala dos dados para posterior aplicação de algoritmos de agrupamento.

Estas operações garantiram a qualidade e a integridade dos dados para a fase seguinte de análise descritiva.

Análise descritiva

Para a análise descritiva, recorreu-se a técnicas de **aprendizagem não supervisionada**, com o intuito de identificar padrões ocultos nos dados e agrupar os municípios com características semelhantes. As técnicas utilizadas foram:

- Visualização da distribuição dos dados com **t-SNE**, uma técnica de redução de dimensionalidade não linear adequada para clusters não lineares.

- **KMeans Clustering** com 3 clusters, aplicado às variáveis normalizadas, seguido de análise visual dos agrupamentos com base em projeções t-SNE e gráficos de dispersão.

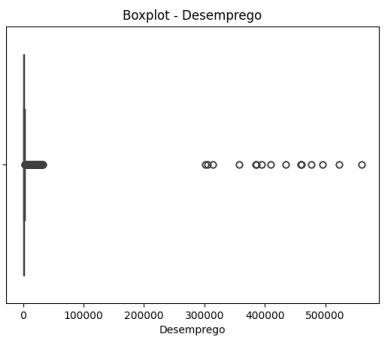
- Análise adicional por região, incluindo cálculo de rendimento per capita (salário/população), normalização e reagrupamento das regiões com KMeans.

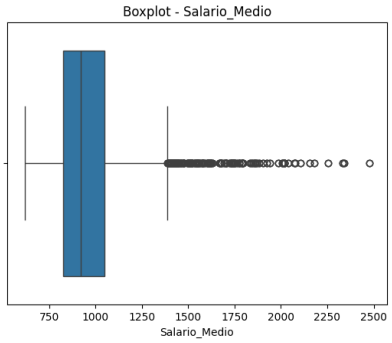
Análise por região

Para aprofundar a análise, foi realizada uma agregação dos dados por região, permitindo:

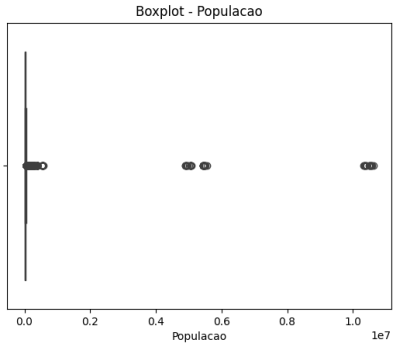
- Calcular o rendimento per capita médio por região;
- Normalizar os dados regionais (desemprego, salário médio, rendimento per capita);
- Aplicar o algoritmo KMeans para identificar agrupamentos de regiões com características socioeconómicas semelhantes;
- Visualizar os agrupamentos através de gráficos de dispersão e matrizes de correlação;
- Verificar relações entre os indicadores através de regressões lineares.

Interpretação dos resultados

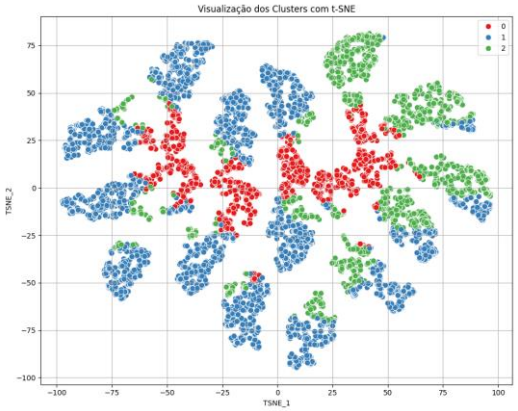




i

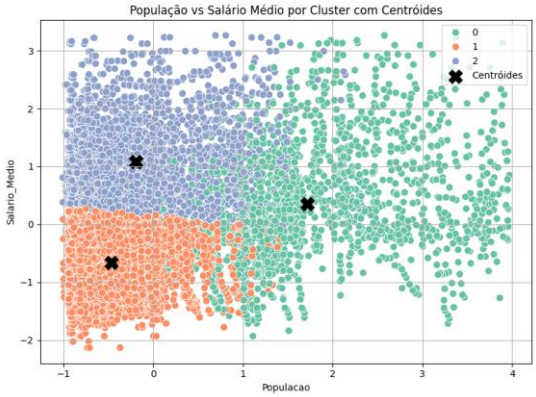


As análises gráficas acima permitiram visualizar a distribuição dos dados antes e depois da normalização, destacando a presença de outliers nas variáveis analisadas. A aplicação do método IQR foi essencial para eliminar valores extremos e garantir uma base de dados mais robusta para o clustering.

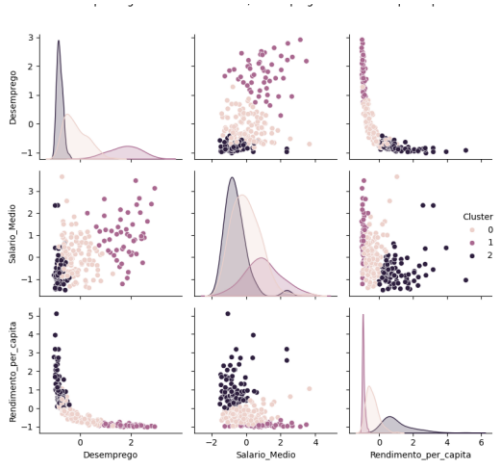


Através deste gráfico, é possível identificar clusters bem definidos, o que sugere a existência de padrões distintos nos dados. Cada ponto representa um município num determinado ano, e a sua posição relativa no espaço t-SNE indica semelhança com os outros pontos próximos. Os diferentes clusters (representados por cores distintas) agrupam municípios com características socioeconómicas semelhantes:

- Municípios com taxas de desemprego elevadas e salários médios baixos tendem a agrupar-se num mesmo cluster.
- Outros grupos podem representar zonas urbanas com alta população, maior rendimento médio e menor desemprego.



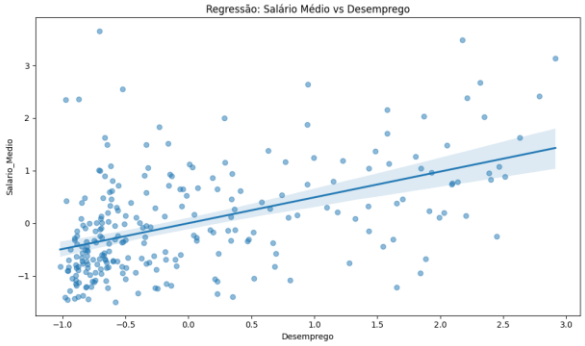
A análise de dispersão entre população e salário médio, com centróides destacados, reforça a distinção entre os clusters, mostrando que municípios mais populosos tendem a apresentar maiores salários médios.



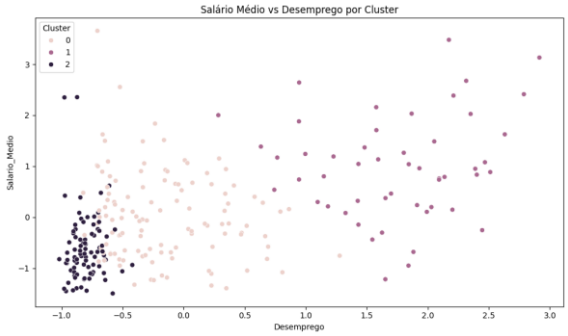
A partir deste gráfico conseguimos identificar três clusters principais:

- **Cluster 0**: Regiões com salários médios elevados e baixo desemprego, levando naturalmente a maior rendimento per capita. Este grupo pode representar regiões urbanas ou mais desenvolvidas.
- **Cluster 1**: Agrupa regiões com valores intermédios nos três indicadores, sugerindo equilíbrio relativo.

- **Cluster 2**: Inclui regiões com desemprego mais elevado, salários mais baixos e, consequentemente, menor rendimento per capita – potencialmente zonas mais vulneráveis. As distribuições KDE nas diagonais mostram as densidades dentro de cada cluster, e os gráficos de dispersão evidenciam correlações:
    - Salario_Medio e Rendimento_per_capita mostram relação positiva clara;
    - Desemprego parece inversamente relacionado com os outros dois indicadores nos clusters mais extremos.



Aqui observa-se uma tendência negativa, ainda que não extremamente acentuada, indicando que à medida que o desemprego aumenta, o salário médio tende a diminuir. Esta relação está alinhada com a teoria económica: regiões com maior desemprego costumam ter menos dinamismo económico e, por conseguinte, salários mais baixos. Apesar da dispersão dos pontos, o modelo linear capta uma direção clara na relação entre as variáveis.



Clusters distintos ocupam diferentes zonas do gráfico, indicando que o algoritmo de clustering conseguiu separar regiões com perfis económicos específicos.

Por exemplo:

- O **cluster 2** agrupa regiões com baixo desemprego e salários médios mais altos (indicando maior desenvolvimento económico);
- O **cluster 0** representa regiões com desemprego elevado e salários mais baixos, indicando fragilidade económica;
- O **cluster 1** mostra valores intermédios, servindo como grupo de transição entre os extremos.

A distribuição não é completamente linear, mas evidencia tendências visíveis de separação socioeconómica.

Conclusão

Este projeto analisou dados de desemprego, salário médio e população dos municípios portugueses. Após a limpeza e normalização dos dados, aplicou-se o algoritmo KMeans para identificar padrões e agrupar municípios com características semelhantes. A visualização com t-SNE e a análise por região revelaram relações claras entre os indicadores, como a correlação negativa entre desemprego e salário. A abordagem adotada permitiu extrair insights úteis para compreensão das desigualdades regionais.

---

i