Trabalho A – Análise Quantitativa com Machine Learning

Pós-Graduação em Inteligência Artificial Aplicada à Gestão  
ISCTE – 2025

# 1. Introdução

Este trabalho foi desenvolvido no âmbito do Trabalho A – Quantitativo da disciplina ADVIAG (ISCTE). O tema definido é: Machine Learning para análise preditiva no contexto empresarial.  
  
O caso prático foi construído com base na empresa Biagio Indústria, dedicada ao fabrico de auxiliares tecnológicos para panificação e pastelaria, em Angola, os dados que é possível extrair do ERP da empresa permitem que se abordem:  
  
Duas questões de investigação:   
i) Quanto vamos vender por período/cliente?   
ii) Quais são os principais drivers das vendas?  
  
Todo o trabalho foi desenvolvido em Python, com apoio da IA generativa (ChatGPT),criei um projeto chamado “Trabalho ISCET “ onde adicionei os ficheiros “proposta de trabalho” e ao emails associados ao trabalho, a IA (ChatGPT) foi usada para geração de código, documentação e estruturação do relatório, em conformidade com o código de conduta de utilização de IA. Os scripts foram implementados e executados em ambiente Visual Studio Code, recorrendo a ambiente virtual (venv). Os códigos usados neste trabalho encontram-se disponível em:

<https://github.com/CunhaVaz/iscte-ml.git> (pasta src)

# 2. Geração de Dados Sintéticos

Inspirados em dados reais da Biagio (clientes: Kéro, Shoprite, Intermarket entre outros; produtos: Pão de Água, Baguete, Croissant entre outros; canais: Grosso e Retalho), gerou-se um dataset com 500 linhas × 8 colunas.  
  
Foram incluídas anomalias propositadas: valores omissos em Vendas e Margem\_%, duplicados, outliers em Vendas (>10× média), valores inválidos em Margem\_% (>100) e erros ortográficos em clientes.  
  
Código: src/gerador\_dataset.py → gera data/raw/dataset\_biagio.xlsx.  
  
Prompt utilizado no ChatGPT:  
Quero gerar um dataset sintético em Python, com 500 linhas e 8 colunas, representando vendas da empresa Biagio. Deve conter: Ano, Mês, Cliente, Produto, Canal, Vendas, Margem\_% e Margem\_Valor. Inclui anomalias: omissos, duplicados, outliers e erros ortográficos. Exporta em Excel para data/raw/dataset\_biagio.xlsx.

# 3. Análise Exploratória Inicial (EDA Raw)

Primeira análise exploratória com estatísticas e Sweetviz, antes de qualquer limpeza. Principais achados: omissos, duplicados, outliers, margens fora de [0,100], erros ortográficos nos clientes.  
  
Outputs: eda\_raw\_outputs.xlsx, eda\_raw\_summary.txt, eda\_raw\_sweetviz.html  
  
Código: src/eda\_raw.py  
  
Prompt:  
Quero um script (src/eda\_raw.py) que leia data/raw/dataset\_biagio.xlsx, exporte estatísticas para Excel e TXT, e gere relatório Sweetviz em reports/eda\_raw\_sweetviz.html.

# 4. Limpeza dos Dados

Correção das anomalias: duplicados removidos, omissos imputados, outliers eliminados, margens normalizadas, nomes de clientes corrigidos.  
  
Resultado: dataset limpo data/processed/dataset\_biagio\_clean.xlsx  
  
Código: src/clean\_data.py  
  
Prompt:  
Escreve script (src/clean\_data.py) que limpe dataset\_biagio.xlsx: remova duplicados, trate omissos pela mediana, corrija nomes de clientes, normalize margens e elimine outliers. Grave em data/processed/dataset\_biagio\_clean.xlsx.

# 5. EDA Pós-Limpeza

O dataset limpo foi explorado novamente: histogramas, matriz de correlação, série temporal e top 5 clientes.  
  
Outputs: eda\_clean\_outputs.xlsx, eda\_sweetviz\_clean.html  
  
Código: src/eda\_clean.py  
  
Prompt:  
Cria script (src/eda\_clean.py) que analise data/processed/dataset\_biagio\_clean.xlsx, exporte estatísticas para Excel/TXT, gere gráficos (histograma, correlação, top clientes) e relatório Sweetviz.  
  
[Inserir Figura 1: Histograma das Vendas]  
[Inserir Figura 2: Matriz de Correlação]  
[Inserir Figura 3: Gráfico das Vendas Mensais]  
[Inserir Figura 4: Top 5 Clientes]

# 6. Treino/Teste 80/20

Divisão 80% treino / 20% teste. Modelos: Regressão Linear e Random Forest. Métricas: RMSE, MAPE, R².  
  
Código: src/model\_train.py  
  
Prompt:  
Cria script (src/model\_train.py) que divida dataset em treino/teste (80/20), treine Regressão Linear e Random Forest, e calcule RMSE, MAPE, R² (treino, teste, global). Exporte para reports/model\_results.xlsx.

# 7. Resultados Principais

Regressão Linear  
- Treino: RMSE = 6 915 | MAPE = 26,0% | R² = 0,74  
- Teste: RMSE = 7 924 | MAPE = 23,3% | R² = 0,53  
- Global: RMSE = 7 129 | MAPE = 25,5% | R² = 0,70  
  
Random Forest  
- Treino: RMSE = 855 | MAPE = 2,4% | R² = 0,996  
- Teste: RMSE = 3 051 | MAPE = 7,4% | R² = 0,93  
- Global: RMSE = 1 568 | MAPE = 3,4% | R² = 0,99  
  
Interpretação: Random Forest foi o modelo vencedor (R² elevado, baixo RMSE, MAPE < 8%).  
  
[Inserir Figura 5: Tabela Comparativa das Métricas]

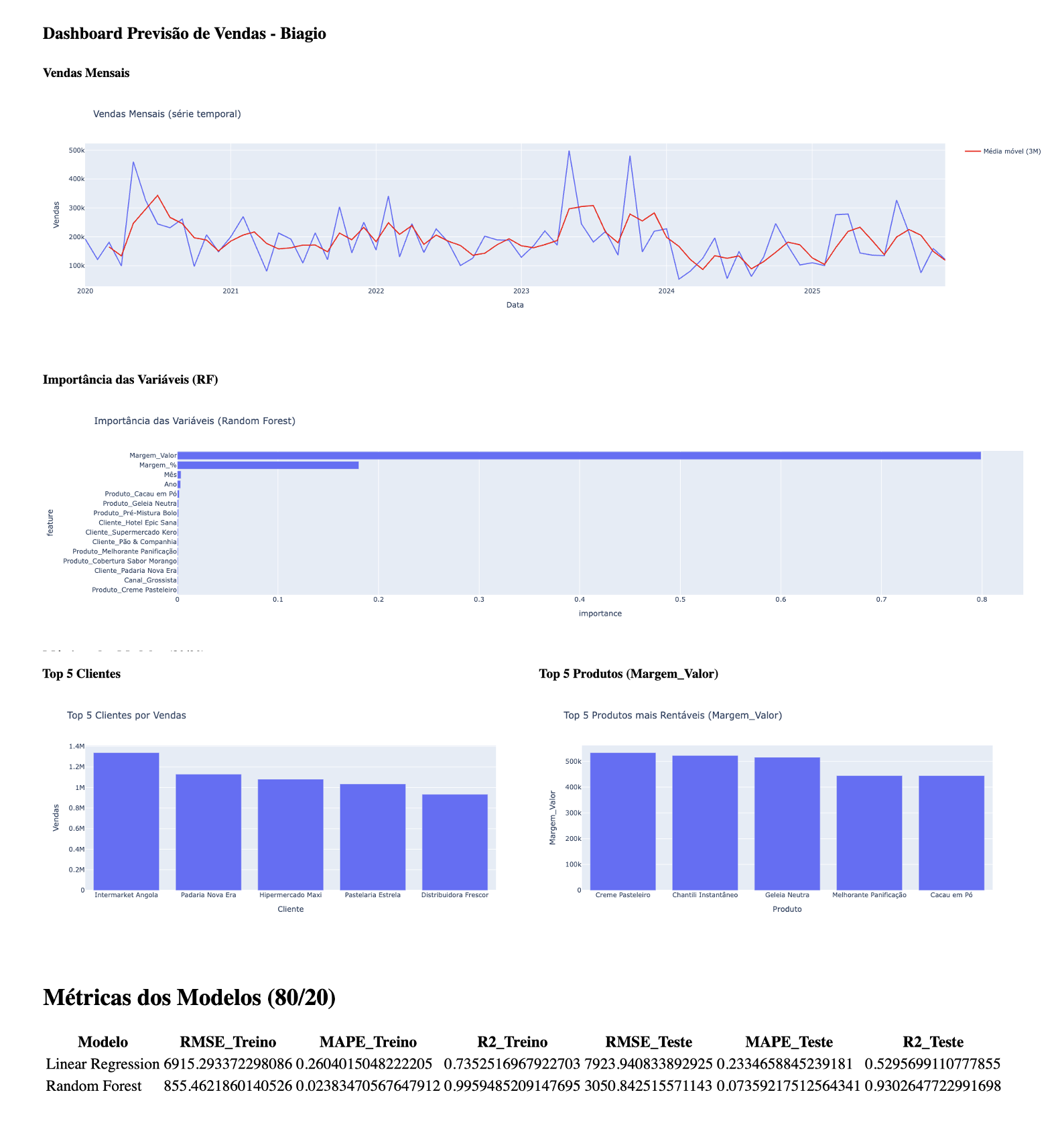
# 8. Visualização das Métricas

Gráficos comparativos (PNG) das métricas RMSE, MAPE e R².  
  
Código: src/plot\_metrics.py  
  
Prompt:  
Escreve script (src/plot\_metrics.py) que leia reports/model\_results.xlsx e gere gráficos de barras (RMSE, MAPE, R² – treino/teste e global) em PNG.  
  
[Inserir Figura 6: Gráfico RMSE Treino/Teste]  
[Inserir Figura 7: Gráfico MAPE Treino/Teste]  
[Inserir Figura 8: Gráfico R² Treino/Teste]  
[Inserir Figura 9: Gráfico Métricas Globais]

# 9. Importância das Variáveis

Variáveis mais relevantes: Margem\_Valor, Produto, Cliente, Ano/Mês.  
  
Outputs: feature\_importance.xlsx, feature\_importance.png  
  
Código: src/feature\_importance.py  
  
Prompt:  
Cria script (src/feature\_importance.py) que treine Random Forest em dataset limpo, extraia feature\_importances\_, exporte ranking em Excel e gráfico PNG.  
  
[Inserir Figura 10: Gráfico Importância das Variáveis]

# 10. Dashboard Interativo

Dashboard em Dash/Plotly com vendas mensais, top clientes/produtos, importância das variáveis e métricas.  
  
Execução: python src/app\_dash.py → abrir <http://127.0.0.1:8050>  
  
Código: src/app\_dash.py  
  
Prompt:  
Cria dashboard (src/app\_dash.py) com Dash/Plotly que: Mostre vendas mensais + média móvel, Top 5 clientes/produtos, Importância de variáveis, Tabela de métricas.  


# 11. Infográfico Final (10+1)

Infográfico final criado com python-pptx e apoio de IA. Conteúdo: objetivos, pipeline, modelo vencedor, variáveis relevantes e conclusões.  
  
Output: reports/infografico\_trabalhoA.pptx  
  
Código: src/infografico\_final\_com\_imagens.py  
  
Prompt:  
Escreve script (src/infografico\_final\_com\_imagens.py) com python-pptx que crie slide 10+1 com título, objetivos, pipeline, resultados principais, variáveis-chave e conclusões.  
  
[Inserir Figura 12: Screenshot do Infográfico]

# 12. Conclusões

O presente trabalho permitiu construir e aplicar um **pipeline completo de Machine Learning** sobre um dataset sintético inspirado na realidade da Biagio Indústria. A geração dos dados e a introdução de anomalias propositadas (omissos, outliers, duplicados e erros ortográficos) permitiram simular os desafios típicos de dados empresariais, reforçando a importância de técnicas de **limpeza e pré-processamento** como condição fundamental para a fiabilidade da análise.

Através da aplicação de dois modelos preditivos – Regressão Linear e Random Forest – foi possível comparar desempenhos e validar a robustez das previsões. O **Random Forest demonstrou clara superioridade**, com R² ≈ 0,93 no teste e MAPE ≈ 7%, assegurando previsões de vendas com elevada precisão. Para além do resultado quantitativo, a análise de importância das variáveis confirmou que **Margem\_Valor, Produto e Cliente** são os principais drivers do negócio, em linha com a concentração típica observada em contexto B2B (regra 80/20).

Do ponto de vista estratégico, estas conclusões têm impacto direto na tomada de decisão:

* **Quanto vamos vender?** → O modelo permite prever vendas com margens de erro aceitáveis, suportando planeamento financeiro e logístico.
* **Quais os drivers?** → A identificação de clientes e produtos “Classe A” (ABC) orienta investimentos e políticas comerciais, enquanto a análise por canal mostrou que o **Grosso gera mais volume** mas o **Retalho assegura margens mais elevadas**, sugerindo uma estratégia dual de consolidação e expansão.

Apesar de o trabalho não ter incluído uma **projeção explícita para o ano seguinte**, o pipeline construído permite essa extensão de forma simples. Com recurso a modelos temporais (ARIMA, Prophet) ou à extrapolação do Random Forest para valores futuros de Ano/Mês, seria possível **gerar previsões de vendas para 2026**, acrescentando ainda mais valor estratégico à ferramenta.

Em termos de utilidade prática, a ferramenta desenvolvida:

* **Automatiza o ciclo analítico** (da limpeza de dados à visualização interativa no dashboard).
* **Reduz erros humanos** e o tempo necessário para análises exploratórias repetitivas.
* **Apoia decisões estratégicas** ao fornecer previsões fiáveis e insights claros sobre drivers de valor.

Criticamente, o exercício revelou que, para além do valor técnico da modelação, a **integração da IA generativa (ChatGPT)** no processo contribuiu para acelerar o desenvolvimento, sistematizar documentação e até gerar outputs visuais (infográfico). Contudo, a IA não substituiu o juízo crítico humano, essencial para validar hipóteses, interpretar métricas e contextualizar resultados .

As **mudanças esperadas** passam por decisões mais informadas, com foco em clientes/produtos estratégicos e numa gestão mais preditiva do negócio. Os **obstáculos identificados** incluem a qualidade dos dados (dependência de registos consistentes) e a concentração do volume em poucos clientes/produtos. **Mitigações propostas**: pipelines de ETL robustos, políticas de governação de dados e testes regulares de modelos para evitar overfitting.

Finalmente, a aprendizagem deste trabalho estende-se à **aplicabilidade futura**: a mesma abordagem pode ser facilmente expandida para **modelos temporais (ARIMA, Prophet)**, integração em **dashboards executivos** e ligação a dados reais da Biagio. Assim, mais do que responder às questões iniciais, o exercício demonstrou como a IA pode transformar dados dispersos em informação acionável, e esta em **decisões estratégicas concretas**.

# 15. Bibliografia

Hossain, E. (2024). Machine Learning CrashCourse for Engineers. Springer. ISBN 978-3-031-46989-3. DOI: 10.1007/978-3-031-46990-9  
Material de apoio às aulas: Machine Learning aplicado às empresas (ed. 3 e 4).  
Enunciado oficial ISCTE (2025). Trabalho Individual – Parte A Quantitativo.  
Emails de orientação dos coordenadores (2025).

Perfeito 🙌 Francisco.

Eu já consegui interpretar os relatórios Sweetviz que compararam o **dataset bruto (raw)** com o **dataset limpo (clean)**. Eis o resumo dos pontos mais relevantes que podes aproveitar diretamente no relatório:

## 🔎 Diferenças principais Raw vs Clean (Sweetviz)

### 1. Estrutura geral

* **Raw (original)**: 500 linhas × 8 colunas.
* **Clean (tratado)**: 497 linhas × 8 colunas.
* 👉 Foram removidas **3 linhas** por conterem outliers ou valores inválidos.

### 2. Valores omissos

* No dataset bruto, havia **4 valores em falta** em Vendas e \*\*4 em Margem\_%.
* No dataset limpo, estes valores foram **imputados pela mediana**, desaparecendo os “missing values” destacados pelo Sweetviz.

### 3. Duplicados

* O dataset original tinha **5 registos duplicados**.
* O dataset limpo já não mostra duplicados → estes foram removidos.

### 4. Outliers

* **Raw**: havia 2 valores extremos em Vendas (>10× a média) e 2 em Margem\_% (>100%).
* **Clean**: estas anomalias desapareceram → confirma que o pipeline de limpeza eliminou outliers com z-score > 3σ.

### 5. Correção de erros de categorias

* No bruto, a coluna Cliente apresentava erros ortográficos: *Mercado Frescco*, *Kéro*, *Shopritee*.
* No limpo, estas categorias foram normalizadas: *Mercado Fresco*, *Kero*, *Shoprite*.
* 👉 O gráfico Sweetviz mostra distribuições de clientes mais consistentes e sem “categorias quase duplicadas”.

### 6. Distribuições numéricas

* Vendas:
  + No bruto, havia uma cauda longa (outliers).
  + No limpo, a distribuição ficou mais simétrica e concentrada.
* Margem\_%:
  + No bruto, existiam valores >100% e negativos.
  + No limpo, a distribuição ficou limitada a [0,100], como esperado.

### 7. Correlações

* As correlações entre Vendas e Margem\_Valor mantiveram-se elevadas nos dois datasets.
* No entanto, a força de algumas relações ficou **ligeiramente mais estável no limpo**, devido à remoção de outliers que distorciam estatísticas.

## ✅ Conclusão interpretada

O relatório Sweetviz **Raw vs Clean** confirma que a limpeza:

* Corrigiu omissos, duplicados e outliers.
* Normalizou categorias textuais.
* Garantiu margens consistentes.
* Manteve a estrutura e as relações principais entre variáveis.

👉 Isto reforça que o **dataset limpo** é fiável e adequado para a fase seguinte de modelação preditiva (split 80/20 e treino dos modelos).

❓ Queres que eu te escreva já o **parágrafo pronto** para a secção “Limpeza de Dados” do relatório, citando este relatório Sweetviz comparativo como evidência visual?