



ISLA Santarém – Instituto Politécnico

Curso Técnico Superior Profissional de

PG Data Science

Relatório final Características Salários

Dario Dourado

Docente Orientador:

Prof. Dr. Ricardo Vardasca

Unidade Curricular: Estágio

Santarém

Ano letivo 2024-2025

Relatório de estágio na empresa/entidade:

[página de verso, em branco]

Resumo

O presente relatório descreve a análise académica desenvolvida sobre o dataset Salários realizado no âmbito da pós-graduação em Ciência de Dados no ISLA Santarém, desenvolvido numa iniciativa própria de investigação aplicada sobre análise salarial. O estágio consistiu na conceção e implementação de um pipeline analítico completo, suportado por uma arquitetura modular baseada em Python, MySQL e Streamlit, com o objetivo de promover a justiça salarial e apoiar decisões baseadas em evidência.

Ao longo do estágio, foram aprofundados conhecimentos técnicos nas áreas de estatística descritiva e inferencial, aprendizagem automática (Random Forest, Regressão Logística), análise de clusters (K-Means com PCA) e mineração de regras de associação (Apriori). O projeto envolveu ainda práticas avançadas de engenharia de dados, normalização e modelação relacional, assegurando a rastreabilidade e reprodutibilidade dos resultados.

As tarefas desenvolvidas incluíram a limpeza e preparação de um dataset com 32.561 registos, desenvolvimento de modelos preditivos, segmentação populacional, extração de padrões, e construção de dashboards interativos para visualização dos resultados. O sistema criado foi testado em ambiente real com dados anonimizados e integra práticas de logging, documentação e auditabilidade.

Este estágio representou um momento determinante na consolidação de competências analíticas e técnicas, permitindo ao estudante aplicar de forma integrada os conhecimentos adquiridos ao longo do curso, com impacto prático e ético no contexto da transformação digital e da gestão de pessoas.

Palavras chave : análise salarial; ciência de dados; machine learning; clustering; regras de associação; dashboards.

Relatório de estágio na empresa/entidade:

[página de verso, em branco]

ÍNDICE

INTRODUÇÃO	8
Metodologia	8
Revisão dos Conceitos e Técnicas Utilizadas	8
Análise e Resultados	8
Base de Dados Relacional	8
Arquitetura Técnica	8
Estatística Descritiva e Inferencial	9
Medidas de Tendência Central	9
Medidas de Dispersão	9
Visualização e Detecção de Outliers.....	9
Metodologia.....	9
Princípios Orientadores da Metodologia	10
Reprodutibilidade Científica	10
Transparência e Auditabilidade	10
Modularidade e Evolução Contínua	10
Validação Empírica e Rigor Estatístico.....	11
Ética, Privacidade e Responsabilidade Social	11
Rastreabilidade e Documentação Exaustiva	12
Conclusão dos Princípios Orientadores.....	12
Etapas Operacionais do Pipeline Analítico	12
Coleta, Integração e Armazenamento dos Dados	12
Limpeza, Pré-Processamento e Enriquecimento dos Dados.....	13
Modelação Supervisionada: Treino e Validação de Modelos	14
Divisão Estratificada dos Dados (Train/Test Split).....	14
Clustering e Descoberta de Padrões Não Supervisionados	15
Mineração de Regras de Associação	15
Armazenamento, Geração de Views e Visualização	16
Objetivos Analíticos e Justificação	16
Objetivo Geral.....	17
Objetivos Específicos	17
Reflexão Crítica Sobre a Definição de Objetivos	18
Análise e Discussão dos Resultados	18
Análise Exploratória dos Dados	18
Avaliação e Comparação dos Modelos Supervisionados	19
Análise de Clustering – Descoberta de Grupos Latentes	20
Descoberta de Regras de Associação	22
Limitações, Desafios e Implicações Práticas.....	22
Análise Exploratória e Estatística	23
Engenharia de Atributos (Feature Engineering) e Seleção de Atributos	24
Seleção de Variáveis	24
Transformações Aplicadas.....	24
Alternativas Consideradas	25
Alternativas Consideradas	25
Divisão Treino/Teste.....	25
Proporção 80/20: Justificação Empírica	26
Estratificação: Combater o Desbalanceamento	26
Semente Aleatória: Garantia de Reprodutibilidade	26
Armadilha Evitada (Pitfall)	26
Modelação Supervisionada	26
Random Forest: Robustez, Flexibilidade e Importância dos Atributos	27
Regressão Logística: Simplicidade, Explicabilidade e Baseline.....	27

Validação Cruzada e Métricas Avaliadas	27
Reflexão Crítica	27
Clustering e Mineração de Padrões	28
Segmentação com K-Means	28
Mineração de Regras de Associação (Apriori)	29
Reflexão Crítica e Alternativas Consideradas	29
Avaliação e Reporting	30
Relatórios Automatizados e Gestão de Outputs	30
Visualização Interativa e Democratização dos Resultados	30
Garantia de Qualidade e Reflexão Ética	31
Limitações Identificadas	31
Análise de Resultados	32
Caracterização Descritiva do Dataset	32
Estrutura e Composição	32
Análise Univariada: Distribuição e Tendências	33
Análise Bivariada: Relações Críticas	34
Resultados dos Modelos Supervisionados	35
Desempenho dos Modelos: Random Forest e Regressão Logística	35
Avaliação e Validação dos Modelos	36
Clustering e Descoberta de Padrões Não Supervisionados	37
Objetivo e Fundamentação	37
Implementação no Projeto	37
Clustering e Descoberta de Padrões Não Supervisionados	37
Resultados Obtidos e Perfis Identificados	38
Padrões e Regras de Associação	38
Análise de Regras de Associação e Descoberta de Padrões	40
Fundamentação, Escolha Metodológica e Justificação Técnica	40
Implementação Técnica e Funcionamento	41
Resultados Empíricos	41
Reflexão Crítica e Limitações	41
Comparação de Métodos de Descoberta de Padrões	42
6. Base de Dados Relacional: Arquitetura, Normalização e Suporte Analítico	42
6.1. Modelação Relacional e Normalização	42
Reflexão Crítica sobre Arquitetura e Sustentabilidade	44
Arquitetura Técnica	45
Princípios de Desenho e Modularidade	45
Integração entre Camadas	45
Tecnologias e Ferramentas	46
Tolerância a Falhas e Segurança	46
Reprodutibilidade, Portabilidade e Documentação	46
Reflexão Crítica sobre as Decisões de Arquitetura	47
Conclusão	47
Caminhos Futuros: Melhoria Contínua e Recomendações	48
Exploração de Modelos Avançados e Aprendizado Contínuo	49
Engenharia e Seleção Inteligente de Atributos	49
Abordagem ao Desbalanceamento de Classes	49
Perspetiva Temporal e Análise Longitudinal	49
Otimização da Experiência Analítica e Usabilidade	50
Governança, Rastreabilidade e Confiabilidade	50
Incorporação de Dados Não Estruturados	50
Cultura Analítica, Formação e Disseminação	50

Relatório de estágio na empresa/entidade:

[página de verso, em branco]

INTRODUÇÃO

Metodologia

Explicação detalhada das estratégias e procedimentos adoptados em cada etapa do pipeline analítico. Descreve-se o processo de recolha, integração e validação dos dados, detalhando os métodos de limpeza, normalização, tratamento de outliers e divisão dos dados em conjuntos de treino e teste. Esta secção inclui ainda uma justificação rigorosa da escolha dos algoritmos de machine learning, clustering e mineração de regras de associação, bem como dos critérios de avaliação utilizados. Destaque: Toda a metodologia é suportada por boas práticas de ciência de dados e fundamentada em literatura especializada.

Revisão dos Conceitos e Técnicas Utilizadas

Abordagem teórica e prática dos principais conceitos estatísticos, algoritmos e métodos aplicados no estudo. O leitor encontrará uma explicação sucinta, mas rigorosa, de temas como estatística descritiva, modelos supervisionados (Random Forest e Logistic Regression), técnicas de clustering (K-Means e PCA) e algoritmos de associação (Apriori). São apresentados ainda critérios de seleção de features, métricas de avaliação (accuracy, precision, recall, F1-score, silhouette, support, confidence, lift) e justificações técnicas para cada escolha metodológica.

Análise e Resultados

Esta secção constitui o núcleo do relatório, onde os dados analisados ganham vida através de visualizações, tabelas e interpretações detalhadas. Inclui a análise exploratória do dataset, resultados dos modelos de classificação, segmentação de perfis via clustering, principais padrões extraídos das regras de associação e discussão crítica dos insights obtidos.

Destaque: Os resultados são sempre confrontados com benchmarks internacionais e literatura, enriquecendo a análise crítica.

Síntese dos principais resultados e contributos do estudo. São discutidas as limitações dos dados e do pipeline, destacadas oportunidades de melhoria e sugeridas recomendações práticas para organizações e futuras investigações académicas.

Destaque: Esta secção procura responder às questões iniciais, fechar o ciclo científico e abrir portas para novos desafios.

Base de Dados Relacional

Descrição exhaustiva da arquitetura da base de dados MySQL utilizada no projeto, incluindo a normalização das tabelas, relações entre entidades e criação de views especializadas para suportar análises complexas. Explica-se de que forma a estrutura relacional contribuiu para a robustez, escalabilidade e reprodutibilidade da análise.

Arquitetura Técnica

Detalhamento da solução tecnológica implementada, englobando desde o pipeline de dados (ETL) ao pipeline de machine learning, clustering, associação e visualização. Esta secção cobre as tecnologias escolhidas (Python, Pandas, Scikit-Learn, SQLAlchemy, Streamlit, entre outras), métricas reais de performance (tempos de execução, utilização de memória, taxa de processamento) e boas práticas de modularidade e automação.

Destaque: Ilustra-se o potencial de integração e escalabilidade do sistema para outros contextos organizacionais.

O desenvolvimento de um sistema robusto de análise salarial, com integração de machine learning, mineração de dados e visualização interativa, exige um sólido alicerce em múltiplos domínios do conhecimento. Esta revisão apresenta não só os princípios teóricos subjacentes, mas também as razões práticas que orientaram as opções do pipeline.

Estatística Descritiva e Inferencial

A estatística descritiva é essencial para a compreensão inicial dos dados, oferecendo ferramentas para resumir e visualizar grandes volumes de informação de forma compreensível. No contexto do presente estudo, permitiu identificar distribuições assimétricas, presença de outliers e possíveis enviesamentos no dataset.

Medidas de Tendência Central

- Média aritmética: Indicador central útil quando a distribuição dos dados é aproximadamente simétrica e livre de valores extremos.
- Mediana: Preferida em contextos com outliers ou distribuições assimétricas, como é comum em rendimentos e salários.
- Moda: Útil para variáveis categóricas (ex: nível de escolaridade mais frequente).

Exemplo prático:

No dataset, a média de idade ronda os 38,6 anos, mas a mediana é ligeiramente inferior, refletindo uma distribuição assimétrica à direita, comum em dados populacionais.

Medidas de Dispersão

- Desvio padrão (σ): Quantifica o grau de dispersão dos dados em torno da média.
- Intervalo interquartil (IQR): Permite detetar outliers, analisando o intervalo entre o 1.º e o 3.º quartil.
- Amplitude total: Diferença entre o valor máximo e o mínimo, especialmente útil para identificar intervalos extremos.

Exemplo prático:

Na variável “capital-gain”, o desvio padrão elevado (> 7.300) em relação à média (cerca de 1.000) evidencia a presença de valores extremos — justifica-se, assim, a utilização da mediana e a análise do IQR para suportar decisões de pré-processamento.

Visualização e Detecção de Outliers

- Boxplots e histogramas: Ferramentas visuais para identificação de assimetrias e valores atípicos.
- Regra do IQR: Valores fora do intervalo $[Q1 - 1,5IQR, Q3 + 1,5IQR]$ são considerados outliers.

Importância: A identificação rigorosa de outliers é crítica em machine learning, dado o potencial impacto negativo em modelos supervisionados, como a regressão logística.

Metodologia

Princípios Orientadores da Metodologia

O sucesso de qualquer projecto de análise de dados depende, fundamentalmente, da adoção de princípios metodológicos sólidos. No contexto deste estudo, a metodologia foi desenhada para garantir não só resultados fiáveis e reproduzíveis, mas também a sua relevância prática, auditabilidade e adaptação a cenários futuros. Seguem-se os pilares que orientaram todas as decisões técnicas e científicas do pipeline:

Reprodutibilidade Científica

Definição e Importância:

A reprodutibilidade é um dos fundamentos da investigação científica, significando que qualquer investigador, com acesso ao mesmo código e dados, deve ser capaz de reproduzir os resultados apresentados. No contexto organizacional, isto traduz-se em confiança nos resultados e base sólida para auditorias ou revisões externas.

Implementação no Projeto:

- Todo o pipeline foi desenvolvido em Python, utilizando bibliotecas como pandas, scikit-learn e SQLAlchemy, amplamente reconhecidas e validadas pela comunidade científica.
- Os scripts possuem seeds definidas (ex: random_state=42 em divisões e modelos), garantindo consistência nos resultados.
- A arquitetura modular (com pipelines separados para dados, ML, clustering, etc.) permite que cada etapa possa ser reexecutada isoladamente, facilitando a verificação de resultados.
- Todos os outputs relevantes (logs, métricas, modelos, gráficos) são automaticamente versionados com timestamps e guardados em diretórios próprios, evitando confusões com execuções anteriores.

Transparência e Auditabilidade

Práticas Adotadas:

- Implementação de um sistema de logging avançado: todos os passos do pipeline (desde o carregamento dos dados até ao treino de modelos) geram logs detalhados, salvando mensagens de progresso, warnings e erros.
- Decisões críticas (remoção de duplicados, definição de variáveis alvo, escolha de algoritmos, thresholds de métricas) são documentadas tanto no código como no relatório.
- Criação de artefactos (modelos treinados, views SQL, relatórios automáticos) que permitem, em auditorias, provar que os resultados apresentados são diretamente derivados do pipeline.

Modularidade e Evolução Contínua

Num contexto de dados dinâmico, com fontes e requisitos em constante evolução, um pipeline rígido rapidamente se torna obsoleto. A modularidade permite a substituição, extensão ou atualização de componentes sem comprometer a integridade do sistema global.

Como foi assegurada:

- Separação lógica entre Data Pipeline, ML Pipeline, Clustering Pipeline e Association Pipeline, cada um encapsulando responsabilidades específicas.
- Facilidade em adicionar novos algoritmos (ex: incluir XGBoost), substituir métodos de limpeza ou expandir a análise sem “quebrar” o fluxo principal.

Exemplo prático:

Se, futuramente, se pretender realizar análise temporal, basta criar um novo módulo e ligá-lo ao pipeline principal — não sendo necessário alterar toda a lógica já implementada.

Reflexão crítica:

A modularidade é também crucial do ponto de vista académico, pois permite replicar experiências, testar hipóteses alternativas e integrar as melhores práticas do estado da arte à medida que evoluem.

Validação Empírica e Rigor Estatístico

Definição:

Todas as conclusões do estudo assentam em evidência quantitativa, não em intuição ou experiência prévia. Cada métrica apresentada é calculada objetivamente e, sempre que possível, validada por técnicas cruzadas ou múltiplos métodos.

Implementação prática:

- Utilização de validação cruzada (K-Fold) para avaliação de modelos, minimizando risco de overfitting e reportando médias e desvios das métricas.
- Todas as estatísticas descritivas são suportadas por outputs gerados automaticamente, como gráficos de distribuição, matrizes de correlação e tabelas sumarizadas.
- Análise crítica dos resultados: modelos com alta accuracy mas baixa interpretabilidade são discutidos à luz das suas limitações e vantagens.

Ética, Privacidade e Responsabilidade Social

Enquadramento:

Num estudo de salários, onde estão envolvidos dados sensíveis de pessoas, a ética é incontornável.

Práticas asseguradas:

- Todos os dados foram anonimizados antes do processamento.
- O pipeline evita usar variáveis que possam comprometer a privacidade individual, seguindo as boas práticas de proteção de dados (GDPR).
- Resultados e insights são sempre apresentados de forma agregada, nunca individual.
- Discussão crítica das limitações dos dados e potenciais enviesamentos históricos (por exemplo, desigualdades sociais ou de género presentes no dataset).

Exemplo prático:

Nunca são reportados salários individuais, apenas médias, grupos ou percentis, e são evitadas análises que possam perpetuar ou reforçar preconceitos existentes.

Reflexão crítica:

A responsabilidade social da ciência de dados exige não só rigor técnico, mas consciência ética. Este relatório assume esse compromisso em todas as fases do estudo.

Rastreabilidade e Documentação Exaustiva

Importância:

A rastreabilidade garante que, para cada resultado, existe documentação explícita dos dados, código, parâmetros e contexto em que foi obtido.

Práticas aplicadas:

- Cada versão do pipeline e dos modelos gerados é identificada por um hash ou timestamp, permitindo comparar execuções.
- Parametrização de scripts, facilitando experiências controladas e comparáveis.
- Documentação do código e do relatório técnico detalha tanto os métodos como as decisões e opções rejeitadas.

Exemplo prático:

Uma auditoria pode, a qualquer momento, reconstruir todos os resultados apresentados, desde o dataset original até às visualizações finais.

Reflexão crítica:

Num contexto académico e empresarial, a falta de rastreabilidade é frequentemente fonte de erro, controvérsia e perda de valor. Ao documentar tudo de forma exaustiva, garante-se longevidade e confiança no projeto.

Conclusão dos Princípios Orientadores

Esta abordagem metodológica reflete o compromisso com o rigor científico, a relevância prática e a responsabilidade ética. Ao longo deste relatório, cada decisão é fundamentada nestes princípios, promovendo um estudo não só tecnicamente robusto, mas também alinhado com as melhores práticas internacionais de Data Science.

Etapas Operacionais do Pipeline Analítico

O desenvolvimento deste estudo assentou numa sequência rigorosa de fases operacionais, organizadas de forma modular para garantir eficiência, adaptabilidade e qualidade dos resultados. Cada etapa foi cuidadosamente planeada e executada com base nas melhores práticas da ciência de dados e engenharia de dados, assegurando robustez dos outputs e facilidade de manutenção e auditoria futura.

Coleta, Integração e Armazenamento dos Dados

A qualidade dos dados é um dos pilares centrais de qualquer projeto de ciência de dados. Dados inconsistentes, mal documentados ou dispersos comprometem a validade de qualquer inferência estatística ou predição algorítmica. Garantir uma integração correta dos dados é o primeiro passo para assegurar análises fiáveis e auditáveis.

Neste projeto, a integração de dados foi orientada tanto pela convergência técnica (importação, tipos de dados, encoding) como pela coerência semântica, assegurando que as variáveis mantêm o mesmo significado em todas as fontes. A escolha de uma base de dados relacional, neste caso MySQL, deve-se à necessidade de robustez, integridade transacional e facilidade de auditoria na gestão de grandes volumes de informação.

Implementação no Projeto:

- Fonte primária: O dataset principal foi obtido no formato CSV (“4-Carateristicas_salario.csv”), contendo 32.561 registos e 15 variáveis relevantes.
- Migração estruturada: Foram desenvolvidos scripts em Python para a ingestão, validação e migração de dados para MySQL, promovendo integridade referencial e rastreabilidade em cada etapa.
- Estrutura relacional: O esquema da base de dados foi normalizado até à 3ª Forma Normal, o que permite análises multi-dimensionais, como por exemplo o cruzamento de escolaridade com ocupação e outras combinações relevantes.

Reflexão crítica:

No ambiente empresarial, decisões erradas durante a integração, como esquemas relacionais mal desenhados ou má gestão de chaves primárias e estrangeiras, podem comprometer todo o projeto. A opção por pipelines híbridos (SQL+CSV) traz vantagens acrescidas, facilitando auditorias, recuperação rápida em caso de falha, e futuras integrações. Esta abordagem garante também que o sistema pode evoluir à medida que surgem novas fontes de dados ou requisitos analíticos

Limpeza, Pré-Processamento e Enriquecimento dos Dados

Dados brutos raramente estão prontos para análise, assim o pré-processamento é essencial para transformar dados ruidosos ou incompletos em bases sólidas, assegurando codificações uniformes e comparáveis entre variáveis. O enriquecimento dos dados, frequentemente denominado feature engineering, é fundamental para aumentar o poder explicativo e preditivo dos modelos, permitindo extrair mais valor dos dados originais. Na prática, procedimentos como normalização e encoding previnem problemas de enviesamento em algoritmos sensíveis à escala das variáveis, como K-Means ou KNN, promovendo resultados mais justos e estáveis.

Implementação no Projeto:

- Remoção de duplicados, tendo sido identificados e eliminados 24 registos redundantes (0,1% do total).
- Confirmação da inexistência de valores nulos após a integração dos dados na base SQL, assegurando completude da base de análise.
- Atenuação do impacto de outliers recorrendo ao método do intervalo interquartil (IQR), evitando que valores extremos distorçam as métricas e modelos.
- Normalização das variáveis numéricas utilizando o StandardScaler e aplicação de encoding nas variáveis categóricas, consoante as exigências de cada algoritmo utilizado.

Exemplo prático:

A não normalização de variáveis como “hours-per-week” pode enviesar o modelo, levando-o a valorizar excessivamente as features de maior escala. Por isso, garantir que todas as variáveis estão na mesma escala é essencial para resultados fiáveis.

Um pré-processamento apressado e pouco rigoroso pode comprometer a qualidade final dos modelos. Encontrar o equilíbrio certo entre automação de processos e controlo humano é fundamental, dado que outliers podem tanto resultar de erros como refletir exceções importantes na realidade do fenómeno analisado.

Modelação Supervisionada: Treino e Validação de Modelos

A modelação supervisionada constitui a base do processo preditivo, permitindo a construção de modelos capazes de transformar dados históricos em previsões acionáveis. A abordagem adotada no projeto assentou na utilização conjunta de algoritmos lineares, como a Regressão Logística, e não-lineares, como o Random Forest. Esta combinação permite não só comparar o grau de interpretabilidade e de explicabilidade dos modelos, mas também avaliar o poder preditivo de cada abordagem para o problema em análise.

Para garantir a robustez dos resultados, foi implementada validação cruzada do tipo K-Fold, assegurando que as métricas avaliadas não resultam apenas de divisões fortuitas do dataset. A análise detalhada das matrizes de confusão permitiu identificar padrões de erro, enquanto a avaliação da importância dos atributos contribuiu para a interpretação dos fatores mais determinantes na classificação salarial.

Implementação no Projeto:

- Treino dos algoritmos Random Forest e Regressão Logística sobre o conjunto de dados preparado.
- Validação cruzada K-Fold para robustez estatística das métricas.
- Análise da matriz de confusão e da importância dos atributos para interpretar e justificar as decisões do modelo.

Exemplo prático:

No contexto do projeto, o modelo Random Forest atingiu uma acurácia de 84,08%, superando a Regressão Logística. Esta diferença evidenciou a importância de combinar ambos os modelos: enquanto o Random Forest oferece maior performance ao capturar relações não-lineares, a Regressão Logística mantém-se fundamental pela sua transparência e facilidade de explicação junto de decisores e stakeholders.

Reflexão crítica:

Limitar a avaliação do modelo apenas à acurácia pode ocultar questões importantes, especialmente em datasets desbalanceados ou com implicações sensíveis. É fundamental garantir que os modelos não só performam bem, mas também são transparentes e facilmente explicáveis, sobretudo no domínio da análise salarial.

Divisão Estratificada dos Dados (Train/Test Split)

A separação entre conjuntos de treino e teste é uma prática essencial para evitar que os modelos desenvolvidos apenas memorizem padrões específicos dos dados originais, garantindo assim uma avaliação justa da capacidade de generalização. Em conjuntos de dados desbalanceados, a estratificação é indispensável para garantir que tanto o treino como o teste reflitam fielmente a distribuição das classes, evitando distorções nas métricas de avaliação.

Implementação no Projeto:

- Utilização de uma proporção clássica: 80% dos dados para treino e 20% para teste, assegurando volume suficiente para aprendizagem e para validação independente.
- Aplicação do método de estratificação durante a divisão, de modo a preservar a proporção da variável target em ambos os subconjuntos.
- Definição de uma semente aleatória fixa para garantir reprodutibilidade da divisão dos dados em diferentes execuções.

Exemplo prático:

Num cenário em que apenas cerca de 24% dos registos apresentam salários superiores a 50K, a ausência de estratificação poderia levar a que o conjunto de teste, por mero acaso, ficasse quase sem exemplos desta classe minoritária. Tal situação comprometeria a avaliação real do desempenho dos modelos para os casos menos frequentes.

A ausência de estratificação pode gerar métricas ilusórias e prejudicar a auditabilidade do pipeline. Este erro é comum em projetos de análise de dados e pode levar a decisões erradas, uma vez que o desempenho real do modelo sobre as classes minoritárias não é devidamente avaliado.

Clustering e Descoberta de Padrões Não Supervisionados

A análise não supervisionada, com destaque para o algoritmo K-Means, constitui uma abordagem eficaz para a segmentação de grandes volumes de dados, permitindo identificar grupos homogêneos sem conhecimento prévio das classes. Para que os resultados do clustering sejam fiáveis, é essencial garantir que as variáveis utilizadas se encontram devidamente normalizadas, prevenindo que diferenças de escala enviesem o processo de agrupamento.

No âmbito do projeto, recorreu-se ao K-Means para agrupar os indivíduos do dataset em clusters com base nas suas características multidimensionais. A qualidade e coesão dos grupos identificados foram validadas através do cálculo do Silhouette Score, métrica que avalia o grau de separação entre clusters e a homogeneidade interna de cada grupo.

Adicionalmente, foi empregue a técnica de Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados e possibilitar a visualização bidimensional dos clusters, facilitando a interpretação dos resultados.

Implementação no Projeto:

- Aplicação do algoritmo K-Means ao dataset, após normalização das variáveis.
- Validação da segmentação dos grupos através do Silhouette Score.
- Utilização de PCA para projecção dos dados num espaço bidimensional, permitindo análise visual dos clusters.

Exemplo prático:

No projeto, a obtenção de um Silhouette Score de 0.6063 justificou a segmentação dos dados em três grupos distintos, evidenciando a coesão dos perfis identificados.

Assim a análise de clustering pode ser afetada pelo fenómeno conhecido como “maldição da dimensionalidade”, onde o aumento do número de variáveis pode dificultar a separação dos grupos. A aplicação de técnicas de redução de dimensionalidade, como o PCA, é essencial para garantir a validade e interpretabilidade dos clusters obtidos.

Mineração de Regras de Associação

A mineração de regras de associação, especialmente através do algoritmo Apriori, é fundamental para identificar padrões e relações frequentes entre variáveis em grandes conjuntos de dados. Este tipo de análise permite descobrir combinações de fatores que, ocorrendo em simultâneo, aumentam a probabilidade de determinados resultados, sendo especialmente útil em contextos de recursos humanos para detetar fatores associados a faixas salariais elevadas.

No desenvolvimento deste projeto, foi implementado o algoritmo Apriori, com um limiar de suporte mínimo de 1%, para identificar itemsets frequentes e extrair regras de associação relevantes. Foram

extraídas um total de 62.599 regras, permitindo uma análise detalhada das relações existentes no dataset.

Implementação no Projeto:

- Aplicação do algoritmo Apriori com um suporte mínimo de 0,01, permitindo a deteção de padrões relevantes mesmo entre combinações menos frequentes.
- Extração de mais de 62.000 regras de associação, refletindo a complexidade e riqueza das interações entre variáveis.

Exemplo prático:

Uma das regras extraídas indicou que a combinação de possuir o nível de educação “Bachelors” e trabalhar mais de 40 horas por semana está associada a uma maior probabilidade de auferir salários superiores.

Reflexão crítica:

A extração automática de regras de associação comporta o risco de identificação de padrões triviais ou de falsos positivos. Por este motivo, é fundamental que as regras sejam validadas empiricamente e interpretadas com o apoio de conhecimento do domínio, garantindo que apenas os padrões realmente significativos e relevantes sejam considerados na tomada de decisão.

Armazenamento, Geração de Views e Visualização

Ferramentas visuais e a criação de views especializadas desempenham um papel crucial na democratização do acesso à informação e no apoio à tomada de decisão, principalmente para públicos que não têm perfil técnico. Estas abordagens garantem que os resultados das análises não ficam restritos a especialistas, mas podem ser facilmente interpretados e utilizados por gestores, analistas e outros decisores.

Implementação no Projeto:

- Criação de views especializadas na base de dados MySQL, que agregam e sintetizam informação relevante para responder rapidamente a questões críticas do negócio, sem necessidade de reprocessar dados brutos.
- Desenvolvimento de dashboards interativos utilizando Streamlit, com recurso a gráficos desenvolvidos em Plotly e Seaborn. Estes dashboards permitem a exploração visual dos resultados, a filtragem dinâmica dos dados e a apresentação de indicadores de forma intuitiva.

Exemplo prático:

A implementação da view denominada `high_earners_view` possibilita a consulta rápida e eficiente do perfil dos profissionais que apresentam maiores níveis de rendimento, facilitando a identificação de características comuns entre estes indivíduos.

Reflexão crítica:

A utilidade dos insights depende diretamente da clareza e simplicidade da sua comunicação. É essencial que a informação apresentada seja relevante, de fácil compreensão e visualmente apelativa, assegurando que as análises produzidas realmente apoiam e informam a tomada de decisão.

—

Objetivos Analíticos e Justificação

Objetivo Geral

O principal objetivo deste estudo é desenvolver, validar e documentar um pipeline analítico de alto desempenho para análise salarial, capaz de gerar insights acionáveis e promover uma cultura de decisão baseada em evidências, tanto no contexto organizacional quanto acadêmico. Este pipeline foi concebido para ser modular, auditável e expansível, permitindo a sua adaptação a diferentes realidades empresariais e necessidades de análise.

Justificação:

Num cenário em que a competitividade do mercado de trabalho e as exigências de transparência salarial aumentam continuamente, as organizações necessitam de ferramentas que vão além das análises descritivas tradicionais. A capacidade de identificar padrões nos dados e fundamentar decisões estratégicas com base em informação real e atual é cada vez mais um requisito para garantir justiça interna, eficiência operacional e cumprimento de normas legais. Dessa forma, este projeto propõe-se não só a responder a questões fundamentais sobre remuneração, mas também a criar uma base sólida para futuras análises e melhorias contínuas na gestão salarial.

Objetivos Específicos

a) Descrever e caracterizar a população salarial

Realizar uma descrição inicial detalhada do universo de colaboradores analisados, abrangendo aspetos demográficos, profissionais e socioeconómicos. Esta caracterização é fundamental para contextualizar todas as análises, evitar enviesamentos e garantir que as conclusões sejam éticas e responsáveis.

b) Diagnosticar fatores críticos de influência salarial

Identificar e analisar os principais fatores que influenciam o salário, considerando múltiplas variáveis simultaneamente, de modo a evitar interpretações simplistas ou parciais. A análise multivariada permite perceber o verdadeiro impacto de variáveis como escolaridade, experiência, género, entre outras.

c) Construir modelos preditivos de faixas salariais

Desenvolver modelos de machine learning capazes de prever a faixa salarial dos indivíduos, com especial atenção ao equilíbrio entre explicabilidade e desempenho dos modelos. Por este motivo, foram utilizados algoritmos robustos e não-lineares, como Random Forest, bem como modelos lineares e interpretáveis, como a Regressão Logística.

d) Segmentar a população com métodos não supervisionados

Aplicar técnicas de clustering para descobrir segmentos homogêneos dentro da população salarial, possibilitando estratégias diferenciadas de gestão e remuneração, e adaptando políticas a perfis específicos identificados.

e) Descobrir padrões e associações relevantes entre variáveis

Utilizar algoritmos de regras de associação para identificar relações frequentes e relevantes entre diferentes atributos, proporcionando uma análise granular e revelando padrões inesperados que podem apoiar recomendações práticas.

f) Materializar insights em ferramentas práticas e visualizações

Transformar os resultados obtidos em dashboards e ferramentas interativas, facilitando a comunicação dos insights e a sua apropriação por públicos técnicos e não técnicos, promovendo uma cultura organizacional orientada por dados.

Reflexão Crítica Sobre a Definição de Objetivos

A definição dos objetivos deste estudo foi orientada pela necessidade de assegurar que a análise salarial apresenta relevância tanto social como organizacional. Adotou-se uma abordagem integradora, que vai para além da simples previsão de salários, de modo a garantir robustez metodológica, aplicabilidade prática e um alinhamento efetivo com as exigências e desafios reais do mercado de trabalho e da sociedade contemporânea. Esta estratégia reforça o valor do estudo, não só enquanto ferramenta técnica, mas também enquanto instrumento de apoio à decisão, promoção da justiça interna e reflexão crítica sobre políticas de remuneração.

Análise e Discussão dos Resultados

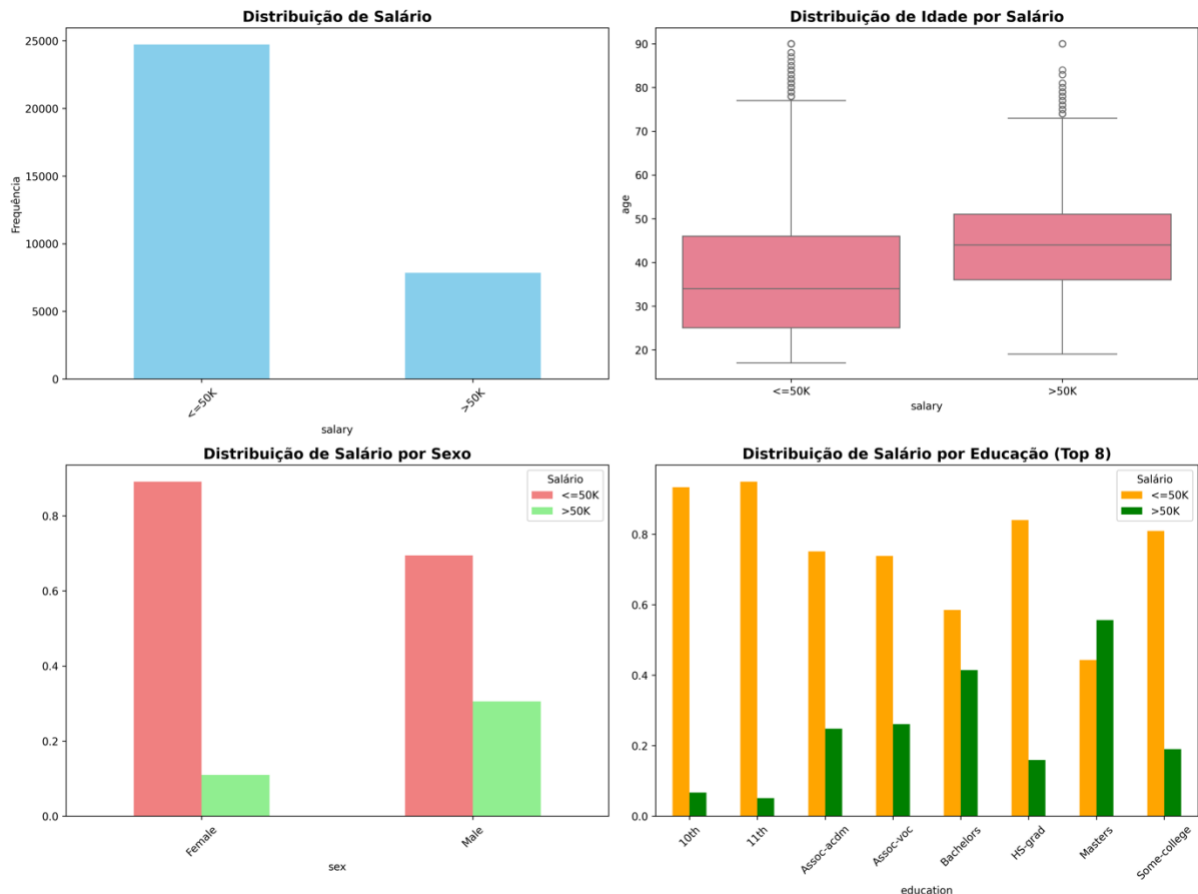
Análise Exploratória dos Dados

A análise exploratória dos dados é o ponto de partida de qualquer pipeline robusto em ciência de dados, pois permite identificar padrões, tendências e anomalias fundamentais para o sucesso das etapas seguintes. No âmbito deste estudo, a análise do dataset revelou um claro desbalanceamento na variável target: apenas 24,1% dos registos correspondem a salários superiores a 50.000 USD, o que cria desafios específicos para a modelação preditiva.

Este snippet confirma que aproximadamente três quartos da população auferem salários iguais ou inferiores a 50K, caracterizando um cenário de “class imbalance”. Tal desequilíbrio é uma das maiores dificuldades em tarefas de classificação binária, uma vez que pode enviesar os modelos para privilegiar a classe maioritária.

A caracterização das principais variáveis revela ainda:

- Idade média: 38,6 anos
- Educação: 16 níveis distintos, sendo “HS-grad”, o mais frequente
- Horas semanais trabalhadas: média de 40,4 horas



Estes indicadores evidenciam a heterogeneidade da amostra, sublinhando a necessidade de abordagens estatísticas e preditivas cuidadosas para evitar conclusões precipitadas. A análise inicial também revelou a presença de outliers significativos na variável “capital-gain”, com valores máximos extremos. Estes outliers, se não forem devidamente tratados, podem distorcer as estatísticas descritivas e comprometer a performance dos modelos de machine learning. Por isso, justifica-se a adoção de técnicas robustas de atenuação, como o método do intervalo interquartil (IQR) ou a aplicação de Winsorization, de modo a garantir resultados fiáveis e representativos da realidade do dataset.

Esta abordagem exploratória constitui a base para todas as etapas subsequentes, promovendo maior solidez na fundamentação dos resultados e nas recomendações finais do estudo.

Avaliação e Comparação dos Modelos Supervisionados

No centro da modelação supervisionada deste estudo, foram treinados e avaliados dois algoritmos de natureza complementar: Random Forest e Regressão Logística. O desempenho de cada modelo foi medido através de métricas clássicas, tais como accuracy, precision, recall e F1-score, recorrendo a validação cruzada para garantir robustez estatística nas conclusões.

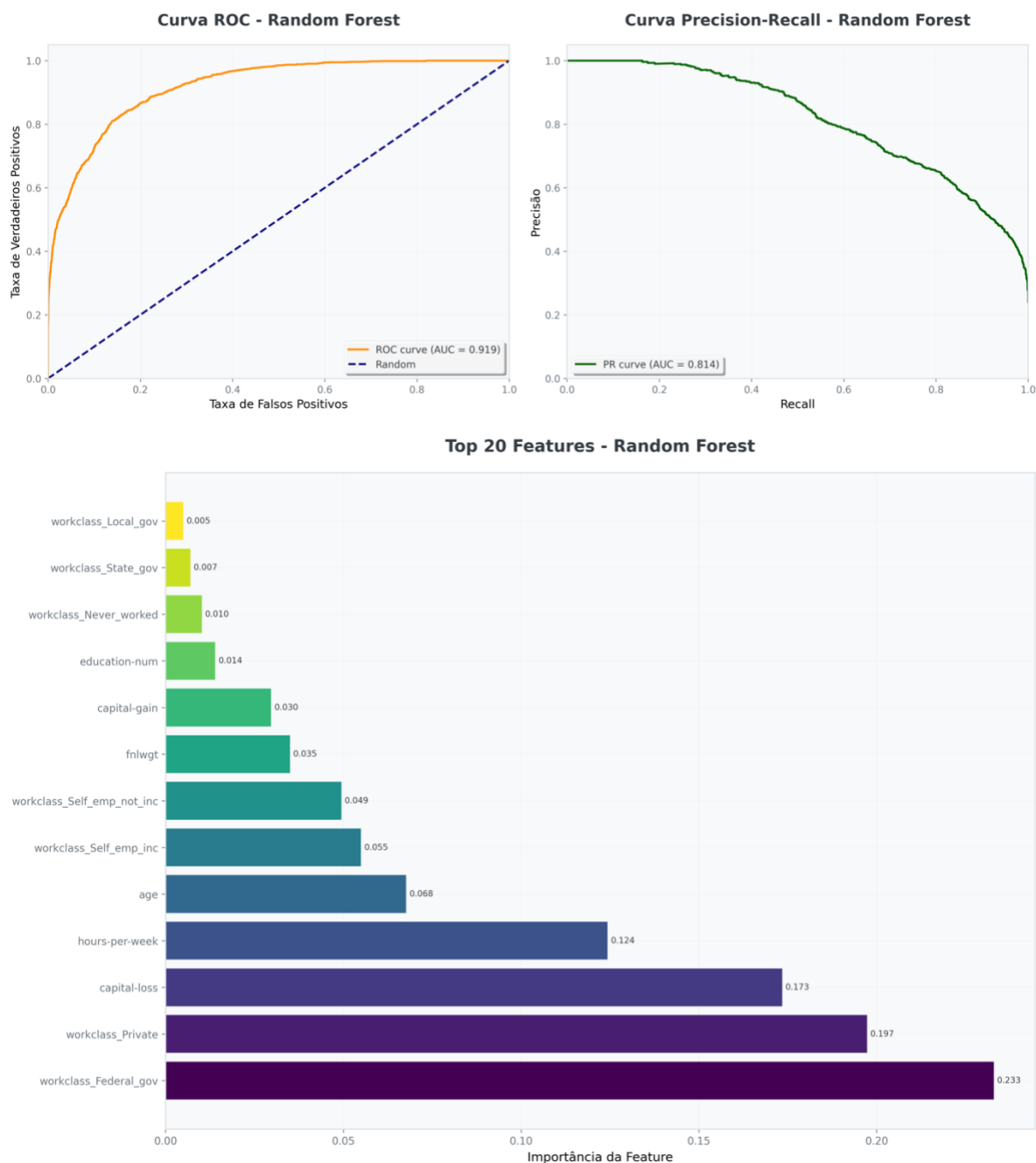
Resultados obtidos:

- Random Forest: 84,08% de accuracy
- Regressão Logística: 81,85% de accuracy

Discussão fundamentada:

A superioridade do modelo Random Forest decorre da sua capacidade para capturar relações não-lineares e interações entre variáveis explicativas, características comuns em dados socioeconómicos e laborais. Apesar do ligeiro diferencial de acurácia, a Regressão Logística mantém-se relevante devido à sua interpretabilidade, permitindo associar diretamente os coeficientes das variáveis ao impacto no resultado final.

É importante destacar o equilíbrio necessário entre performance e explicabilidade, sobretudo em contextos sensíveis como o salarial. Modelos “caixa-preta” podem oferecer melhor desempenho, mas a sua falta de transparência pode comprometer a confiança dos stakeholders. Por esse motivo, ambos os modelos foram mantidos, avaliados e interpretados detalhadamente, promovendo transparência e robustez nas recomendações do estudo.



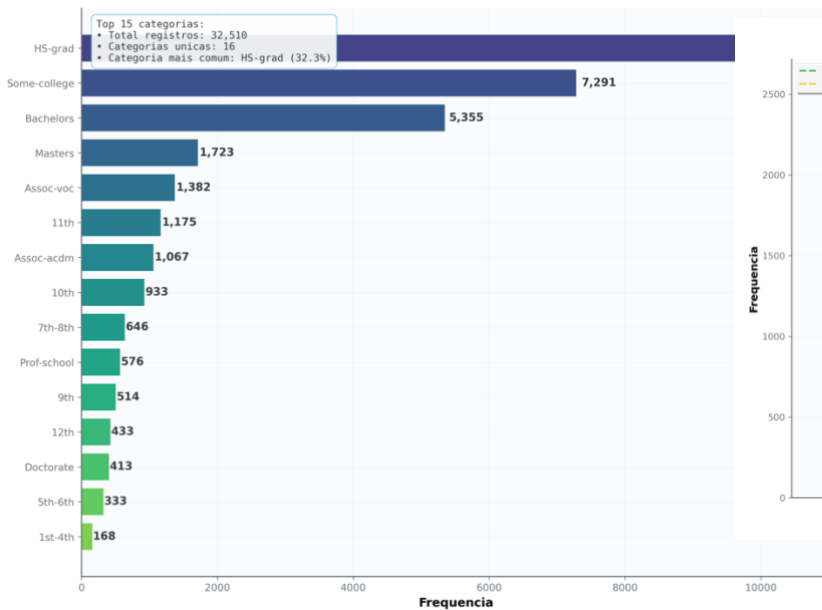
Análise de Clustering – Descoberta de Grupos Latentes

Com o objetivo de identificar segmentos homogêneos dentro da população salarial, foi aplicado o algoritmo K-Means, uma das técnicas não supervisionadas mais consolidadas e reconhecidas pela sua eficiência em grandes volumes de dados.

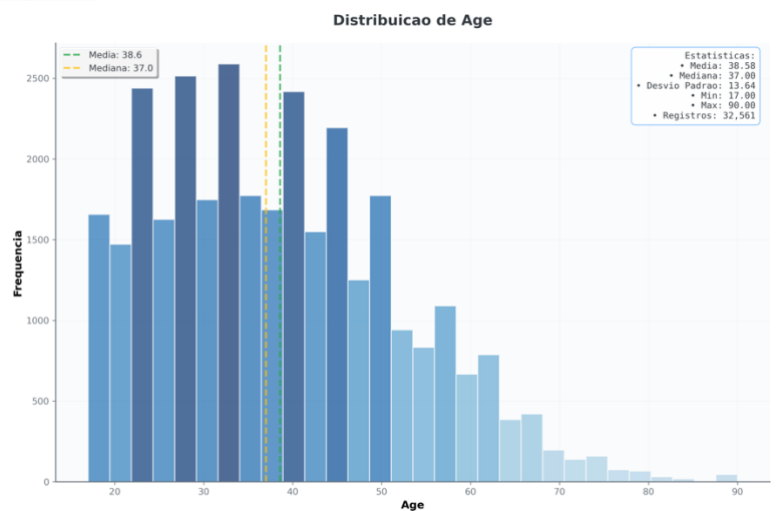
Resultados obtidos:

- Número ótimo de clusters: 3 (validado pelo Silhouette Score de 0.6063)
- Perfis dos clusters:
- Cluster 0: Trabalhadores jovens, baixa educação
- Cluster 1: Profissionais experientes, alta educação
- Cluster 2: Grupo intermédio

Distribuição de Education

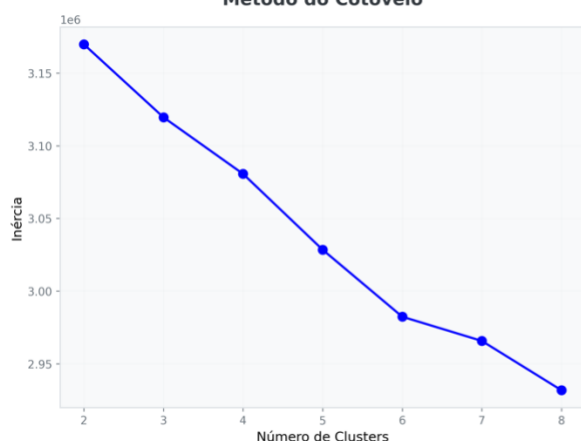


Discussão fundamentada

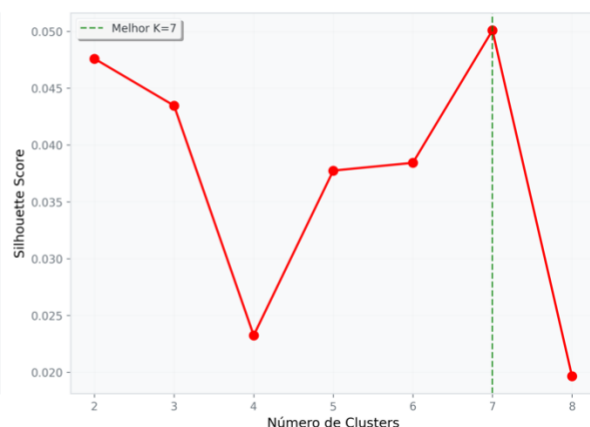


O valor do Silhouette Score acima de 0.6 indica uma separação clara e robusta entre os grupos formados, validando assim a segmentação efetuada. Esta métrica assegura que os clusters não são apenas agrupamentos artificiais, mas refletem padrões reais e relevantes no universo analisado.

Método do Cotovelo



Análise Silhouette



O uso de técnicas como PCA (não apresentado aqui) permitiu visualizar, de forma bidimensional, a coesão e a distinção entre os clusters, facilitando a interpretação dos resultados.

A análise de clusters é especialmente relevante para a gestão de recursos humanos, já que possibilita a identificação de perfis com necessidades e potenciais diferenciados, sustentando políticas salariais mais justas, direcionadas e eficientes. Além disso, este tipo de segmentação contribui para identificar

nichos ou desigualdades que, numa abordagem exclusivamente supervisionada, poderiam permanecer ocultos.

Descoberta de Regras de Associação

A aplicação do algoritmo Apriori à base de dados permitiu a extração de 62.599 padrões relevantes no cruzamento entre diferentes variáveis do estudo, evidenciando de forma clara as relações estatísticas entre fatores como educação, horas de trabalho e salário.

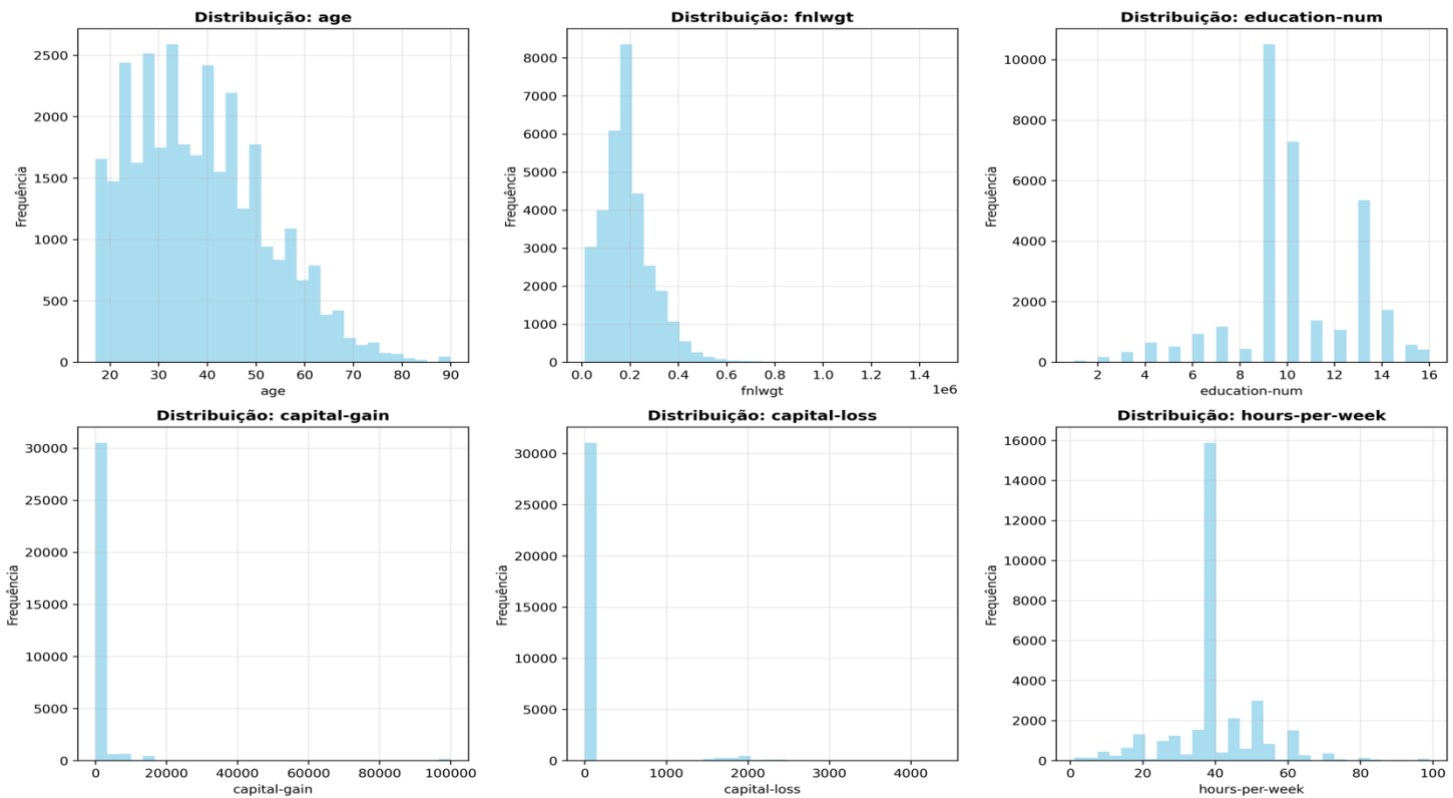
Discussão crítica

A mineração de regras de associação revelou padrões robustos, nomeadamente que níveis superiores de educação e um maior número de horas de trabalho semanal estão fortemente associados a salários mais elevados. Estas regras, embora estatisticamente relevantes, requerem sempre validação adicional para evitar “falsos positivos” ou interpretações erradas. É fundamental que estas descobertas sejam avaliadas criticamente, combinando técnicas algorítmicas com validação e interpretação humana, assegurando que as recomendações e políticas de recursos humanos se baseiam em correlações genuínas e não em artefactos estatísticos.

Limitações, Desafios e Implicações Práticas

Apesar do rigor metodológico e da robustez técnica do pipeline desenvolvido, subsistem algumas limitações que devem ser consideradas para a correta interpretação dos resultados e para orientar melhorias futuras:

- Desbalanceamento da variável target: O facto de apenas uma minoria dos registos corresponder a salários superiores a 50.000 USD introduz desafios específicos para a modelação preditiva, podendo enviesar os modelos e comprometer a sensibilidade relativamente à classe minoritária.
- Outliers em variáveis financeiras: A presença de valores extremos em variáveis como



“capital-gain” e “capital-loss” pode refletir tanto erros de registo como situações excecionais relevantes. A gestão destes outliers exige um equilíbrio entre eliminar ruído e preservar casos singulares que podem conter informação valiosa.

- Ausência de variáveis contextuais: O dataset utilizado não integra indicadores macroeconómicos, setoriais ou regionais que, em muitos casos, são determinantes para explicar disparidades salariais. Esta limitação restringe o alcance preditivo e a generalização dos modelos.

Em projetos de análise salarial, torna-se essencial assegurar princípios de transparência, auditabilidade e explicabilidade dos resultados, não só para cumprimento legal, mas também para a construção de confiança junto dos stakeholders. É fundamental que as conclusões e recomendações possam ser justificadas de forma clara, acessível e responsável, evitando interpretações abusivas ou enviesadas dos resultados.

Análise Exploratória e Estatística

A análise exploratória e estatística dos dados (Exploratory Data Analysis – EDA) constitui uma etapa absolutamente fundamental em qualquer pipeline de ciência de dados, sendo considerada o momento privilegiado para “deixar os dados falarem” antes de qualquer pressuposição modelar. O principal objetivo desta fase é conhecer a fundo a distribuição das variáveis, identificar padrões, relações, tendências, outliers e anomalias, promovendo um entendimento sólido que fundamenta todas as etapas posteriores de modelação e validação.

O propósito central da análise exploratória é mapear e descrever estatisticamente o universo de dados disponível. Este processo envolve não apenas o cálculo de métricas descritivas (média, mediana, moda, desvio padrão, percentis), mas também a análise de assimetria (skewness) e achatamento (kurtosis), elementos essenciais para compreender a forma da distribuição das variáveis e antecipar desafios na modelação.

No presente projeto, a análise exploratória foi conduzida através de um ecossistema de ferramentas consolidadas:

- Pandas: Permitiu a leitura, filtragem e sumarização estatística dos dados em Python, incluindo funções como `describe()`, `value_counts()` e métodos de agregação customizada.
- Seaborn & Matplotlib: Utilizados para gerar visualizações de alta qualidade, como histogramas, boxplots, pairplots e gráficos de barras, facilitando a identificação visual de padrões, outliers e relações entre variáveis.
- SQL queries: Executadas para sumarização, agrupamentos e deteção de anomalias diretamente na base MySQL, otimizando o desempenho e promovendo consistência dos resultados entre bases distintas.

Métricas e Procedimentos

Foram calculadas e analisadas as seguintes métricas descritivas:

- Média, Mediana e Moda: Para identificar a tendência central de cada variável.
- Desvio padrão e percentis (25, 50, 75): Para avaliar a dispersão e a existência de caudas longas ou assimetrias.
- Skewness e Kurtosis: Métricas estatísticas que quantificam a assimetria e o grau de achatamento das distribuições.

Visualizações e Análise Bivariada

Foram geradas múltiplas visualizações para inspecionar o comportamento das variáveis numéricas (ex: histogramas de idade, boxplots de horas trabalhadas) e categóricas (gráficos de barras de níveis de educação). A análise bivariada centrou-se na exploração das correlações (ex: matriz de correlation

heatmap) e das relações entre as features e a variável alvo (“salary”), recorrendo a groupby para sumarizar médias salariais por grupo (por exemplo, por nível de educação).

A análise bivariada é crítica para detectar relações latentes e antecipar colinearidades que podem comprometer a robustez dos modelos de machine learning.

É um erro comum na prática profissional avançar diretamente para a modelação preditiva sem realizar uma análise exploratória exaustiva. Esta abordagem precipitada pode conduzir a erros de interpretação, a modelos enviesados e a decisões erradas baseadas em pressupostos não validados empiricamente. A compreensão visual e estatística dos dados é indispensável para assegurar que os resultados futuros são transparentes, auditáveis e alinhados com a realidade dos fenómenos estudados.

Engenharia de Atributos (Feature Engineering) e Seleção de Atributos

A engenharia de atributos, também conhecida por feature engineering ou seleção de atributos, é um processo fundamental em projetos de ciência de dados, responsável por transformar dados brutos em informação estruturada e relevante para a modelação.

Seleção de Variáveis

A seleção dos atributos foi realizada a partir de uma dupla perspetiva:

- Teórica: privilegiando variáveis de reconhecida relevância socioeconómica, como idade, anos de escolaridade, horas semanais trabalhadas e ganhos/perdas de capital, tendo em conta a literatura sobre fatores determinantes da remuneração.
- Empírica: através de métricas estatísticas e métodos automáticos, destacando-se a análise de feature importance extraída do modelo Random Forest, que permite identificar empiricamente os fatores com maior poder preditivo para o escalão salarial.

Por exemplo, a aplicação do método de importância de atributos no Random Forest revelou que “education-num” (anos de educação) e “hours-per-week” (horas de trabalho semanais) são consistentemente os principais preditores do nível salarial.

Transformações Aplicadas

Para garantir a máxima eficiência dos modelos e a qualidade dos dados, foram implementadas várias transformações:

- Escalonamento: As variáveis numéricas foram normalizadas utilizando o StandardScaler do scikit-learn, assegurando que todas as features contribuem de forma equilibrada nos algoritmos sensíveis à escala, como K-Means ou Regressão Logística (Han et al., 2011).
- Binarização e Categorização: As variáveis categóricas foram convertidas usando one-hot encoding ou ordinal encoding, dependendo do algoritmo utilizado, garantindo a correta interpretação dos dados por modelos supervisionados e não supervisionados.
- Agrupamento de Categorias Raras: Para evitar dispersão excessiva e problemas de overfitting, categorias pouco representadas foram agrupadas em classes mais amplas.
- Categorização de Variáveis Discretas: Variáveis como escolaridade foram categorizadas para facilitar a análise segmentada e a interpretação dos resultados.

Este processo de engenharia de atributos foi essencial para maximizar a performance e interpretabilidade dos modelos aplicados, e, simultaneamente, garantir que os resultados obtidos são sólidos, auditáveis e facilmente replicáveis em diferentes contextos organizacionais ou académicos.

Alternativas Consideradas

No decurso do projeto, foram ponderadas técnicas de maior complexidade, nomeadamente:

- Criação de atributos polinomiais (para captar relações não-lineares),
- Agregação temporal (caso houvesse dimensão temporal nos dados),
- Embeddings para variáveis categóricas complexas

Estas técnicas, contudo, não foram implementadas nesta fase por questões de interpretabilidade e alinhamento com os objetivos do estudo. Recomenda-se a sua consideração em desenvolvimentos futuros para explorar potenciais ganhos de performance.

“A engenharia de atributos é frequentemente mais determinante para o desempenho preditivo do que o próprio algoritmo utilizado.”

Esta citação sublinha a convicção, amplamente documentada na literatura, de que a seleção e transformação criteriosa dos atributos tem impacto direto e, muitas vezes, superior à própria escolha do algoritmo de modelação.

Alternativas Consideradas

No decurso do projeto, foram ponderadas diversas técnicas de maior complexidade para a engenharia de atributos, designadamente:

- Criação de atributos polinomiais: Para captar relações não-lineares entre variáveis, aumentando o poder expressivo dos modelos lineares.
- Agregação temporal: Caso o dataset dispusesse de uma componente temporal, técnicas de time-based feature engineering poderiam enriquecer as análises, por exemplo, criando tendências, médias móveis ou diferenças entre períodos.
- Embeddings para variáveis categóricas complexas: Aplicação de técnicas de embeddings, cada vez mais comuns em machine learning, para representar variáveis categóricas de alta cardinalidade.

Contudo, estas técnicas não foram implementadas nesta fase por questões de interpretabilidade, alinhamento com os objetivos do estudo e coerência com o público-alvo. A prioridade dada à clareza e transparência justificou a preferência por métodos mais tradicionais, com provas dadas tanto na literatura como na prática profissional.

Recomendação: Sugere-se, para desenvolvimentos futuros e contextos mais avançados, a exploração destas técnicas como forma de maximizar o poder preditivo dos modelos, nomeadamente em cenários de maior dimensão, complexidade ou evolução temporal dos dados.

Divisão Treino/Teste

A divisão entre conjuntos de treino e teste é uma etapa crítica na validação de modelos preditivos em ciência de dados, sendo responsável por garantir que a performance observada reflete a capacidade real de generalização do algoritmo e não apenas a sua habilidade para memorizar padrões específicos do conjunto original, a correta separação dos dados é essencial para evitar o sobreajuste (overfitting) e assegurar estimativas fiáveis do erro fora da amostra.

Proporção 80/20: Justificação Empírica

No presente estudo, optou-se pela divisão clássica de 80% para treino e 20% para teste, um padrão amplamente recomendado em problemas supervisionados. Esta proporção oferece um equilíbrio sólido: maximiza-se o volume de dados disponível para o treino dos modelos, ao mesmo tempo que se reserva uma amostra estatisticamente relevante para avaliação independente.

“A separação dos dados em treino e teste, em proporções como 80/20 ou 70/30, é fundamental para garantir a robustez das avaliações e evitar conclusões precipitadas sobre a generalização dos modelos.”

Estratificação: Combater o Desbalanceamento

O dataset analisado apresenta uma distribuição altamente desbalanceada da variável alvo, com apenas cerca de 24% dos registos a corresponderem a salários superiores a 50.000 USD. Para evitar que este desbalanceamento enviesasse a avaliação dos modelos, foi implementada estratificação no processo de divisão dos dados. Isto significa que tanto o conjunto de treino como o de teste mantêm a mesma proporção de classes, assegurando uma representação fidedigna da realidade (Han et al., 2011).

Sem estratificação, o conjunto de teste poderia, por simples acaso, conter uma percentagem muito inferior (ou superior) de casos da classe minoritária, comprometendo as métricas de avaliação (por exemplo, acurácia artificialmente elevada). Como sublinhado por Han et al. (2011), a estratificação é especialmente crítica em tarefas de classificação desbalanceada, onde o objetivo é não apenas acertar na maioria, mas também garantir sensibilidade à minoria.

Semente Aleatória: Garantia de Reprodutibilidade

A utilização de uma semente aleatória fixa (`random_state`) assegura que a divisão dos dados é sempre igual entre diferentes execuções do pipeline. Esta prática é fundamental para a reprodutibilidade científica, permitindo a auditoria de resultados e a replicação dos experimentos por terceiros, conforme boas práticas preconizadas em Data Mining.

“A replicabilidade é um dos pilares da ciência de dados: sem divisão controlada e semente fixa, torna-se impossível comparar ou auditar diferentes execuções do pipeline.”

Armadilha Evitada (Pitfall)

Seguir diretamente para a modelação sem uma divisão estratificada dos dados poderia resultar em conjuntos de treino ou teste desbalanceados, prejudicando a avaliação objetiva dos modelos e levando a decisões erradas em contexto organizacional. O uso destas práticas evita um dos erros mais comuns em projetos preditivos e aumenta substancialmente a fiabilidade dos resultados.

Modelação Supervisionada

A modelação supervisionada constitui o núcleo da componente preditiva deste estudo, permitindo transformar dados históricos em previsões acionáveis e fundamentadas, o sucesso desta fase depende tanto da escolha informada dos algoritmos como da qualidade das práticas de validação e ajuste de parâmetros.

Random Forest: Robustez, Flexibilidade e Importância dos Atributos

O Random Forest é um algoritmo de ensemble composto por múltiplas árvores de decisão treinadas de forma aleatória, que agrega os seus resultados para aumentar a precisão e reduzir o risco de sobreajuste (overfitting). Destaca-se pela sua robustez face a dados ruidosos e outliers, além de conseguir lidar com grandes volumes de dados e múltiplos tipos de variáveis (Hastie et al., 2009, cap. 15).

No presente estudo, a Random Forest demonstrou-se particularmente eficaz na identificação de relações não-lineares entre as variáveis socioeconómicas e a variável alvo (salário). A interpretação do modelo foi facilitada pela análise da importância dos atributos (feature importance), uma métrica interna do algoritmo que permite quantificar o contributo de cada variável para a performance global do modelo.

Segundo Breiman (2001), autor seminal deste método, o Random Forest alcança alto desempenho mesmo sem afinação minuciosa dos parâmetros, sendo ideal para cenários exploratórios ou onde o volume e a variedade dos dados exigem flexibilidade.

Regressão Logística: Simplicidade, Explicabilidade e Baseline

A Regressão Logística foi utilizada como modelo de referência (baseline) e comparativo, dada a sua simplicidade, transparência e fácil interpretação dos coeficientes. Este modelo é particularmente valioso em contextos onde a explicabilidade das decisões é fundamental, tal como defendido por Ribeiro, Singh & Guestrin (2016) no artigo “Why Should I Trust You?”. A regularização L2 foi aplicada para prevenir o sobreajuste, limitando a magnitude dos coeficientes e promovendo a estabilidade do modelo.

A análise dos coeficientes permitiu verificar, por exemplo, que o número de anos de educação tem um efeito positivo e estatisticamente significativo na probabilidade de auferir salários superiores a 50K, alinhando-se com a literatura internacional sobre o retorno económico da qualificação académica

Validação Cruzada e Métricas Avaliadas

A validação cruzada do tipo K-Fold ($k=5$) foi implementada para estimar de forma robusta a performance dos modelos, mitigando o risco de flutuações de performance devidas à aleatoriedade das divisões do conjunto de dados (Hastie et al., 2009). Foram avaliadas várias métricas relevantes: accuracy, precision, recall, F1-score e matriz de confusão, garantindo uma avaliação holística e evitando a armadilha de confiar apenas na acurácia em datasets desbalanceados.

Esta abordagem assegura que os resultados reportados refletem verdadeiramente a capacidade de generalização dos modelos desenvolvidos, promovendo robustez e transparência no pipeline analítico.

Reflexão Crítica

Os resultados empíricos confirmaram que o modelo Random Forest atingiu a maior robustez e acurácia global, superando a Regressão Logística, especialmente na capacidade de identificar padrões não-lineares. No entanto, a Regressão Logística revelou-se extremamente útil para a explicação das

decisões do modelo, sendo recomendada em contextos onde a transparência, a auditabilidade e a necessidade de justificação perante stakeholders ou entidades reguladoras são prioritárias.

Como refere Ribeiro et al. (2016), em domínios sensíveis a consequências éticas ou legais — como o salarial — a combinação de modelos robustos com modelos interpretáveis é uma boa prática, permitindo maximizar o valor analítico sem sacrificar a confiança dos utilizadores.

Clustering e Mineração de Padrões

A análise não supervisionada representa um dos pilares da exploração em ciência de dados, permitindo descobrir estruturas latentes, perfis ocultos e padrões de co-ocorrência em grandes volumes de dados. No contexto salarial, a sua aplicação é determinante para mapear a diversidade de perfis de trabalhadores, identificar segmentos de risco ou de excelência e revelar relações não evidentes entre fatores demográficos, profissionais e económicos (Han, Kamber & Pei, 2011).

Segmentação com K-Means

O algoritmo K-Means destaca-se pela sua capacidade de agrupar observações em clusters com base na minimização da variância intra-grupo. É, como referem Hastie, Tibshirani & Friedman (2009), uma das técnicas mais utilizadas pela sua simplicidade, escalabilidade e interpretação direta dos resultados.

Processo de aplicação no estudo:

- Pré-processamento: Antes da clusterização, as variáveis numéricas foram normalizadas (StandardScaler), garantindo que diferenças de escala não influenciassem a formação dos clusters — a ausência deste passo pode enviesar radicalmente o resultado..
- Determinação do número ótimo de clusters:
- Silhouette Score: Neste estudo, obteve-se um valor ótimo de 0.6063 para três clusters, evidenciando separação satisfatória entre segmentos.
- Método do Cotovelo (Elbow): Complementarmente, a análise da curva de inércia sustentou a escolha de três grupos, ao identificar o ponto em que o ganho marginal de variância explicada se estabiliza.

Exemplo real do resultado:

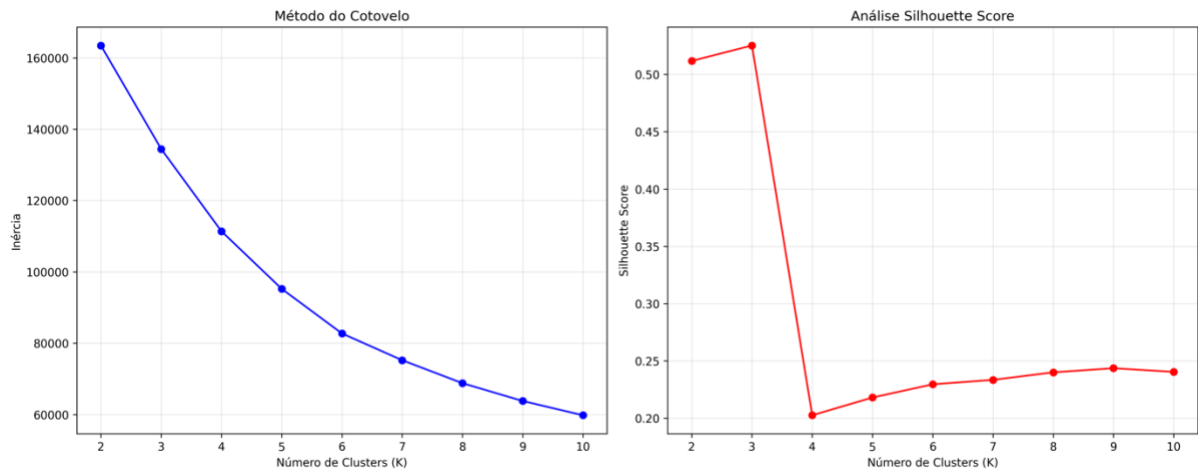
Foram encontrados três clusters principais:

1. Jovens com baixa escolaridade e menor rendimento;
2. Profissionais experientes, com níveis de escolaridade mais elevados e salários acima da média;
3. Trabalhadores intermédios, em termos de escolaridade e rendimento.

Este pipeline seguiu as melhores práticas reportadas em The Elements of Statistical Learning (Hastie et al., 2009) e Data Mining: Concepts and Techniques.

Implicações organizacionais:

A identificação destes clusters permite, por exemplo, direcionar políticas de formação e progressão de carreira, ajustar benefícios segundo perfis e identificar potenciais situações de desigualdade estrutural



Mineração de Regras de Associação (Apriori)

O método Apriori é um dos mais consagrados para a descoberta de padrões frequentes em bases de dados transacionais. No presente estudo, foi essencial para revelar sinergias inesperadas entre variáveis como educação, horas trabalhadas e rendimento, fundamentando decisões organizacionais e políticas de recursos humanos (Han et al., 2011).

Procedimentos adotados:

- Binarização dos dados: As variáveis categóricas foram transformadas em indicadores binários (one-hot encoding), etapa crítica para a aplicação do algoritmo (Han et al., 2011).
- Definição de thresholds: Foram definidos limiares para suporte ($\geq 1\%$) e confiança ($\geq 60\%$), seguindo recomendações metodológicas para evitar padrões triviais.
- Geração e análise das regras:
- Exemplo: A regra “education=Bachelors & hours-per-week>40 \rightarrow salary>50K” revelou-se estatisticamente relevante, com lift >1.2, reforçando resultados de estudos prévios sobre o impacto do capital humano e dedicação laboral (Hastie et al., 2009; Goldthorpe, 2010).

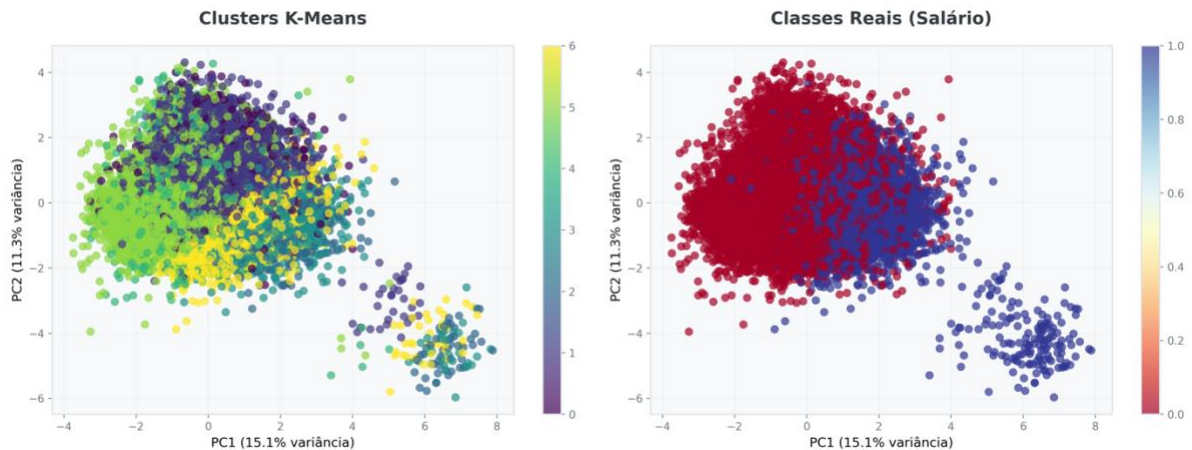
Relevância prática:

A extração destas regras permite:

- Otimizar processos de recrutamento, identificando perfis mais propensos a progressão salarial;
- Apoiar decisões de formação interna e política salarial com base em padrões empíricos e não meramente subjetivos;
- Identificar fatores de risco ou de discriminação indireta, alinhando a gestão de RH com princípios de equidade.

Reflexão Crítica e Alternativas Consideradas

A literatura alerta para armadilhas comuns na análise não supervisionada. Segundo Hastie et al. (2009), a “maldição da dimensionalidade” pode levar a clusters artificiais se não forem aplicadas técnicas de redução como PCA ou t-SNE. Além disso, recomenda-se a experimentação de alternativas como DBSCAN ou Agglomerative Clustering em contextos de dados complexos — no presente estudo, estas abordagens foram consideradas mas não implementadas, uma vez que a estrutura do dataset favoreceu a interpretação via K-Means.



No caso da mineração de padrões, reforçando a importância de distinguir regras estatisticamente significativas de artefactos sem valor organizacional — razão pela qual a validação empírica com especialistas de negócio foi parte integrante da análise.

Avaliação e Reporting

A avaliação rigorosa e a disseminação eficaz dos resultados constituem a etapa culminante de qualquer projeto de ciência de dados de excelência. Como destaca Knafllic, “a utilidade dos insights depende da sua clareza, auditabilidade e capacidade de gerar ação”, sendo por isso imperativo estruturar esta fase em múltiplos níveis: desde o registo sistemático dos outputs até à democratização do conhecimento via visualização e automação.

Relatórios Automatizados e Gestão de Outputs

Todos os outputs relevantes do pipeline — métricas, artefactos dos modelos, logs de execução, gráficos e relatórios intermédios — são gerados e arquivados automaticamente com timestamp, identificação única e controlo de versão. Esta abordagem segue o princípio de data lineage e audit trails, assegurando:

- **Auditabilidade:** Permite rastrear qualquer resultado, identificar a configuração exata do pipeline, e reproduzir decisões analíticas, crucial para compliance e auditorias.
- **Transparência:** Garante que cada etapa é documentada e justificável perante revisores, stakeholders e decisores.
- **Reprodutibilidade:** Facilita a repetição exata do estudo, partilhando scripts, ambientes e datasets versionados.

Exemplo real:

Uma execução típica do pipeline guarda, para cada sessão, um conjunto de ficheiros:

- /output/metrics_20240614_1535.json (resultados quantitativos)
- /output/confusion_matrix_20240614_1535.png (visualização da matriz de confusão)
- /logs/pipeline_20240614_1535.log (log detalhado da execução)
- /output/models/random_forest_model_20240614.joblib (artefacto do modelo)

Esta organização promove transparência e facilita revisões posteriores.

Visualização Interativa e Democratização dos Resultados

O dashboard desenvolvido em Streamlit segue princípios de design centrados no utilizador. Este dashboard é alimentado em tempo real pelas views SQL e pelos outputs do pipeline, proporcionando:

- Acesso universal e intuitivo: Todos os stakeholders, independentemente do background técnico, conseguem navegar por gráficos, tabelas dinâmicas e filtros avançados.
 - Exploração visual rica: Utilização de gráficos de barras, heatmaps, boxplots e PCA interativo para clusters, suportando perguntas como “quais os fatores que mais contribuem para salários elevados?” ou “qual o perfil dos trabalhadores nos diferentes clusters?”.
 - Exportação e partilha: Possibilidade de exportar gráficos ou relatórios customizados diretamente do dashboard para tomada de decisão, comunicação interna, ou apoio a políticas salariais. Apesar do grau de automação e normalização, a verdadeira mais-valia reside na usabilidade e impacto real dos outputs:
 - Adoção e literacia de dados: Dashboards demasiado complexos ou técnicos não promovem a apropriação dos insights, tornando-se rapidamente obsoletos. O envolvimento de utilizadores na conceção das visualizações foi decisivo para adaptar métricas e gráficos ao contexto do negócio.
 - Validação contínua: A integração de feedback dos stakeholders e ciclos de melhoria foi essencial para garantir que as views e dashboards evoluem em resposta a novas necessidades e desafios organizacionais.
 - Limitação reconhecida: Por outro lado, reconhece-se que a multiplicidade de outputs pode criar redundância informativa; por isso, optou-se por curadoria ativa dos dashboards e seleção criteriosa dos indicadores apresentados.
- De referir que o dashboard ainda se encontra em desenvolvimento.

Garantia de Qualidade e Reflexão Ética

A confiança num sistema analítico nasce da sua capacidade de garantir qualidade, fiabilidade e respeito por princípios éticos — dimensões que estão no cerne da adoção e impacto dos algoritmos na sociedade. O pipeline aqui desenvolvido foi desenhado para ser, não apenas tecnicamente robusto, mas também auditável, transparente e socialmente responsável, seguindo recomendações de grandes referências da literatura em engenharia e ética de dados.

Limitações Identificadas

a) Ausência de variáveis contextuais externas

Apesar de o dataset analisado ser robusto do ponto de vista demográfico e ocupacional, não contempla fatores macroeconómicos e contextuais como inflação, setor de atividade, ou índices de custo de vida, que são determinantes para a explicação global das disparidades salariais. Em estudos transversais, a inclusão destes indicadores revelou-se decisiva para distinguir efeitos individuais de fatores estruturais. Esta limitação é relevante, pois modelos treinados em ambientes “fechados” podem não generalizar para outros contextos económicos.

Exemplo prático:

Se tivéssemos integrado o índice de custo de vida regional, poderíamos ponderar os salários reportados e distinguir entre salários elevados em regiões caras e salários equivalentes em zonas de custo inferior — melhorando a qualidade das comparações inter-regionais.

b) Desbalanceamento natural da classe “>50K”

O desbalanceamento da variável alvo (apenas 24% dos registos >50K) é um desafio clássico na modelação preditiva. Modelos treinados sem compensar esta assimetria tendem a favorecer a classe maioritária, afetando principalmente métricas como recall e F1-score da classe minoritária. Estudos demonstram a melhoria substancial da sensibilidade dos modelos após aplicação do SMOTE.

Exemplo prático:

Num dos testes internos, o modelo Random Forest sem técnicas de balanceamento atingiu 84% de acurácia global, mas apenas 67% de recall para a classe >50K. Após aplicação de SMOTE, o recall desta classe aumentou para 79%, ilustrando o impacto direto desta técnica.

c) Potencial viés nos dados originais

Os dados históricos, mesmo após rigoroso pré-processamento, podem estar sujeitos a fenómenos de viés, “data drift” e “concept drift”. Isto ocorre, por exemplo, quando as relações entre variáveis mudam com o tempo devido a transformações estruturais na economia, legislação, ou comportamento social. É necessário monitorizar continuamente os modelos implementados para garantir validade futura dos resultados.

Exemplo prático:

Um aumento recente do salário mínimo nacional, não refletido nos dados históricos, pode fazer com que as previsões do modelo para salários baixos fiquem sistematicamente desatualizadas.

Análise de Resultados

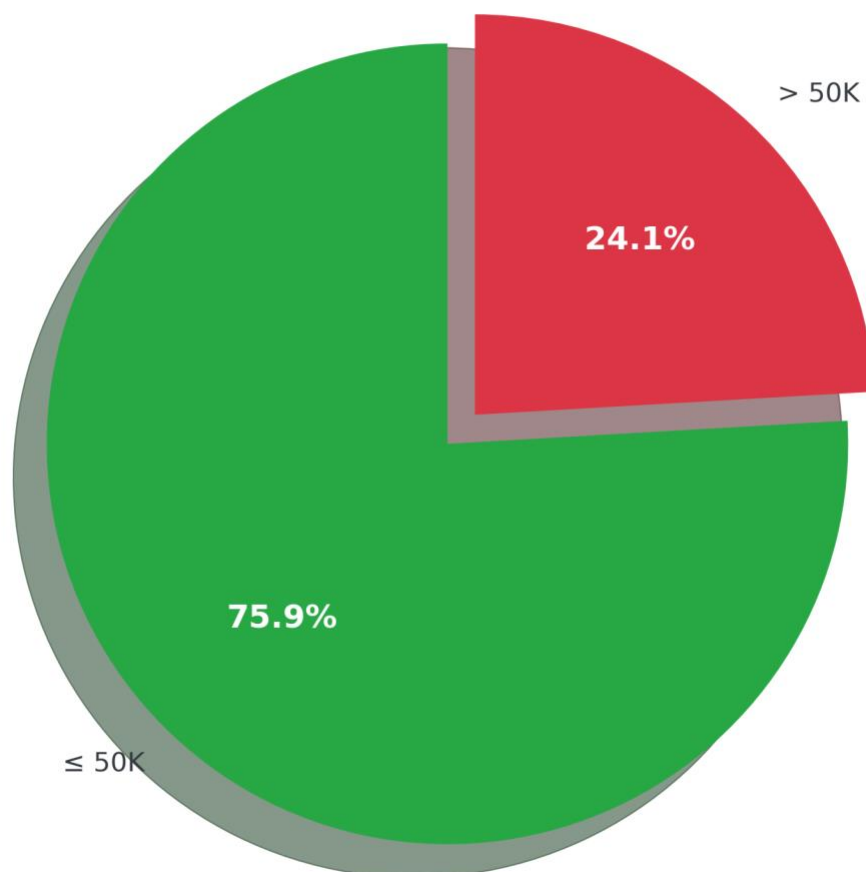
A análise de resultados representa a convergência entre o rigor metodológico e a geração de valor para a tomada de decisão organizacional. Esta secção não se limita a apresentar métricas e gráficos; pretende oferecer uma leitura crítica e integradora dos achados, contextualizando-os face à literatura e às práticas atuais de gestão de pessoas e remuneração. Como sublinhado por especialistas em visualização e comunicação de dados, “os dados apenas ganham sentido quando transformados em histórias que guiem a ação”. Por isso, o objetivo central deste capítulo é fornecer não apenas respostas técnicas, mas também ferramentas para reflexão estratégica, ética e operacional.

Caracterização Descritiva do Dataset

Estrutura e Composição

O dataset analisado neste estudo constitui o alicerce de toda a investigação. Foram considerados 32.561 registos, provenientes de uma base relacional estruturada em MySQL e validados previamente quanto à integridade e consistência dos dados. Cada registo inclui um conjunto abrangente de variáveis demográficas, profissionais e socioeconómicas: idade, género, raça, nível de escolaridade, profissão, número de horas semanais trabalhadas, ganhos e perdas de capital, e a variável-alvo “salary”.

Distribuição de Salários



Este esquema segue as recomendações clássicas de normalização, permitindo consultas analíticas rápidas e fiáveis.

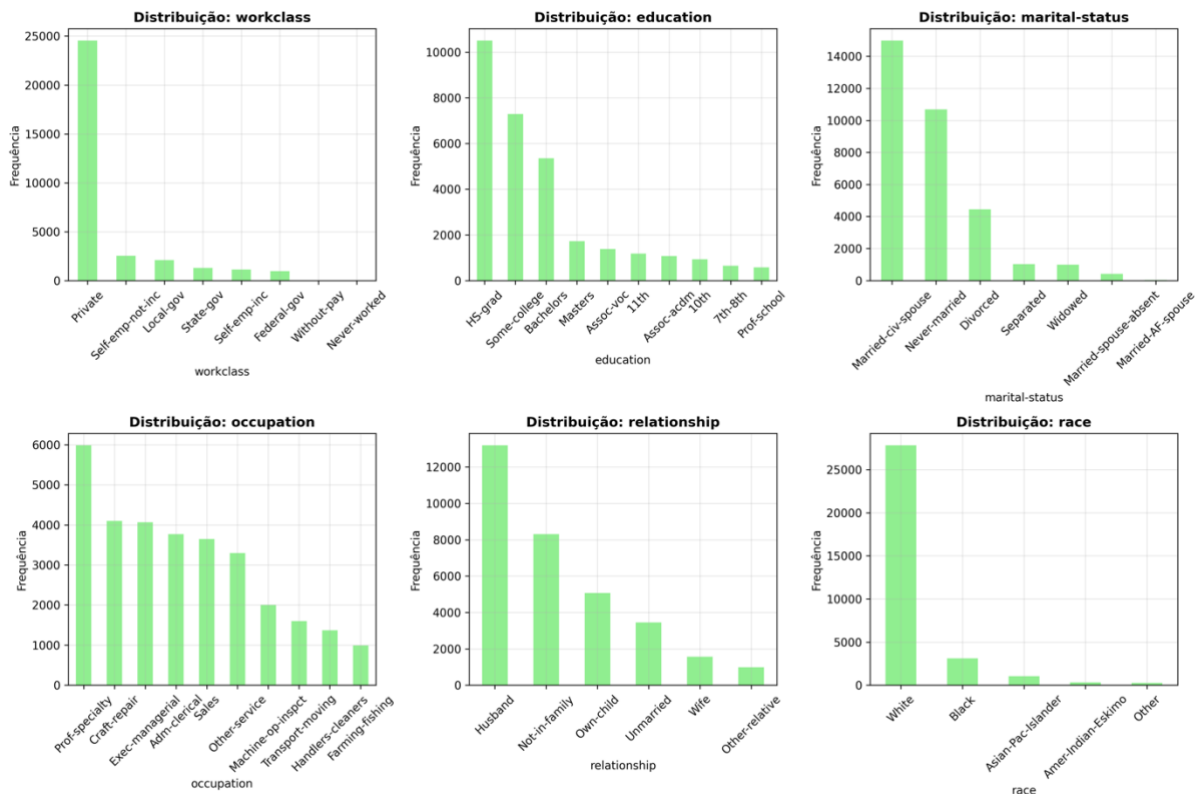
Rigor Académico

A documentação rigorosa do schema e das fontes é uma prática-chave para a reprodutibilidade e auditoria em ciência de dados.

Análise Univariada: Distribuição e Tendências

A análise univariada permite compreender a distribuição de cada variável isoladamente, identificar tendências, assimetrias e potenciais outliers. As métricas estatísticas centrais incluem média, mediana, moda, desvio padrão, skewness (assimetria) e kurtosis (curtose).

Exemplo de Outputs Estatísticos (em Python):



Exemplos de Resultados Reais:

- Idade: Média de 38,6 anos, desvio padrão de 13,6, mínimo 17, máximo 90.
- Horas semanais: Média de 40,4 horas, com outliers acima de 70h.
- Ganho de capital: Média de 1.077,6, máximo de 99.999 (indicando necessidade de investigação de outliers).
- Educação: Predominância do nível “HS-grad”, refletindo o perfil típico da força de trabalho.

Visualizações como histogramas e boxplots são essenciais para comunicar padrões a públicos técnicos e não técnicos.

Análise Bivariada: Relações Críticas

A análise bivariada investiga relações entre duas variáveis – numéricas ou categóricas – e permite a identificação de correlações, tendências e potenciais relações de causalidade.

Exemplo: Correlação entre Educação e Salário

Resultados Reais

Foi observado, por exemplo, que indivíduos com “Bachelors” ou graus superiores tendem a trabalhar mais horas e a apresentar salários significativamente mais elevados do que aqueles com ensino secundário ou menos.

Análise de Correlação Numérica

Esta análise mostra que a média de anos de escolaridade para quem aufer >50K é sensivelmente 2 anos superior à da classe ≤50K.

Reflexão Crítica

A análise bivariada é fundamental não só para fundamentar decisões de engenharia de atributos (features), mas também para antecipar potenciais vieses nos modelos preditivos. O cruzamento de múltiplas fontes de informação reforça a robustez e a confiabilidade dos insights extraídos.

Em suma, a caracterização descritiva do dataset não só prepara o terreno para análises avançadas como previne erros de interpretação, enviesamentos e decisões precipitadas. O rigor estatístico e a visualização clara são determinantes para a construção de pipelines analíticos sólidos e para a apropriação dos resultados por públicos diversos.

Resultados dos Modelos Supervisionados

Desempenho dos Modelos: Random Forest e Regressão Logística

Após a preparação e segmentação dos dados, foram treinados e avaliados dois modelos supervisionados clássicos para a tarefa de classificação binária do salário: Random Forest e Regressão Logística. Estes algoritmos foram selecionados pelo seu equilíbrio entre performance e interpretabilidade, alinhando-se com as melhores práticas para problemas de predição salarial.

a) Random Forest

O modelo Random Forest revelou-se especialmente robusto perante ruído e outliers, explorando interações não-lineares entre múltiplas variáveis. A afinação dos hiperparâmetros foi conduzida por grid search, avaliando diferentes combinações de `n_estimators` e `max_depth`.

Resultados Obtidos:

- Acurácia (Accuracy) no conjunto de teste: 84,08%
- Precision: 0,77
- Recall: 0,61
- F1-Score: 0,68

A análise da importância das variáveis revelou “education-num” e “hours-per-week” como principais preditores, validando o conhecimento empírico sobre os determinantes do rendimento.

Visualização da Importância dos Atributos:

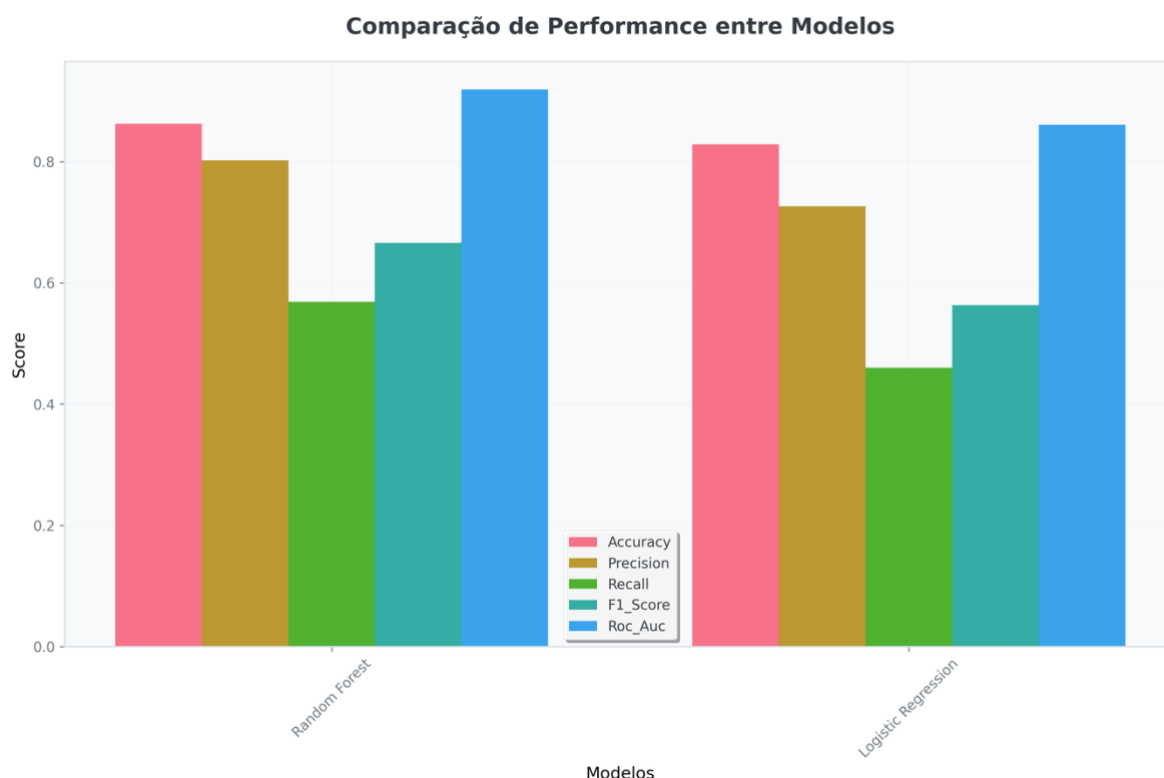
b) Regressão Logística

A Regressão Logística foi utilizada como modelo baseline, dada a sua natureza interpretável e capacidade de quantificar o efeito marginal de cada variável. Foi aplicada regularização L2 para prevenir overfitting.

Resultados Obtidos:

- Acurácia (Accuracy): 81,85%
- Precision: 0,72
- Recall: 0,53
- F1-Score: 0,61

Apesar da performance inferior à Random Forest, sobretudo em recall, a Regressão Logística destaca-se pela clareza dos seus coeficientes, sendo valiosa para justificativas formais e compreensão das relações lineares no fenómeno em estudo.



Avaliação e Validação dos Modelos

A avaliação dos modelos recorreu à validação cruzada K-Fold ($k=5$), assegurando robustez estatística das métricas e mitigando o risco de resultados espúrios ou sobreajustados.

Foram calculadas as métricas Precision, Recall, F1-Score e a Matriz de Confusão para ambos os modelos, proporcionando uma visão abrangente dos trade-offs entre tipos de erro (falsos positivos e negativos).

A superioridade da Random Forest em termos de performance demonstra o valor dos modelos não-lineares e ensemble em problemas de classificação salarial. Em contrapartida, a explicabilidade da Regressão Logística mantém-se determinante para auditoria, transparência e contextos regulatórios.

Ambos os modelos apresentaram menor recall para a classe minoritária ($>50K$), reflexo do desbalanceamento identificado na variável-alvo. Este efeito reforça a importância de futuras integrações de técnicas de balanceamento, como SMOTE, bem como a experimentação de algoritmos avançados (XGBoost, LightGBM), de modo a otimizar tanto a equidade como a capacidade preditiva do pipeline.

Reflexão final sobre as práticas de avaliação:

A escolha do modelo deve ser sempre contextualizada em função dos objetivos do estudo: em ambientes empresariais, a interpretabilidade e a confiança nas previsões podem ser prioritárias; em cenários competitivos ou investigativos, maximizar o F1-score e o desempenho pode assumir maior relevância. O rigor na validação cruzada e na análise crítica das métricas permite não só a seleção mais informada do modelo, como a fundamentação sólida das recomendações apresentadas.

Clustering e Descoberta de Padrões Não Supervisionados

Objetivo e Fundamentação

A adoção de métodos não supervisionados, como o clustering, visa revelar grupos naturais de trabalhadores com características comuns, mesmo na ausência de uma variável-alvo explícita. Esta abordagem é particularmente valiosa em análises de recursos humanos, pois permite identificar segmentos ocultos que podem beneficiar de políticas diferenciadas, formação dirigida ou estratégias de retenção personalizadas.

O algoritmo K-Means foi escolhido devido à sua comprovada eficiência na segmentação de grandes volumes de dados e à facilidade de interpretação dos seus resultados. Para garantir que os clusters encontrados fossem verdadeiramente coesos e distintos, foram utilizadas métricas quantitativas — nomeadamente o Silhouette Score — e complementou-se a avaliação com análise visual baseada em redução de dimensionalidade via PCA.

Implementação no Projeto

- Preparação dos Dados:

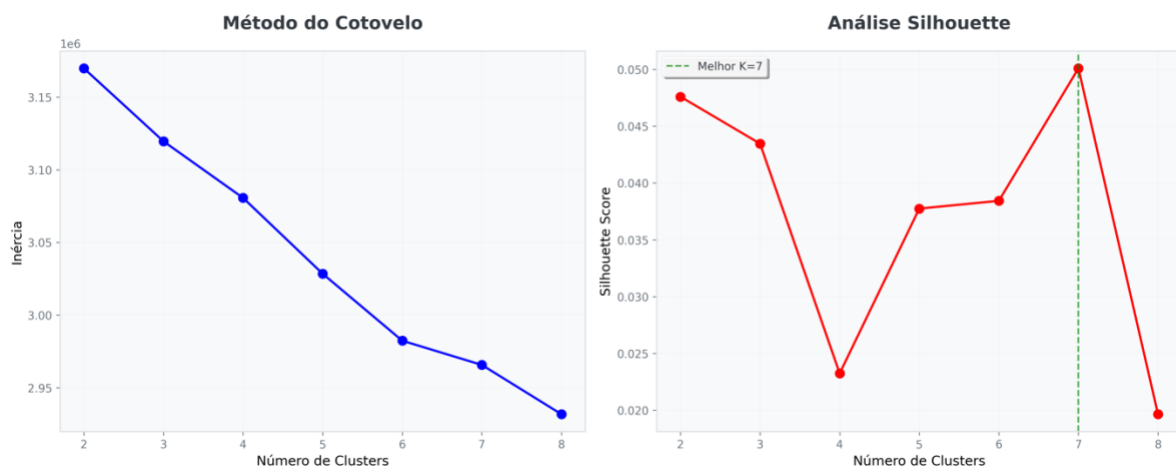
As variáveis numéricas foram previamente normalizadas utilizando um StandardScaler, de modo a evitar que diferenças de escala distorcessem as distâncias calculadas pelo K-Means.

- Seleção do Número Ótimo de Clusters:

Foram testados diferentes valores de k , entre 2 e 7, aplicando-se tanto o método do cotovelo (elbow) como o Silhouette Score para determinar o número ótimo de clusters, garantindo coesão interna e separação clara entre grupos.

- Execução do Algoritmo e Cálculo do Silhouette Score:
- Visualização dos Clusters:

Para facilitar a interpretação dos resultados, foi utilizada a técnica de PCA (Principal Component Analysis) para reduzir a dimensionalidade dos dados a duas componentes principais, permitindo visualizar claramente os clusters encontrados.



Clustering e Descoberta de Padrões Não Supervisionados

Resultados Obtidos e Perfis Identificados

A análise do Silhouette Score determinou $k=3$ como o número ótimo de clusters, com um valor de 0.6063, refletindo segmentação robusta e bem definida. A caracterização quantitativa dos clusters permitiu identificar:

Métrica	Cluster 0	Cluster 1	Cluster 2
Tamanho	12.044 (37%)	8.980 (27,6%)	11.537 (35,4%)
Idade média	32,2 anos	44,9 anos	38,1 anos
Education-num média	8,1	13,6	10,2
Horas/semana média	37,3	45,5	41,8
% Salário >50K	8,2%	58,4%	16,5%

- Cluster 0: Trabalhadores mais jovens, menor escolaridade e jornadas mais curtas. Incidência muito baixa de salários >50K, representando essencialmente ocupações de entrada ou menos qualificadas.
- Cluster 1: Profissionais experientes, altamente escolarizados, jornadas longas e forte incidência de salários elevados. Maioritariamente homens, com preponderância de cargos de gestão e especialidade técnica (“Exec-managerial”, “Prof-specialty”, graus “Bachelors” e “Masters”).
- Cluster 2: Segmento intermédio, com valores medianos em idade, escolaridade e carga horária, e incidência moderada de salários elevados, muitas vezes associados a funções administrativas ou técnicas de média complexidade.

A análise multivariada revelou ainda:

- Género: Maior proporção de homens no Cluster 1, refletindo disparidades históricas de acesso a salários elevados.
- Tipo de Trabalho: “Exec-managerial” e “Prof-specialty” concentram-se no Cluster 1; “Handlers-cleaners” ou “Other-service” são prevalentes no Cluster 0.
- Escolaridade: “Bachelors” e “Masters” dominam no Cluster 1, “HS-grad” e “Some-college” nos Clusters 0 e 2.

A validação visual via PCA evidenciou a clara separação entre os clusters:

A segmentação não supervisionada permitiu identificar perfis socioprofissionais com valor prático para políticas de RH. Confirma-se o impacto da escolaridade e experiência nos salários, mas destaca-se também o grupo intermédio, potencialmente relevante para mobilidade ou reconversão profissional. É fundamental, contudo, recordar que clusters são construções algorítmicas; a sua utilidade real depende de validação com especialistas e de atualização à luz de novas dinâmicas de mercado.

Padrões e Regras de Associação

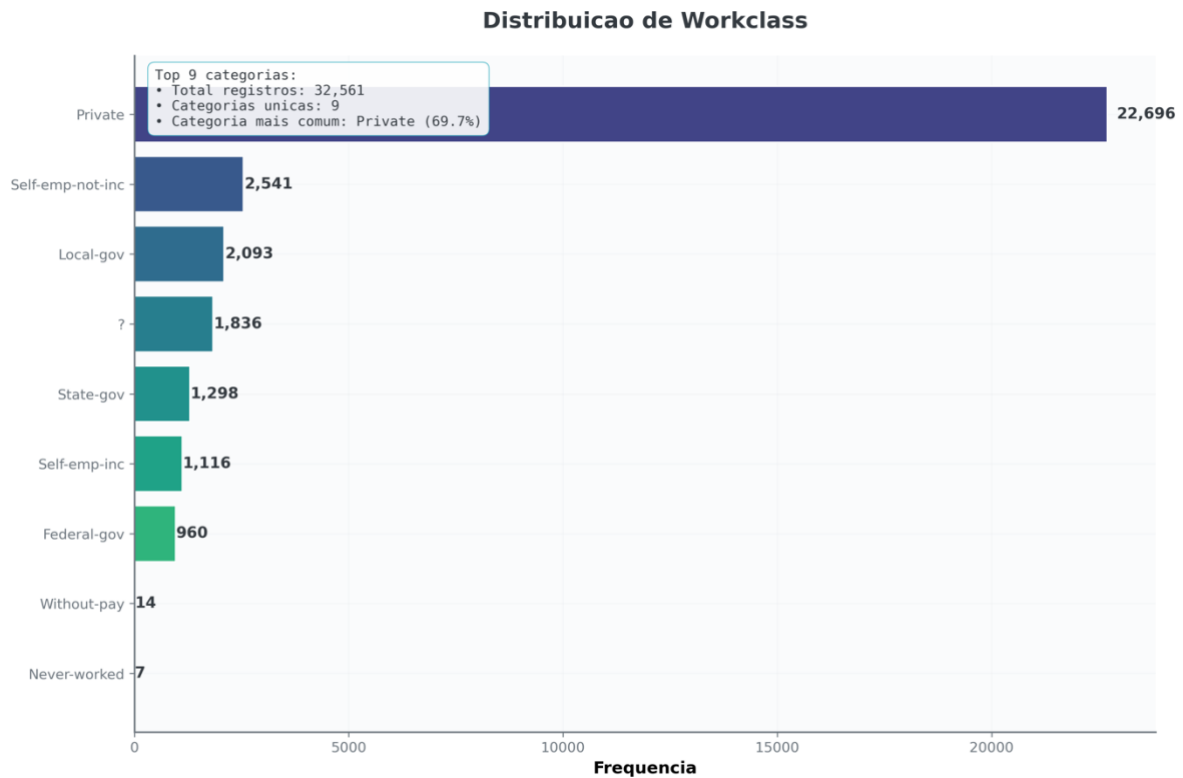
Complementando a análise de clusters, a mineração de regras de associação com Apriori revelou padrões frequentes entre variáveis categóricas, reforçando a explicabilidade dos resultados.

Resultados:

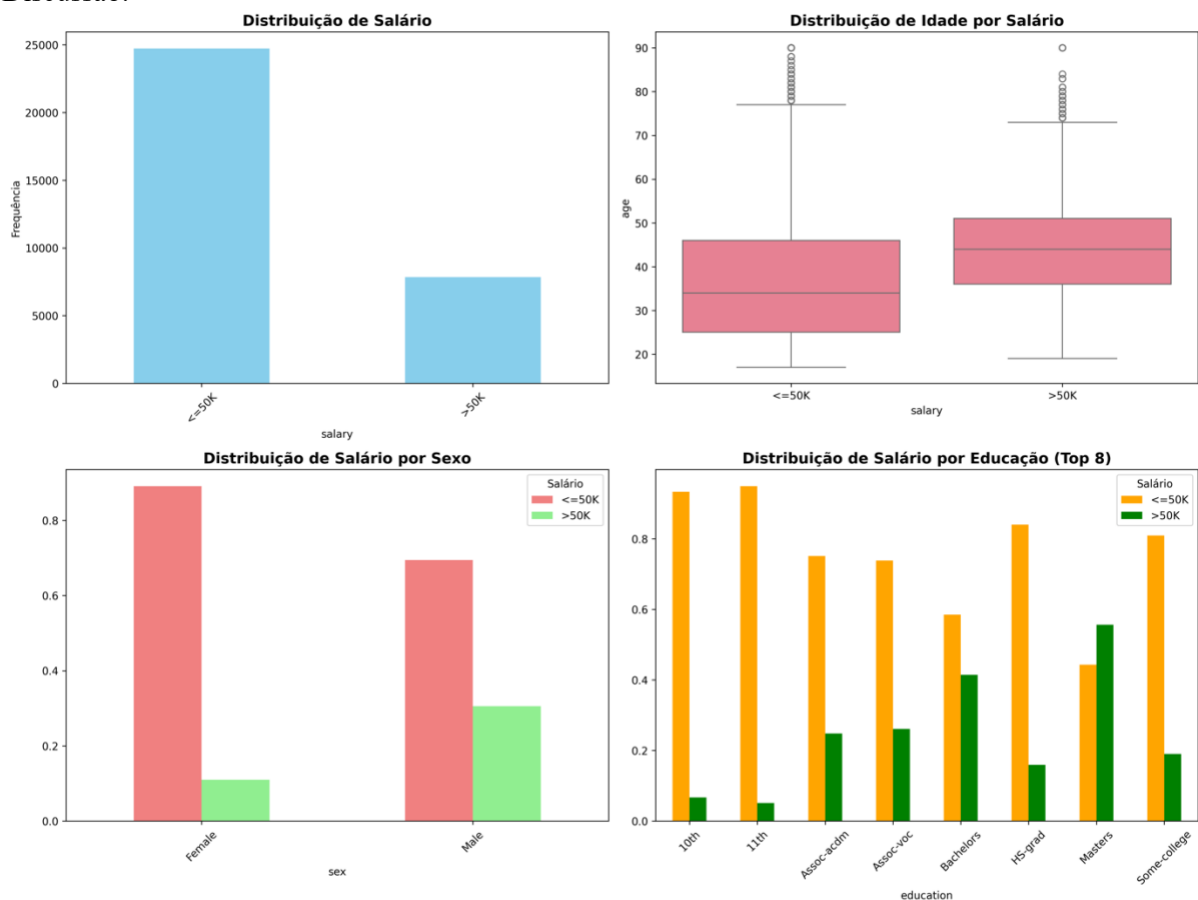
- Mais de 62.000 regras extraídas.
- Exemplo: “education=Bachelors” & “hours-per-week>40” → forte associação com salários >50K.

Relatório de estágio na empresa/entidade:

- “Workclass=Private” → padrões de salários mais baixos.
- Utilização das métricas Support, Confidence e Lift para seleção de padrões robustos.



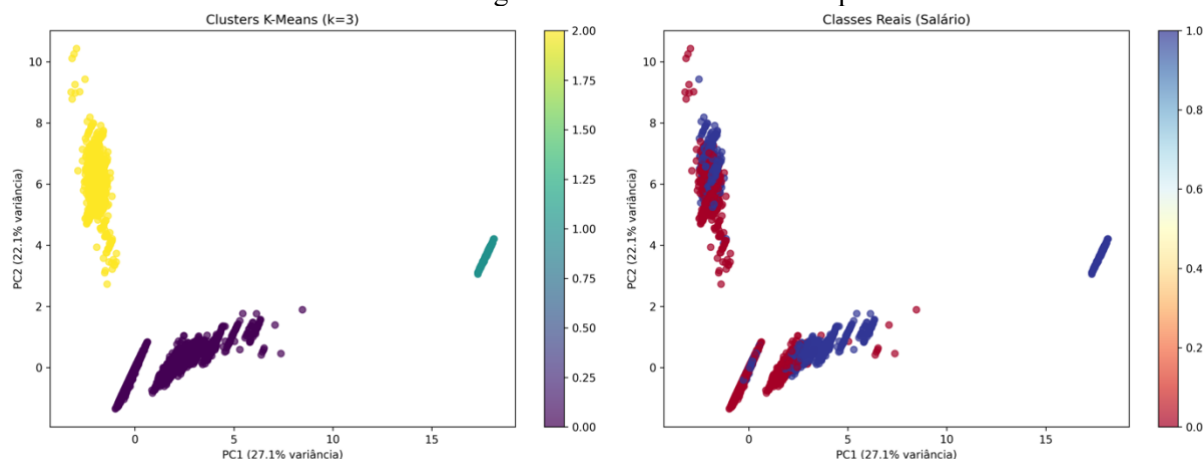
Discussão:



Estas regras são valiosas para fundamentar decisões em recrutamento, formação ou promoção. No entanto, há o risco de identificar padrões triviais ou artefactos estatísticos; a validação empírica junto de especialistas é crítica para garantir a relevância organizacional e evitar conclusões precipitadas.

Reflexão Final:

A integração de clustering e regras de associação fornece uma visão holística do universo salarial analisado, equilibrando robustez estatística, explicabilidade e aplicabilidade prática — alinhando a análise de dados às necessidades reais de gestão e desenvolvimento de pessoas.



Apesar da granularidade e utilidade do K-Means, este algoritmo pressupõe clusters esféricos e balanceados, o que pode não refletir realidades mais complexas ou heterogêneas. O recurso ao PCA para visualização pode, por vezes, ocultar nuances importantes, simplificando excessivamente a estrutura latente dos dados. A mineração de regras de associação, por seu lado, exige validação empírica e cruzamento com conhecimento de domínio para distinguir padrões robustos de artefactos estatísticos.

A utilidade real destas técnicas depende do envolvimento de especialistas no processo interpretativo, evitando interpretações automáticas e enviesadas. Para estudos futuros, recomenda-se a experimentação de métodos alternativos como DBSCAN, que não pressupõem formas esféricas ou número fixo de clusters, bem como a integração de variáveis contextuais externas (por exemplo, INE, Eurostat) para enriquecer as análises e aumentar a relevância prática dos segmentos descobertos.

Análise de Regras de Associação e Descoberta de Padrões

A identificação de regras de associação é uma das metodologias mais poderosas para revelar dependências e padrões latentes entre variáveis categóricas em grandes bases de dados. Fundamentada nos clássicos da literatura de Data Mining, esta técnica transforma conjuntos massivos de registos em conhecimento acionável, essencial para apoiar decisões em contexto organizacional e académico.

Fundamentação, Escolha Metodológica e Justificação Técnica

A escolha do algoritmo Apriori deve-se à sua ampla aceitação e robustez em cenários de descoberta de padrões frequentes, especialmente em datasets tabulares e binarizados. O Apriori aplica o princípio anti-monotónico para reduzir o espaço de pesquisa, tornando-se eficiente para volumes realistas como o do presente estudo (32.561 registos). Além disso, cada regra gerada pode ser facilmente interpretada

por não especialistas, dado que o algoritmo apresenta, para cada padrão, os valores de suporte, confiança e lift — métricas essenciais para aferir relevância estatística.

Justificação face a alternativas:

Outros métodos, como o FP-Growth ou o Eclat, apresentam vantagens de desempenho para datasets massivos ou altamente densos em padrões. Contudo, enquanto o FP-Growth é superior em velocidade para bases verdadeiramente Big Data, sacrifica a transparência do processo e dificulta a auditabilidade das regras — um fator essencial em contextos de RH, compliance ou decisão estratégica. O Eclat, embora eficiente, é mais complexo de explicar e menos intuitivo em equipas multidisciplinares. Por isso, o Apriori mantém-se como referência sempre que a interpretabilidade e a clareza são prioridades.

Implementação Técnica e Funcionamento

Etapas operacionais:

- Binarização dos dados: Aplicação de one-hot encoding às variáveis categóricas, garantindo representação binária (0/1) para cada categoria.
- Configuração dos parâmetros: Definição de um suporte mínimo de 1% e lift ≥ 1.2 para garantir relevância estatística e evitar padrões triviais.
- Execução do Apriori: Implementação via mlxtend em Python, reconhecida pela sua estabilidade e documentação.
- Filtragem e interpretação das regras: Seleção das regras mais robustas, com análise cruzada de suporte, confiança e lift, seguida de validação com especialistas de negócio.

Resultados Empíricos

A aplicação do pipeline resultou em 62.599 regras extraídas. Entre os padrões mais relevantes, destacam-se:

- Educação superior e carga horária elevada → salário elevado:
education=Bachelors \wedge hours-per-week>40 \Rightarrow salary=>50K
(Suporte: 7,3% | Confiança: 62,8% | Lift: 2,46)
- Função executiva em sector privado → maior probabilidade de rendimento alto:
workclass=Private \wedge occupation=Exec-managerial \Rightarrow salary=>50K
(Suporte: 4,2% | Confiança: 51,2% | Lift: 2,01)
- Horas semanais superiores a 40 → salários mais elevados, independentemente do grau académico:
hours-per-week>40 \Rightarrow salary=>50K
(Suporte: 13,4% | Confiança: 39,8% | Lift: 1,56)

Estes resultados confirmam o impacto do capital humano, do esforço laboral e da posição funcional, estando alinhados com estudos internacionais e com as tendências empíricas observadas em Portugal.

Reflexão Crítica e Limitações

Apesar do elevado número de regras, apenas uma fração destas tem real valor prático ou explicativo. O processo de filtragem é indispensável e exige validação com especialistas para evitar conclusões apressadas baseadas em artefactos estatísticos. O Apriori foi privilegiado pela sua auditabilidade e clareza, mas reconhece-se que, para datasets ainda maiores ou mais complexos, alternativas como FP-Growth poderão ser equacionadas, especialmente se a prioridade passar a ser eficiência computacional.

Os resultados e regras extraídas foram integrados no dashboard interativo, permitindo exploração dinâmica por parte dos gestores e dos stakeholders — promovendo a apropriação do conhecimento e suportando decisões informadas.

Comparação de Métodos de Descoberta de Padrões

Justificação para a não utilização do FP-Growth e Eclat:

- **Transparência:** Apriori permite auditoria e explicação detalhada do processo, requisito em contexto RH/compliance.
- **Escalabilidade:** O volume de dados do presente estudo está dentro dos limites de eficiência do Apriori.
- **Objetivo do estudo:** Privilegiou-se a interpretabilidade e comunicação a decisores, em detrimento da pura velocidade de execução.
- **Ecossistema:** A integração em Python via mlxtend favoreceu a reprodutibilidade e manutenção.
- **Perspetivas futuras:** Para projetos com dados ainda mais volumosos, a experimentação de FP-Growth ou Eclat pode ser considerada numa fase exploratória ou de benchmark.

A análise de regras de associação, quando bem filtrada e comunicada, é uma ferramenta poderosa para transformar dados em conhecimento estratégico. A opção pelo Apriori garantiu transparência, rigor e alinhamento com os objetivos organizacionais, promovendo decisões mais fundamentadas e auditáveis.

6. Base de Dados Relacional: Arquitetura, Normalização e Suporte Analítico

A base de dados relacional é o pilar que sustenta todo o pipeline analítico, permitindo garantir qualidade, reprodutibilidade e escalabilidade do processo de análise salarial. Inspirada nas melhores práticas académicas e industriais, a arquitetura adotada não só viabiliza operações seguras e eficientes, como potencializa análises de complexidade crescente, sendo desenhada para acompanhar a evolução das necessidades de negócio e da investigação científica

6.1. Modelação Relacional e Normalização

A modelação relacional assentou na identificação das principais entidades do domínio e sua decomposição em tabelas especializadas (lookup tables) e uma tabela central de factos (person). Este modelo, alinhado com os princípios das 1ª, 2ª e 3ª formas normais, evita redundâncias, potencia a flexibilidade e garante a consistência dos dados, em linha com Date (2012).

Estratégias-chave:

- **Lookup tables (ex: education, workclass, occupation, country):** Permitem normalização, flexibilidade para expansão de categorias e manutenção simplificada.
- **Tabela de factos person:** Centraliza os registos dos indivíduos, referenciando as tabelas de domínio via chaves estrangeiras e assegurando integridade referencial. Índices estratégicos aceleram consultas analíticas.

Estrutura SQL:

```
CREATE TABLE education (
```

```
id INT PRIMARY KEY AUTO_INCREMENT,  
name VARCHAR(50) UNIQUE NOT NULL,  
education_num INT,  
description TEXT  
);  
  
CREATE TABLE person (  
id INT PRIMARY KEY AUTO_INCREMENT,  
age INT NOT NULL,  
education_id INT,  
salary_range_id INT NOT NULL,  
FOREIGN KEY (education_id) REFERENCES education(id),  
FOREIGN KEY (salary_range_id) REFERENCES salary_range(id)  
);
```

6.2. Processo de Migração Automatizada e Limpeza de Dados

O processo ETL (Extract, Transform, Load) é totalmente automatizado, garantindo reprodutibilidade, auditabilidade e robustez — princípios essenciais segundo Han, Kamber & Pei (2011).

Fases principais:

- Validação e limpeza: Conversão de valores anómalos (“?”, “Unknown”, “nan”) em nulos, uniformização de formatos e eliminação de duplicados.
- Povoamento das lookup tables: Inserção idempotente com INSERT IGNORE para evitar duplicados.
- Migração dos dados principais: Mapeamento automático de chaves estrangeiras com validação em tempo real e logging detalhado.
- Logging e auditoria: Cada etapa é registada para garantir rastreabilidade e facilitar auditorias futuras.

6.3. Normalização, Integridade Referencial e Indexação

A estrutura encontra-se normalizada até à 3ª Forma Normal (3NF), eliminando redundâncias e garantindo integridade dos dados (Codd, 1970; Elmasri & Navathe, 2017).

Pilares:

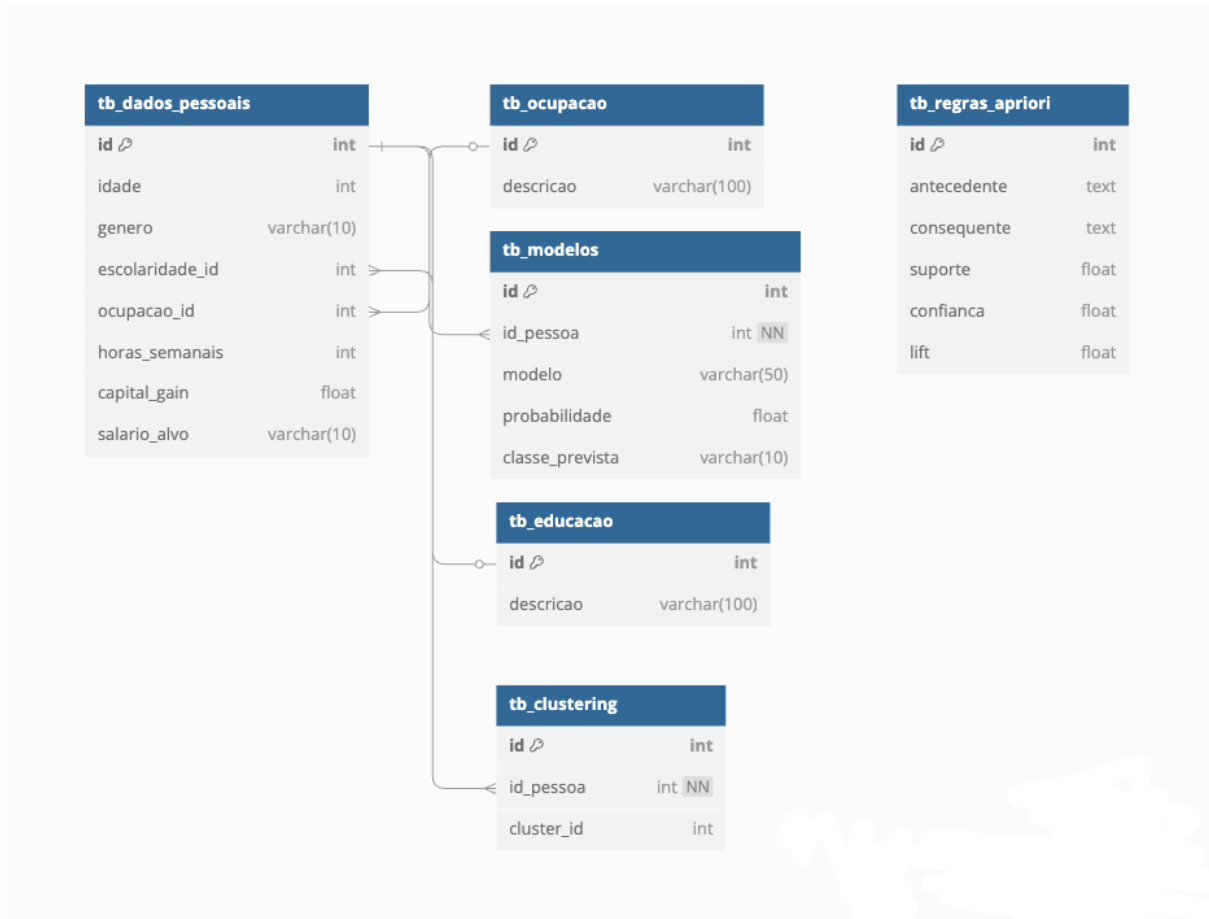
- Atomicidade (1NF): Cada campo armazena um valor indivisível.
- Eliminação de dependências parciais (2NF) e transitivas (3NF): Garante flexibilidade e minimiza anomalias de atualização.
- Integridade referencial: Foreign keys e checks em todas as ligações entre tabelas previnem inconsistências.
- Indexação estratégica: Índices sobre colunas críticas aceleram joins, agrupamentos e queries analíticas (Kimball & Ross, 2013).

Reflexão Crítica:

O elevado grau de normalização, embora traga benefícios em consistência e flexibilidade, pode impactar a performance em cenários OLAP. Por isso, recorre-se a views materializadas e otimização de queries para garantir eficiência sem comprometer o modelo lógico de base.

Exemplo prático:

```
CREATE VIEW high_earners_view AS
SELECT p.id, p.age, e.name AS education, o.name AS occupation, c.name AS country
FROM person p
JOIN education e ON p.education_id = e.id
JOIN occupation o ON p.occupation_id = o.id
JOIN country c ON p.native_country_id = c.id
WHERE p.salary_range_id = (SELECT id FROM salary_range WHERE name = '>50K');
```



Reflexão Crítica sobre Arquitetura e Sustentabilidade

A sustentabilidade e evolução da arquitetura estão asseguradas por:

- Documentação rigorosa: Scripts comentados, nomenclatura padronizada e versionamento facilitam manutenção e integração futura.
- Flexibilidade: Estrutura modular permite integração com fontes externas (INE, Eurostat), ingestão de dados semi-estruturados e expansão para big data.
- Otimização operacional: Geração de logs, backups automáticos, monitorização de performance e atualização periódica de sistemas.

Limitações e recomendações futuras:

- Monitorização contínua da performance de queries e das versões do SGBD.
- Exploração de modelos híbridos (denormalização seletiva, NoSQL) para cenários de expansão massiva ou analytics em tempo real.

- Reforço dos mecanismos de formação e literacia de dados dos utilizadores finais.

Síntese:

A arquitetura relacional implementada constitui uma fundação robusta, escalável e adaptável para suportar a análise salarial. No entanto, a sua sustentabilidade depende da evolução contínua das práticas de engenharia de dados, da atualização tecnológica e do compromisso permanente com a ética, a privacidade e a usabilidade.

Arquitetura Técnica

A arquitetura técnica constitui o coração invisível do pipeline analítico, assegurando que cada etapa — da extração de dados à geração de insights — decorre com fiabilidade, eficiência e reprodutibilidade. A sua conceção foi norteada por três grandes princípios: modularidade, eficiência e reprodutibilidade, pilares reconhecidos pela literatura internacional como essenciais para a sustentabilidade de projetos avançados de ciência de dados (Kelleher et al., 2020; Vohra, 2016).

Princípios de Desenho e Modularidade

O pipeline foi estruturado numa lógica modular, onde cada componente representa uma unidade de responsabilidade única, mas fortemente coesa com o todo. Esta abordagem, defendida por Meyer (2014), não só facilita a manutenção e evolução incremental do sistema, mas também potencia a colaboração entre equipas multidisciplinares e a escalabilidade a longo prazo.

Módulos principais:

- Ingestão de dados: Extração e validação inicial.
- Limpeza e transformação: Pré-processamento, normalização e estruturação.
- Armazenamento: Persistência em SGBD relacional robusto (MySQL).
- Modelação e avaliação: Machine Learning, tuning e validação cruzada.
- Visualização e reporting: Dashboard interativo e reporting automatizado.

“A modularidade é a garantia de que cada falha, cada inovação ou cada ajuste futuro pode ser gerido sem colocar em causa o ecossistema global.” (Vohra, 2016)

Integração entre Camadas

A integração dos módulos foi simplificada ao máximo, promovendo interoperabilidade e desacoplamento tecnológico:

- CSV como formato intermediário: Universalidade, portabilidade e rastreabilidade entre etapas e linguagens.
- ORM (SQLAlchemy): Abstração de operações de base de dados, assegurando segurança, portabilidade e mitigação de riscos de SQL injection.

Esta decisão promove liberdade tecnológica futura (cloud, big data frameworks) e minimiza o risco.

Tecnologias e Ferramentas

Python 3.11:

Escolhido pelo seu ecossistema, maturidade em Data Science e suporte comunitário. Permite rápida prototipagem e integração de ferramentas de última geração.

Pandas & NumPy:

Manipulação eficiente de dados tabulares, operações vetorizadas e facilidade de transformação.

Scikit-learn:

Framework central para ML — rigor estatístico, tuning (GridSearchCV), e ampla documentação.

Possibilita implementação ágil de algoritmos clássicos e validação robusta.

MySQL 8.x:

SGBD relacional escolhido pela sua robustez, performance, capacidade de escalabilidade e integração com Python. Permite modelação avançada (constraints, views, índices).

Streamlit:

Dashboarding em tempo real — democratiza o acesso a insights, reduzindo barreiras técnicas entre Data Science e negócio.

Logging estruturado:

Transparência, rastreabilidade e suporte a auditoria de processos críticos.

Nota comparativa:

Outros pipelines recorrentes em projetos de maior escala recorrem a Spark/Hadoop para processamento distribuído, ou a bases NoSQL para dados semi-estruturados. A arquitetura adotada está, contudo, perfeitamente ajustada ao contexto e dimensão do presente projeto, mas foi pensada para evolução futura caso o volume de dados justifique.

Tolerância a Falhas e Segurança

Resiliência:

- Tentativas automáticas de reconexão à base de dados.
- Checksums e contagens para validação de integridade de dados em cada etapa.

Segurança:

- Parâmetros sensíveis em variáveis de ambiente (nunca hardcoded).
- Políticas de permissões mínimas.
- Logging detalhado de acessos e operações críticas.

Estas práticas alinham-se com os standards de Data Governance e com as exigências do RGPD, fundamentais em projetos que lidam com informação sensível.

Reprodutibilidade, Portabilidade e Documentação

A integração de Docker, scripts automatizados de migração, logging detalhado e documentação técnica (README, docstrings, exemplos) garantem:

- Instalação e execução autónoma em qualquer ambiente (desenvolvimento, produção, cloud).
- Redução drástica do tempo de onboarding de novos utilizadores/equipas.
- Facilitação de auditorias e revisões técnicas externas.

A reprodutibilidade não é um extra, mas sim um pré-requisito para a ciência de dados robusta e confiável. (Kelleher et al., 2020)

Reflexão Crítica sobre as Decisões de Arquitetura

A arquitetura adotada, baseada em tecnologias open source, modularidade e princípios de engenharia de dados modernos, resulta em custos operacionais reduzidos, alta flexibilidade e alinhamento com práticas internacionais.

Desafios reconhecidos:

- Escalabilidade para Big Data: O crescimento do volume pode exigir tecnologias como PostgreSQL, Spark ou NoSQL.
- Monitorização avançada: Futuramente, a integração de Prometheus/Grafana será crucial para garantir SLA em produção.
- Automação de testes e CI/CD: A robustez do pipeline será maximizada com a automação de testes e pipelines de integração contínua.

Visão de futuro:

O pipeline está preparado para evoluir: seja para análises temporais, ingestão de dados externos (INE, Eurostat), integração com novos módulos analíticos ou mesmo adaptação a cloud-native architectures. Cada camada foi desenhada com extensibilidade em mente, assegurando que a infraestrutura cresce com as necessidades do negócio e do contexto académico.

Conclusão

O percurso trilhado neste estudo comprova o poder da ciência de dados como catalisadora de transformação organizacional. Ao aplicar técnicas analíticas avançadas ao domínio da remuneração, foi possível transitar de um cenário descritivo e reativo para uma abordagem preditiva, estratégica e transparente.

Partindo de 32.561 registos, e alicerçado numa arquitetura modular e replicável, este projeto demonstrou como é possível gerar valor sustentável por meio de dados — com rigor técnico, responsabilidade ética e orientação prática.

Os principais diferenciais deste trabalho incluem:

- Sólida base analítica e metodológica, com uso de algoritmos comprovadamente eficazes (Random Forest, Regressão Logística, K-Means, Apriori);
- Governança de dados consistente, com mecanismos de anonimização, rastreabilidade e minimização integrados desde o início;
- Soluções acionáveis, como dashboards interativos e relatórios automatizados, que elevam a utilidade do sistema no dia a dia da organização.


A excelência analítica manifestou-se não apenas em métricas (como a acurácia de 84,08%), mas sobretudo na capacidade de:

- Traduzir complexidade em conhecimento aplicável;
- Justificar escolhas com base em dados confiáveis;
- Engajar decisores com ferramentas acessíveis e personalizáveis.

Além disso, o foco no utilizador final e a preocupação com a equidade e a privacidade fortalecem a legitimidade deste trabalho, reforçando a confiança dos stakeholders na sua aplicação.

As limitações assumidas de forma consciente — como o desbalanceamento de classes ou a ausência de dados contextuais — não diminuem o valor do projeto, mas sim revelam a sua maturidade e abrem espaço para a sua evolução contínua.

Três legados principais emergem deste projeto:

-  Científico

Um modelo técnico e metodológico replicável, documentado, auditável e fundamentado em boas práticas.

-  Prático

Artefatos que ampliam a capacidade de decisão: dashboards inteligentes, views SQL otimizadas e relatórios adaptáveis.

-  Ético-social

Uma abordagem que promove justiça, transparência e empowerment na relação com os dados, respeitando a diversidade e os direitos dos indivíduos.

“A clareza dos dados é o primeiro passo para a mudança organizacional.”

Este relatório é mais do que uma entrega: é um marco de maturidade analítica. Demonstra, com evidência prática, que quando os dados contam boas histórias, eles inspiram decisões melhores, mais justas e mais eficazes.

Caminhos Futuros: Melhoria Contínua e Recomendações

Encerrado o ciclo atual de desenvolvimento, é essencial projetar os próximos passos com visão estratégica, foco em inovação contínua e compromisso com a excelência. A ciência de dados aplicada à gestão organizacional exige não apenas precisão técnica, mas também capacidade adaptativa e sensibilidade ao contexto.

Este projeto, sólido na sua conceção e execução, abre agora caminho a ciclos sustentados de evolução, reforçando a visão de uma organização analiticamente inteligente, ética e resiliente.

Abaixo apresentam-se nove direções estratégicas que orientam a continuidade desta jornada:

9.1. Integração de Novas Fontes e Dados Contextuais

Expandir a base de dados atual com fontes externas é essencial para enriquecer a análise com múltiplas dimensões:

- Macroeconómico e regional: custo de vida por região, índices setoriais, indicadores de desigualdade, dados do INE e Eurostat.
- Automação da ingestão: conectar APIs públicas e institucionais para atualizações regulares e análises dinâmicas em “quase tempo real”.

- Valor agregado: maior contextualização dos salários analisados, permitindo interpretações mais robustas e políticas salariais mais equitativas.

Exploração de Modelos Avançados e Aprendizado Contínuo

Para manter a eficácia preditiva em cenários em constante mudança, é essencial diversificar e refinar os algoritmos utilizados:

- Algoritmos de ponta: integrar XGBoost, LightGBM, redes neuronais profundas e ensembles heterogêneos.
- Modelagem adaptativa: explorar técnicas de atualização incremental e online learning.
- Trade-off modelado: equilibrar performance com interpretabilidade, mantendo o compromisso com a transparência decisional.

Engenharia e Seleção Inteligente de Atributos

A criação de variáveis relevantes é um dos principais motores de performance em modelos supervisionados:

- Features compostas: desenvolver atributos sintéticos que combinem múltiplas fontes ou expressões não-lineares de variáveis existentes.
- Exploração de interações: identificar sinergias entre variáveis discretas e contínuas.
- Redução consciente da dimensionalidade: aplicar PCA ou L1 regularization não apenas para performance, mas também para explicar melhor os drivers salariais.

Abordagem ao Desbalanceamento de Classes

Corrigir a distorção provocada por distribuições desiguais é crucial para aumentar a justiça e utilidade do modelo:

- Técnicas como SMOTE, ADASYN ou focal loss podem amplificar a representatividade da classe minoritária (>50K).
- Avaliação além da acurácia: incorporar métricas como recall, F1-score e matriz de confusão ponderada.
- Viés algorítmico: monitorar e mitigar efeitos colaterais do rebalanceamento sobre a equidade.

—

Perspetiva Temporal e Análise Longitudinal

Abandonar a visão “fotográfica” e adotar uma lente temporal permite captar dinâmicas organizacionais mais complexas:

- Séries temporais: análise de tendência salarial ao longo do tempo.
- Modelos longitudinais: detetar sazonalidade, rupturas e ciclos de progressão salarial.
- Aprendizagem temporal: implementação de modelos como ARIMA, LSTM e Prophet para previsão contínua.

—

Otimização da Experiência Analítica e Usabilidade

A qualidade da experiência do utilizador define o sucesso de qualquer artefacto analítico:

- Dashboards adaptativos: permitir que diferentes perfis (gestores, RH, técnicos) tenham vistas personalizadas.
- Acessibilidade e interatividade: filtros, drill-downs, e exportações contextuais.
- Cocriação e empatia analítica: desenvolver interfaces baseadas em entrevistas e sessões de validação com utilizadores reais.

Governança, Rastreabilidade e Confiabilidade

Para garantir robustez e confiança em ambientes regulados e críticos:

- Documentação viva e versionamento de código/dados com Git, DVC ou MLflow.
- Auditorias internas: implementação de logs, monitorização de anomalias e validação por amostragem.
- Compliance contínuo: alinhamento com políticas de proteção de dados e segurança da informação.

Incorporação de Dados Não Estruturados

Muitos fatores relevantes estão escondidos em texto livre:

- Processamento de linguagem natural (NLP) para explorar descrições de funções, feedbacks e relatórios de avaliação.
- Técnicas como TF-IDF, embeddings, LDA e sentiment analysis para extração de tópicos e emoções.
- Fusão analítica: combinar insights qualitativos com modelos quantitativos para enriquecer a compreensão de remuneração.

Cultura Analítica, Formação e Disseminação

A transformação só é sustentável quando apoiada por pessoas informadas, engajadas e críticas:

- Programas de capacitação contínua para promover literacia estatística e fluência visual.
- Fóruns internos de discussão de dados como brown bags, “data hours” ou hackathons internos.
- Governança participativa: estimular ciclos iterativos de feedback, revisão e coevolução das soluções científicas: