

Agentes Autonomos - Relatorio da Atividade Extra

Discente: Bruno Urbano Rodrigues

1. Framework escolhida

- **Streamlit** para a interface web onde o usuario envia perguntas e visualiza graficos.
- **LangChain** como orquestrador dos agentes de dominio (descricao, padroes, anomalias, visualizacoes).
- **OpenAI Chat API** para interpretar perguntas em portugues e compor respostas estruturadas.
- **Plotly + Kaleido** para gerar graficos interativos e exportar imagens utilizadas no relatorio.
- **pandas, numpy, scikit-learn** para leitura do CSV, estatisticas e analises exploratorias complementares.

2. Estrutura da solucao

- `eda/app.py`: app Streamlit que carrega dados (CSV ou Supabase), aciona o orquestrador LangChain e gera PDF.
- `eda/agents/`: implementacoes dos agentes especializados utilizados nas respostas.
- `eda/domain/analysis.py` e `eda/domain/charts.py`: funcoes reutilizaveis para diagnosticos, correlacoes, outliers e construcao de graficos.
- `eda/Dockerfile` e `eda/docker-compose.yml`: empacotamento e execucao containerizada, lendo variaveis de ambiente via `.env`.
- `lambda/handler.py`: funcao AWS Lambda que processa arquivos do S3 e grava linhas no Supabase.
- `database/*.sql`: funcoes SQL auxiliares usadas pelo Supabase para preparar tabelas durante a ingestao.

3. Perguntas por dominio

3.1 Descricao dos dados

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Quais são os tipos de dados (numéricos, categóricos)?

Perguntar

Domínio acionado: DESCRITIVO

O conjunto de dados contém 31 colunas, todas numéricas. Não há colunas categóricas ou não numéricas além de 'time', 'class' e 'amount'.

Quais sao os tipos de dados (numericos, categoricos)?

Resposta do agente: O dataset contem 31 colunas totalmente numericas. Nao

ha campos categoricos; as variaveis sao `time`, `class`, `amount` e os componentes transformados `v1` a `v28`.

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Qual a distribuicao de cada variavel?

Perguntar

Domnio acionado: DESCRITIVO

O conjunto de dados contém 31 colunas, todas numéricas. Abaixo estão algumas estatísticas descritivas para cada variável:

- **v1:** média = -0.24, mediana = -0.25, desvio padrão = 1.83
- **v2:** média = 0.09, mediana = 0.12, desvio padrão = 1.60
- **v3:** média = 0.46, mediana = 0.71, desvio padrão = 1.78
- **v4:** média = 0.13, mediana = 0.15, desvio padrão = 1.37
- **v5:** média = -0.32, mediana = -0.26, desvio padrão = 1.58
- **v6:** média = 0.08, mediana = -0.21, desvio padrão = 1.44
- **v7:** média = -0.08, mediana = -0.02, desvio padrão = 1.48
- **v8:** média = 0.02, mediana = 0.04, desvio padrão = 1.96
- **amount:** média = 101.15, mediana = 25.73, desvio padrão = 308.05

Todas as colunas têm 0% de valores ausentes. A distribuição das variáveis pode ser influenciada por outliers, que podem distorcer a métrica. Recomenda-se analisar a presença de outliers para uma interpretação mais precisa.

Qual a distribuicao de cada variavel?

Resposta do agente: Todas as colunas sao numericas e mantiveram medias proximas de zero apos a transformacao PCA. Exemplos: `v1` media -0.24 e desvio 1.83; `v2` media 0.09 e desvio 1.60; `amount` media 101.15, mediana 25.73 e desvio 308.05. Nao ha valores ausentes.

Qual o intervalo de cada variavel (minimo, maximo)? **Observacao:** Nao houve captura correspondente nos artefatos fornecidos; o agente nao registrou explicitamente os intervalos nas imagens anexadas.

Quais sao as medidas de tendencia central (media, mediana)?

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Quais são as medidas de tendência central ?

Perguntar

Domnio acionado: DESCRITIVO

As medidas de tendência central para as variáveis numéricas são as seguintes:

- **v1:** média = -0.24, mediana = -0.25
- **v2:** média = 0.09, mediana = 0.12
- **v3:** média = 0.46, mediana = 0.71
- **v4:** média = 0.13, mediana = 0.15
- **v5:** média = -0.32, mediana = -0.26
- **v6:** média = 0.08, mediana = -0.21
- **v7:** média = -0.08, mediana = -0.02
- **v8:** média = 0.02, mediana = 0.04
- **amount:** média = 101.15, mediana = 25.73

Essas medidas indicam a tendência central dos dados, mas é importante considerar a presença de outliers, que podem influenciar a média, especialmente em variáveis com alta variabilidade.

Resposta do agente: As medias e medianas permanecem proximas de zero para `v1` a `v28`, refletindo escalonamento do PCA. `amount` apresenta media 101.15 e mediana 25.73, reforçando cauda longa.

Qual a variabilidade dos dados (desvio padrao, variancia)?

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Qual a variabilidade dos dados ?

Perguntar

Domnio acionado: DESCRITIVO

A variabilidade dos dados pode ser avaliada através do desvio padrão, que indica o quanto os dados se dispersam em relação à média. Aqui estão os desvios padrão das variáveis principais:

- `v1`: desvio padrão = 1.83
- `v2`: desvio padrão = 1.60
- `v3`: desvio padrão = 1.78
- `v4`: desvio padrão = 1.37
- `v5`: desvio padrão = 1.58
- `v6`: desvio padrão = 1.44
- `v7`: desvio padrão = 1.48
- `v8`: desvio padrão = 1.96
- `amount`: desvio padrão = 308.05

Valores de desvio padrão mais altos indicam maior variabilidade nos dados. A presença de outliers pode aumentar o desvio padrão, sugerindo que é importante considerar esses valores ao interpretar a variabilidade.

Resposta do agente: Os desvios padrao dos componentes variam entre ~1.3 e ~2.0 (ex.: `v1` 1.83, `v8` 1.96). `amount` possui desvio padrao 308.05, indicando alta dispersao influenciada por outliers.

3.2 Identificacao de padroes e tendencias

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Existem padrões ou tendências temporais?

Perguntar

Domnio acionado: PADROES

Nenhuma coluna temporal relevante foi detectada nos dados, portanto, não há padrões ou tendências temporais.

Gerar conclusões do agente

Existem padroes ou tendencias temporais?

Resposta do agente: Nao foi identificada coluna temporal relevante alem de `time`, portanto o agente nao detectou padroes ou tendencias temporais significativas.

Pergunte ao agente sobre os dados

Pergunta (ex.: "Qual a media de Amount por Class?")

Quais os valores mais frequentes ou menos frequentes?

Perguntar

Domnio acionado: PADROES

Os valores mais frequentes e menos frequentes por coluna são os seguintes:

- **v1:** valor mais frequente = 1.405 (freq. 2), valor menos frequente = -0.6751 (freq. 1)
- **v2:** valor mais frequente = -0.9583 (freq. 2), valor menos frequente = 0.5378 (freq. 1)
- **v3:** valor mais frequente = -0.6133 (freq. 2), valor menos frequente = 1.532 (freq. 1)
- **v4:** valor mais frequente = -1.798 (freq. 2), valor menos frequente = 0.03022 (freq. 1)
- **v5:** valor mais frequente = 1.01 (freq. 2), valor menos frequente = 0.1993 (freq. 1)
- **v6:** valor mais frequente = 3.536 (freq. 2), valor menos frequente = 0.7019 (freq. 1)
- **v7:** valor mais frequente = -1.583 (freq. 2), valor menos frequente = 0.4429 (freq. 1)
- **v8:** valor mais frequente = 0.9266 (freq. 2), valor menos frequente = 0.3435 (freq. 1)
- **amount:** valor mais frequente = 1 (freq. 36), valor menos frequente = 72.42 (freq. 1)
- **class:** valor mais frequente = 0 (freq. 996), valor menos frequente = 1 (freq. 4)

Esses dados indicam que algumas variáveis têm valores que se repetem com frequência, enquanto outras têm valores raros. Essa análise de padrões e tendências nos dados.

Gerar conclusões do agente

Quais os valores mais frequentes ou menos frequentes?

Resposta do agente: Diversas colunas apresentam valores repetidos com baixa frequência (modo aprox. 1 ou 2 ocorrências). Exemplos: **v1** modo 1.405 (freq. 2) e mínimo -0.6751 (freq. 1); **amount** modo 1 (freq. 36) e valor raro 72.42 (freq. 1). **class** permanece altamente desbalanceada (modo 0, freq. 996).

Pergunte ao agente sobre os dados

Pergunta (ex.: "Qual a media de Amount por Class?")

Existem agrupamentos (clusters) nos dados?

Perguntar

Domnio acionado: PADROES

Os dados apresentam uma correlação forte entre v21 e v22 ($|\rho| \approx 0.655$), o que pode indicar a presença de agrupamentos. Para investigar sobre agrupamentos, é recomendável investigar os valores mais frequentes por coluna.

Gerar conclusões do agente

Existem agrupamentos (clusters) nos dados?

Resposta do agente: O agente observou correlação forte entre **v21** e **v22** ($|\rho| \approx 0.655$), sugerindo possível agrupamento. Recomendou investigar os valores mais frequentes por coluna para aprofundar uma análise de clusters.

3.3 Deteccao de anomalias (outliers)

Pergunte ao agente sobre os dados

Pergunta (ex.: "Qual a media de Amount por Class?")

Existem valores atipicos nos dados ?

Perguntar

Domnio acionado: ANOMALIAS

Foram identificados valores atipicos em várias colunas dos dados. Abaixo estão os detalhes:

- v1: 38 outliers (3.8%) com limites de -4.12 a 4.34 e impacto médio de -0.27.
- v2: 57 outliers (5.7%) com limites de -2.65 a 2.83 e impacto médio de 0.02.
- v3: 54 outliers (5.4%) com limites de -1.78 a 3.19 e impacto médio de -0.31.
- v4: 14 outliers (1.4%) com limites de -3.37 a 3.62 e impacto médio de 0.02.
- v5: 76 outliers (7.6%) com limites de -2.72 a 2.15 e impacto médio de -0.01.
- v6: 97 outliers (9.7%) com limites de -2.39 a 2.15 e impacto médio de 0.27.
- v7: 48 outliers (4.8%) com limites de -2.12 a 2.02 e impacto médio de -0.04.
- v8: 102 outliers (10.2%) com limites de -0.84 a 1.02 e impacto médio de -0.06.
- amount: 96 outliers (9.6%) com limites de -115.90 a 213.32 e impacto médio de 60.02.

Sugestões de tratamento incluem a remoção dos outliers, aplicação de cap/floor (winsorização) ou transformações log/box-cox para lidar com assimétricas. É importante investigar manualmente os registros atipicos antes de decidir removê-los.

Gerar concluses do agente

Existem valores atipicos nos dados?

Resposta do agente: Sim. Exemplos: v1 com 38 outliers (limites -4.12 a 4.34), v6 com 97 outliers (limites -2.39 a 2.15) e amount com 96 outliers (limites -115.90 a 213.32). Foram listados contagens, limites e impacto medio por variavel.

Fonte de dados

Supabase

CSV

SUPABASE_URL

https://tzcybdgggmucqovwfga.supabas

SUPABASE_SERVICE_ROLE_KEY

.....

Schema

public

Tabela

s3_creditcard

Limite de linhas

20000

Carregar dados

Renderizar grafico

Pergunte ao agente sobre os dados

Pergunta (ex.: "Qual a media de Amount por Class?")

Como esses outliers afetam a análise?

Perguntar

Domnio acionado: ANOMALIAS

Os valores atipicos identificados nas colunas podem impactar significativamente a análise. Por exemplo, a coluna v6 apresenta 97 dos dados, com um impacto médio positivo de 0.27. Isso sugere que esses valores extremos podem estar influenciando a média e a distribuição dos dados. Recomenda-se considerar a remoção dos outliers, aplicar técnicas de winsorização para limitar o impacto dos extremos ou realizar transformações log/box-cox para lidar com a assimetria. É importante investigar manualmente os registros atipicos para entender seu contexto antes de decidir sobre a remoção.

Gerar concluses do agente

Gerar PDF do relatrio

Como esses outliers afetam a analise?

Resposta do agente: Os outliers distorcem medias e interpretacoes, especialmente em v6 (impacto medio 0.27). O agente sugere considerar remocao, winsorizacao ou transformacoes logaritmicas para reduzir assimetria antes de conclusoes.

Fonte de dados

Supabase

CSV

SUPABASE_URL

https://lzycybdgggmugovwrfqa.supabas

SUPABASE_SERVICE_ROLE_KEY

.....

Schema

public

Tabela

s3_creditcard

Limite de linhas

20000

-

+

Carregar dados

Renderizar grafico

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Podem ser removidos, transformados ou investigados?

Perguntar

Domnio acionado: DESCRITIVO

Sim, os outliers podem ser removidos, transformados ou investigados. A remoção pode ser apropriada para representarem a realidade do fenômeno estudado. A transformação, como a winsorização ou a aplicação do impacto dos outliers na análise. Além disso, é importante investigar os outliers para entender seu impacto e como devem ser tratados de forma diferente.

Gerar concluses do agente

Podem ser removidos, transformados ou investigados?

Resposta do agente: Sim. Recomenda avaliar remocao quando outliers forem erros, aplicar winsorizacao ou transformacoes logaritmicas para reduzir impacto, alem de investigar manualmente cada caso para decidir o tratamento.

3.4 Relacoes entre variaveis

Como as variaveis estao relacionadas umas com as outras?

Fonte de dados

Supabase

CSV

SUPABASE_URL

https://lzycybdgggmugovwrfqa.supabas

SUPABASE_SERVICE_ROLE_KEY

.....

Schema

public

Tabela

s3_creditcard

Limite de linhas

20000

-

+

Carregar dados

Renderizar grafico

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Como as variáveis estão relacionadas umas com as outras?

Perguntar

Domnio acionado: DESCRITIVO

As variáveis no conjunto de dados estão todas numéricas e apresentam uma variedade de estatísticas descritivas. A correlação entre v21 e v22 é forte ($|rho|=0.655$), indicando que essas variáveis estão relacionadas. Além disso, a correlação entre v2 e amount ($|rho|=0.504$) sugere que há uma relação moderada entre essas duas variáveis. Para entender melhor como as variáveis estão relacionadas, é importante analisar a dispersão e a distribuição de cada uma delas, além de considerar a presença de outliers que podem influenciar essas relações.

Gerar concluses do agente

Gerar PDF do relatirio

Resposta do agente: Todas as colunas sao numericas; ha correlacoes destacadas entre v21 e v22 ($|rho|$ aprox. 0.655) e entre v2 e amount ($|rho|$ aprox. 0.504). O agente ressalta necessidade de analisar dispersoes e outliers para entender as relacoes.

Fonte de dados

Supabase

CSV

SUPABASE_URL

https://lzycybdgggmugovwrfqa.supabas

SUPABASE_SERVICE_ROLE_KEY

.....

Schema

public

Tabela

s3_creditcard

Limite de linhas

20000

-

+

Carregar dados

30

Renderizar grafico

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Existe correlação entre as variáveis?

Perguntar

Domnio acionado: PADROES

Existem correlações significativas entre algumas variáveis. A correlação mais forte é entre v21 e v22 ($|rho|=0.655$), seguida pela correlação entre v2 e amount ($|rho|=0.504$). Outras correlações relevantes incluem v27 com v28 ($|rho|=0.435$), v5 com v7 ($|rho|=0.415$) e v1 com v25 ($|rho|=0.414$). O impacto médio significativo foi detectada.

Gerar concluses do agente

Gerar PDF do relatirio

Existe correlacao entre as variaveis?

Resposta do agente: Sim. Alem da correlacao forte v21-v22, outras relevantes

6

são v27-v28 ($|\rho|$ aprox. 0.435), v5-v7 ($|\rho|$ aprox. 0.415) e v1-v25 ($|\rho|$ aprox. 0.414). Nenhuma categoria com impacto medio significativo foi encontrada.

Quais variaveis parecem ter maior ou menor influencia sobre outras?

Fonte de dados

☒ Supabase ☐ CSV

SUPABASE_URL

https://lzycybdgggtgmuqovwfga.supabas

SUPABASE_SERVICE_ROLE_KEY

.....

Schema

public

Tabela

s3_creditcard

Limite de linhas

20000

Carregar dados

30

Renderizar grafico

Pergunte ao agente sobre os dados

Pergunta (ex.: 'Qual a media de Amount por Class?')

Quais variáveis parecem ter maior ou menor influência sobre outras?

Perguntar

Domínio acionado: PADROES

As variáveis que parecem ter maior influência sobre outras são: v21 e v22, com uma correlação forte de $|\rho|=0.655$. Em seguida, v2 e amount apresentam uma correlação moderada de $|\rho|=0.504$. Outras correlações relevantes incluem v27 com v28 ($|\rho|=0.435$), v5 com v7 ($|\rho|=0.415$) e v1 com v25 ($|\rho|=0.414$). Nenhuma categoria com impacto médio significativo foi detectada.

Gerar conclusões do agente

Gerar PDF do relatório

Resposta do agente: v21 e v22 exercem maior influencia mutua com correlacao forte. Em seguida, v2 com amount ($|\rho|$ aprox. 0.504) e pares v27-v28, v5-v7, v1-v25 mostram correlacoes moderadas. Nao foram encontradas categorias com impacto medio alto.