

Aprendizagem e Extração do Conhecimento

PL01 · CRISP – DM | Business Understanding | Data Understanding

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.

Aulas Práticas da UC



De raiz, um processo de *Data Mining* para Decisão Clínica

Da compreensão do problema até à tomada de decisão.



Grupo

Desenvolvimento de projeto acompanhado



Exercício Comum

Em aula igual para todos



Objetivos Intermédios

Cumprir objetivos definidos

Tópicos

PARTE 1 · FUNDAMENTOS DE DESENVOLVIMENTO DE PROJETOS EM DM

- ✓ CRISP-DM

PARTE 2 · DATA MINING

- ✓ Relembrar Data Mining

PARTE 3 · PROJETO DE AULA

- ✓ Compreensão do Negócio

PARTE 4 · COMPREENSÃO DOS DADOS

- ✓ AI Studio
- ✓ Recolha de Dados
- ✓ Metadados e Qualidade
- ✓ Visualizações e Insights

PARTE 1 · FUNDAMENTOS DE DESENVOLVIMENTO DE PROJETOS EM DM

Um projeto não é um exercício

Exercício de aula

- Problema fechado e já preparado
- Uma resposta certa
- Dura uma aula
- Ninguém o vai usar

Projeto

- Problema real e em aberto
- Várias soluções possíveis
- Tem fases e entregáveis
- É construído para alguém o usar

Sem método, um projeto tem atrasos e um âmbito que nunca fecha.

Processo de Desenvolvimento de Software

- Define como se estrutura o desenvolvimento de software
- Identifica as fases de desenvolvimento e como se passa de umas para outras:
 - Quem faz o quê
 - Quando é feito
 - Durante quanto tempo



Processo de Desenvolvimento de Software

Code and fix

Raramente útil, mas muito utilizado

Não tanto uma estratégia deliberada, mais o resultado de falta de planeamento e tempo:

- **Sem muita análise:** começa-se a produzir código de imediato
- Eventualmente fazem-se testes e os inevitáveis erros têm que ser resolvidos

Vantagens:

- Produção de código é imediata
- Não requer muita competência – qualquer programador pode utilizar

Problemas:

- Não existe forma de avaliar progresso
- Não existe forma de avaliar qualidade e/ou de prever riscos
- Não escala bem para equipas (múltiplos programadores)
- Apenas viável para pequenos projetos
- Usado em provas de conceito ou demos

Processo de Desenvolvimento de Software

Planeamento

- Decisão de avançar com o projeto
- Gestão do Projeto

Análise

- Análise do Domínio do Problema
- Análise de Requisitos

Conceção

- Conceção da Arquitetura
- Conceção do Comportamento

Implementação

- Construção
- Teste
- Instalação
- Manutenção

Diferentes Metodologias



Waterfall

Sequencial: cada fase fecha antes de a seguinte começar.

- **Bom quando:** os requisitos são estáveis e claros
- **Risco:** erros só aparecem lá ao fim



Agile / Iterativo

Ciclos curtos que repetem as fases, com feedback contínuo.

- **Bom quando:** os requisitos evoluem ou são incertos
- **Risco:** exige disciplina e ritmo constante

Abordagens Estruturadas - Waterfall



Define uma série de etapas executadas sequencialmente.

Abordagens Estruturadas - Waterfall

Assume que é sempre possível tomar as decisões mais corretas. Mas....

- como prever situações de alto risco mais à frente?
- como favorecer reutilização?
- como lidar com alterações (p.e. de requisitos) durante o processo?

Quanto mais tarde um problema é encontrado, mas caro será corrigi-lo!

Abordagens Iterativas e Incrementais - Agile

✓ Vantagens:

- Entregas intermédias permitem demonstração/exploração do sistema desde cedo
- Identificação de novos requisitos / identificação de problemas
- Demonstração de progresso (potencia satisfação dos clientes)
- Maior capacidade de lidar com incerteza, maior controlo de risco
- Promovem a inclusão de objetivos de qualidade no processo de desenvolvimento
- Promovem a utilização de ferramentas de automatização do desenvolvimento

Abordagens Iterativas e Incrementais - Agile

→ Problemas:

- Utilizadores começam por trabalhar com sistemas que estão incompletos
- Gestão de expectativas
- Dependência da qualidade da avaliação de risco/identificação de funcionalidades mais relevantes
- Decisões tomadas no início podem revelar-se inadequadas mais tarde
- Produto não deve ser monolítico
- Entregas causam trabalho adicional

Qual a nossa metodologia?

CRIPSP-DM

CRoss Industry **S**tandard **P**rocess for **D**ata **M**ining

Principais objetivos:

- Encorajar a utilização de ferramentas interoperáveis ao longo de todo o processo de Data Mining;
- Retirar conhecimento valioso de tarefas simples de Data Mining.

CRISP-DM



Confiável e Repetível



Diretrizes



Flexível

CRISP-DM



Confiável e Repetível

O processo de Data Mining deve ser confiável e repetível por pessoas com pouco conhecimento em DM!!

CRISP-DM



Diretrizes

CRISP-DM é uma metodologia uniforme com diretrizes,
documentação de experiência!

CRISP-DM



Flexível

A metodologia CRISP- DM é flexível o suficiente para ter em conta problemas de negócio diferentes e dados diferentes

CRISP-DM

→ Características:

- **Metodologia** para o registo de experiências;
- Permite que os projetos sejam **replicados**;
- Ajuda no **planeamento** e na **gestão de projetos**;
- "**Fator de conforto**" para novos utilizadores;
- Demonstra a **maturidade** do Data Mining.

CRISP-DM



Define uma série de etapas executadas de forma iterativa e reversível.

CRISP-DM



CRISP-DM

01



Compreensão do
Negócio

- **Determinar objetivos do negócio**
 - Documentar o background do negócio
 - Definir Sucesso do negócio
- **Determinar objetivos da aplicação de DM**
 - Definir o sucesso do processo
 - Identificar o critério de sucesso para a aplicação de DM
- **Avaliar a situação atual**
 - Avaliar recursos disponíveis
 - Identificar constrangimentos
 - Documentar custos e benefícios
- **Planeamento**
 - Definir o plano de implementação
 - Definir técnicas e ferramentas

O que se pretende atingir com o processo de Data Mining? Qual o critério para o sucesso?

02



Compreensão dos
Dados

- **Recolher os dados iniciais:**
 - Definir quais as fontes e métodos de recolha
- **Explorar os dados:**
 - Produzir um relatório da exploração dos dados
 - Analisar em detalhe atributos de interesse
 - Definir relações e agrupamento de atributos
- **Descrição detalhada dos dados:**
 - Efetuar a avaliação de quantidade
 - Detalhar o tipo de atributos
 - Avaliar estatisticamente: Min,Max,Mean,etc
- **Avaliação de qualidade:**
 - Avaliar a existência de campos nulos ou em falta
 - Avaliar valor <-> significado do atributo

Recolher, organizar e analisar os dados! Garantir a qualidade e a compreensão dos dados disponíveis!

03



Preparação dos
Dados

- **Construir dados:**
 - Derivar atributos
 - Validar a reconstrução de campos nulos
 - Integrar dados obtidos de outras fontes
- **Limpar os dados:**
 - Efetuar a limpeza de dados desnecessários (identificadores)
- **Integração de outras fontes:**
 - Integrar atributos resultantes de outras fontes
- **Formatar os dados:**
 - Organizar atributos dentro do dataset
 - Garantir coerência para a criação de datasets precisos

Garantir que os dados estão prontos para serem adicionados aos modelos de ML.

04



Construção dos
Modelos

- **Selecionar a técnica:**
 - Seleção dos modelos a usar
 - Interpretar a técnica juntamente com as conclusões retiradas sobre os dados
- **Construção do modelo:**
 - Definição de parâmetros iniciais
 - Descrever o modelo e a interpretação para o seu uso
- **Modelo de Teste:**
 - Definir qual o modelo de teste a utilizar
 - Divisão do dataset em dados de treino e teste
- **Avaliação do modelo:**
 - Interpretação inicial dos resultados
 - Comparar com potenciais expectativas e conhecimento existente

Implementação de modelos de DM que apliquem as técnicas selecionadas (algoritmos e testes) aos dados

CRISP-DM

05



Avaliação da
Performance

- **Decisão:**
 - Avaliação detalhada dos resultados
 - Redefinição dos modelos de DM
 - Comparação com os critérios de sucesso
 - Avaliação de potenciais correções nos dados
- **Próximos passos:**
 - Definição de parâmetros iniciais
 - Descrever o modelo e a interpretação para o seu uso

Revisão crítica de resultados. Os critérios para o sucesso definidos foram atingidos?

06



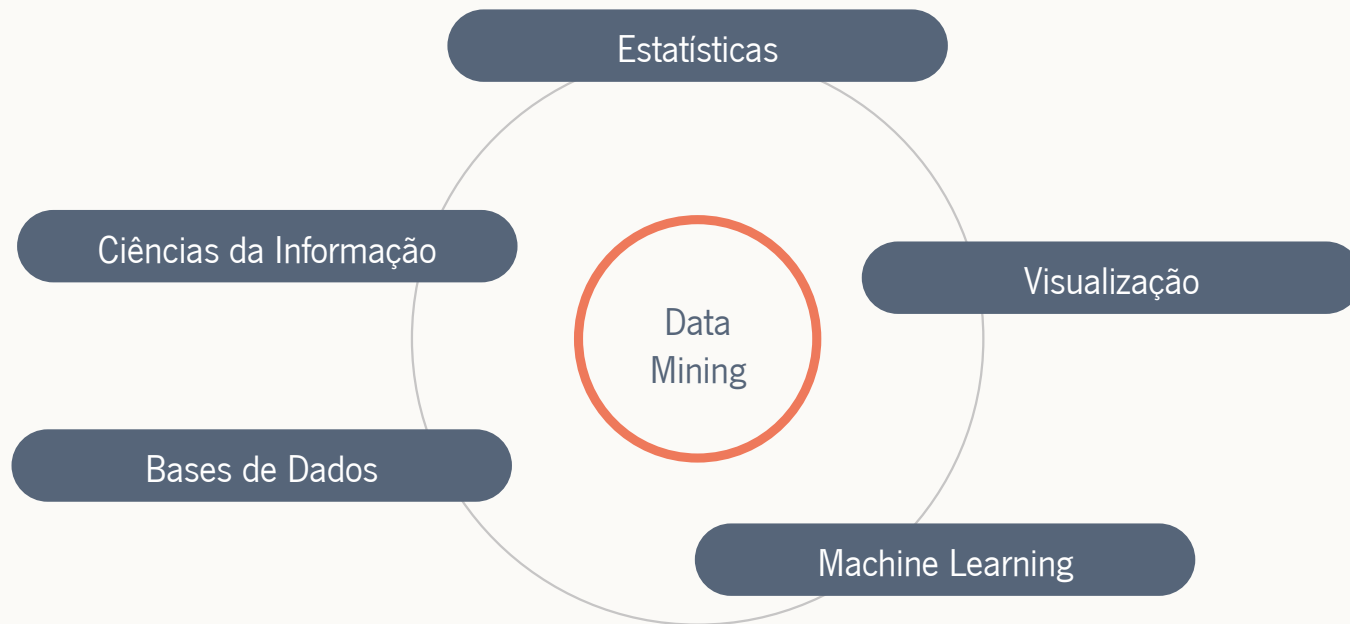
Colocação em
Produção

- **Plano de implementação:**
 - Descrever o plano de implementação do conhecimento gerado
 - Escrever o relatório final e as visualizações do projeto
- **Revisão e Manutenção:**
 - Rever periodicamente o processo implementado
 - Avaliar potenciais pontos de melhoria e estrangimentos

Colocar em produção e avaliar o impacto real!

PARTE 2 • DATA MINING

Data Mining



Data Mining – Técnicas/Modelos

- **Regras de Associação:**
 - Carrinhos de compras em supermercados
- **Classificação:**
 - Construir modelos (funções) que descrevem e distinguem classes ou conceitos para previsão futura
- **Agrupamento (Clustering):**
 - A etiqueta da classe é desconhecida: Agrupar dados para formar novas classes, por exemplo, cluster clientes supermercado (zonas - dias) – Maximização da semelhança intra-classe e minimização da semelhança interclasse.

Data Mining – Técnicas/Modelos

- **Análise de Desvios:**
 - *Outlier*: um objeto de dados que não está de acordo com o comportamento geral dos dados - Ruído ou exceção? Não! útil na detecção de fraudes, análise de eventos raros
- **Análise de tendências e evolução:**
 - Tendência e desvio: análise de regressão
 - Mineração de padrões sequenciais, análise de periodicidade
 - Análise baseada na similaridade

A nossa stack



DataSets

Kaggle, UCI, worldindata, etc



AIStudio

Pré-processamento, modelação + exploração

Antes de sair hoje

Confirmar instalado:



AI Studio a correr

PARTE 03 • PROJETO DE AULA

Contexto

Uma equipa clínica recolhe regularmente dados de pacientes (idade, IMC, tensão, colesterol, hábitos de vida) e exames laboratoriais (glicose, hemoglobina glicada, creatinina, triglicéridos). Os dados estão dispersos em dois ficheiros, com valores em falta e possíveis outliers, e ninguém ainda os analisou em conjunto.

A equipa quer saber: que fatores estão associados ao risco de diabetes, e será possível usar estes dados para identificar pacientes em risco antes do diagnóstico?

O Data Mining pode ajudar a responder a isto em várias etapas: analisar e preparar os dados, descobrir que variáveis se correlacionam com a diabetes, identificar combinações de fatores de risco através de regras de associação, e construir um modelo de classificação que apoie a triagem de novos pacientes.

CRISP-DM – Compreensão do Negócio



DataSet



Pacientes.csv – Informação relativa aos pacientes



Exames.csv – Informação relativa a exames

download via eLearning

Ficha de Exercícios 01

1 • Compreensão do Negócio

- Determinar objetivos do negócio
- Determinar objetivos da aplicação do DM
- Avaliar a situação atual
- Planeamento

PARTE 04 • COMPREENSÃO DOS DADOS

CRISP-DM



AI Studio

O **AI Studio** é uma ferramenta comercial para análise de dados que utiliza machine learning e pode ser considerada uma alternativa para a ferramenta *Weka*.

Esta ferramenta desenvolvida pela empresa com o mesmo nome, tem como principal missão acelerar o processo de criação de análises preditivas e torná-las mais fáceis para serem aplicadas em cenários práticos de negócios.



Compreensão dos Dados

- **Recolha das fontes de dados relevantes**
- **Descrição: nº de registos, tipos de atributos, formato**
- **Exploração: estatística, visualizações, primeiras hipóteses**
- **Verificação da qualidade: falhas, inconsistências, outliers**

Compreensão dos Dados

- **Integridade referencial: algum id_paciente órfão em exames?**
- **Cobertura: há doentes sem qualquer exame?**
- **Cardinalidade: exames por doente, em média e máximo**
- **Um id órfão ou doente sem exames indica erro de recolha**

Compreensão dos Dados

Como escolher estatísticas e gráficos segundo o tipo de variável

Tipo	Exemplos nestes dados	Estatísticas adequadas
Numérica contínua	idade, imc, glicose_jejum, creatinina	média, mediana, desvio padrão, quartis
Numérica discreta	nº de exames por doente	moda, contagens, distribuição
Categórica nominal	sexo	frequências, moda
Binária	diabetico, fumador, sedentarismo	proporção, frequência
Temporal	data_visita	amplitude, ordenação, contagem por período

Compreensão dos Dados

- **Tendência central: média, mediana e moda**
- **Dispersão: desvio padrão, IQR, mínimo e máximo**
- **Forma: assimetria (enviesamento) e curtose**
- **Valores em falta: quantificar o padrão (tratar só na Data Preparation)**

Os problemas de qualidade a caçar

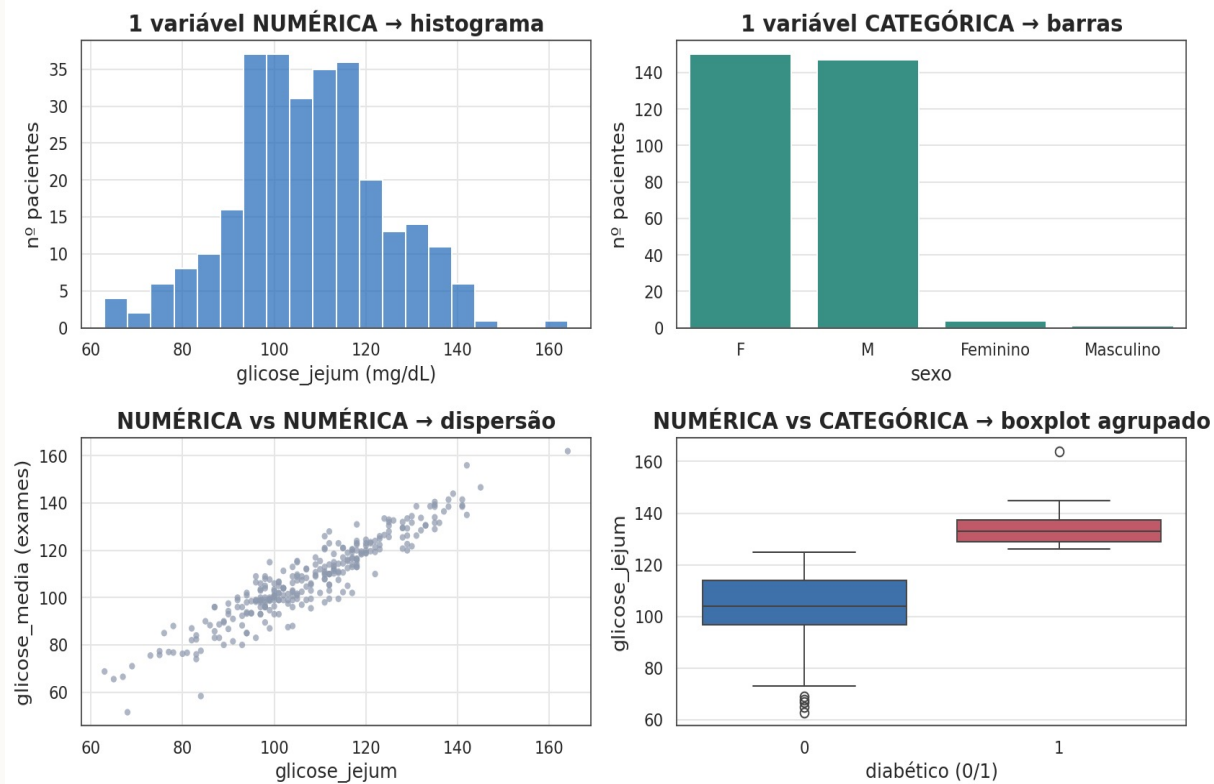
Cada problema tem uma forma simples de deteção na Data Understanding

Problema	Onde aparece nestes dados	Como se deteta
Duplicados	2 linhas repetidas em pacientes (P0004, P0051)	nº de linhas vs id únicos
Categorias inconsistentes	sexo: F/M mas também Feminino/Masculino	gráfico de barras
Outliers implausíveis	idade 137; imc 0 e 87.4; creatinina negativa; triglicéridos 1500	mínimo/máximo, boxplot
Valores em falta	colesterol_total (~10%), imc, hba1c...	% de nulos por coluna
Chave órfã	um id_paciente em exames sem par em pacientes	integridade referencial

Que gráfico usar?

Antes de resumir uma variável num número, olhe para a sua distribuição

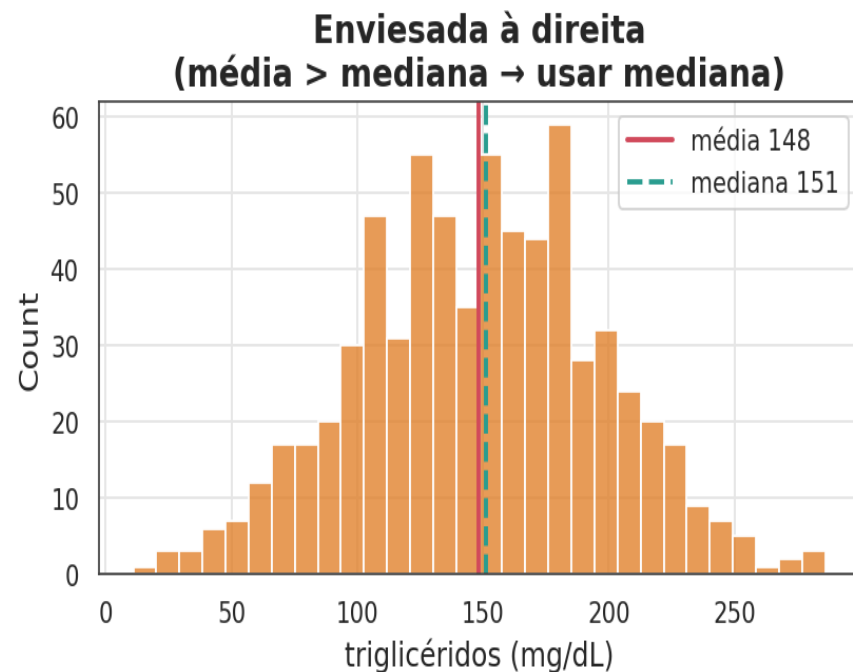
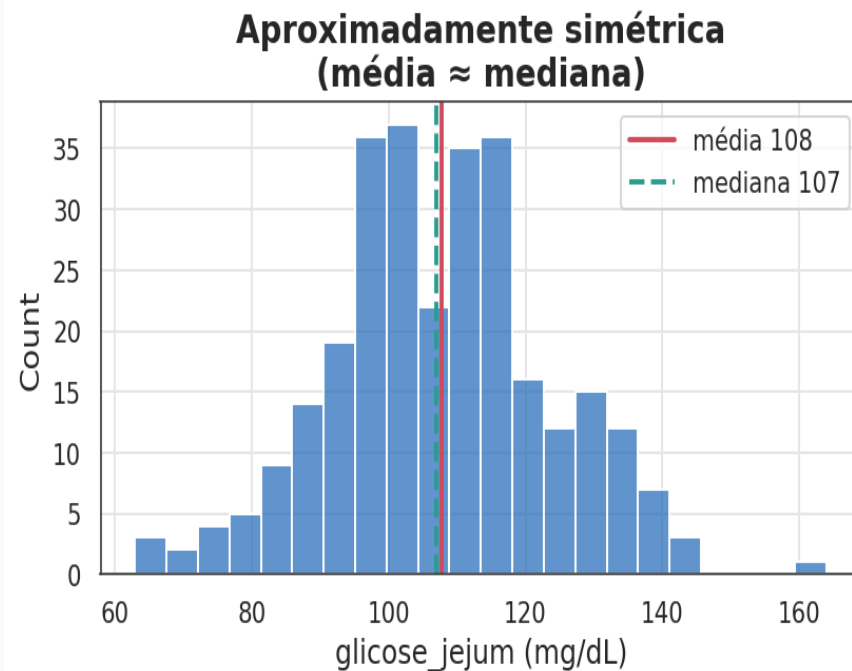
Que gráfico usar? Depende do tipo de variável



Histograma: a forma da distribuição

Simétrica → a média descreve bem; enviesada → a mediana é o valor típico

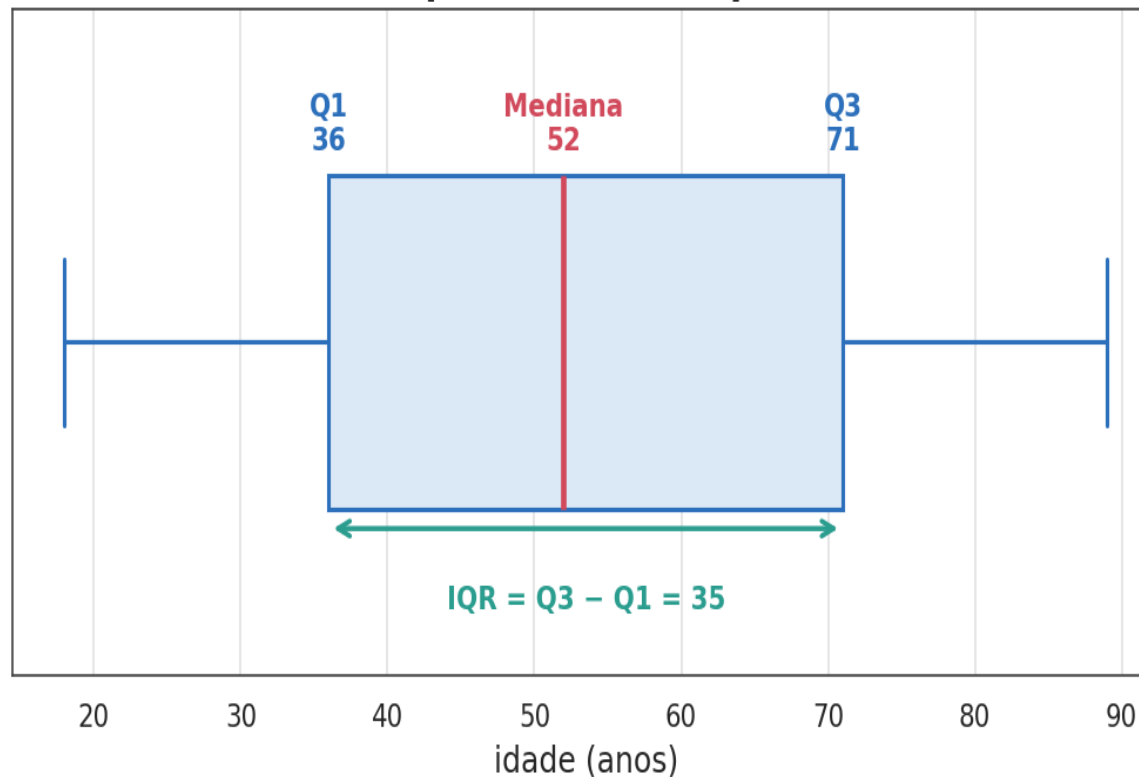
A forma da distribuição condiciona a estatística a usar



Anatomia de um boxplot

Num só gráfico: mediana, quartis, IQR e potenciais outliers

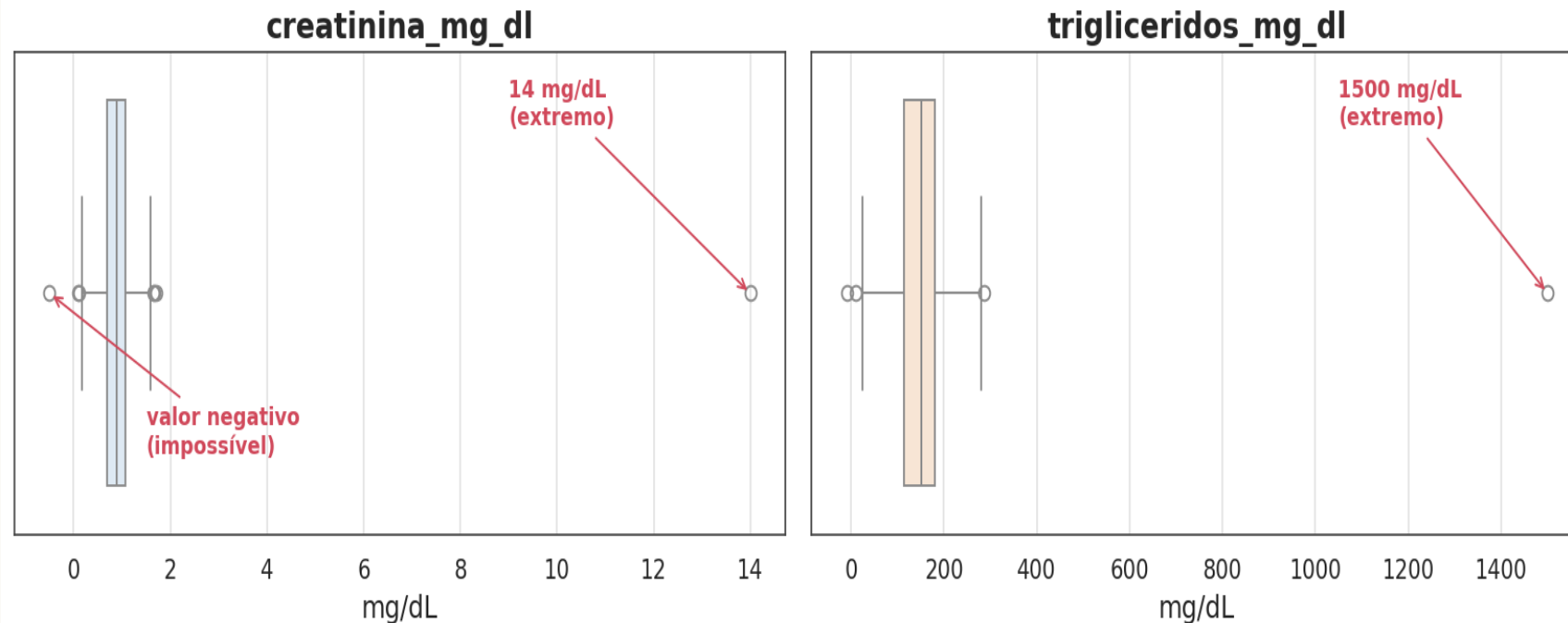
Anatomia de um boxplot: mediana, quartis, IQR e outliers



O boxplot revela valores implausíveis

Sem limpeza, uma creatinina negativa e triglicéridos de 1500 saltam à vista

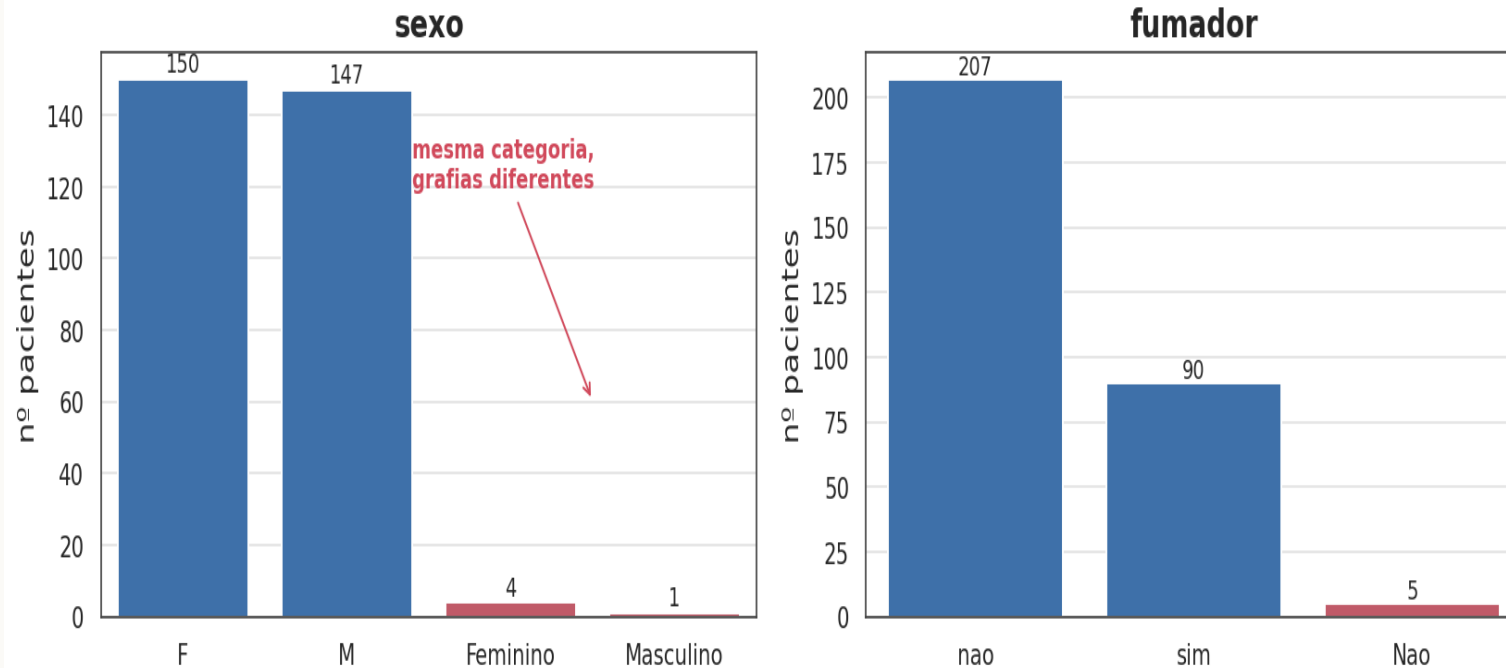
O boxplot revela valores implausíveis (dados por limpar)



Barras: categorias inconsistentes

Quatro categorias de sexo onde deviam existir apenas duas

Gráfico de barras expõe categorias inconsistentes



Compreensão dos Dados

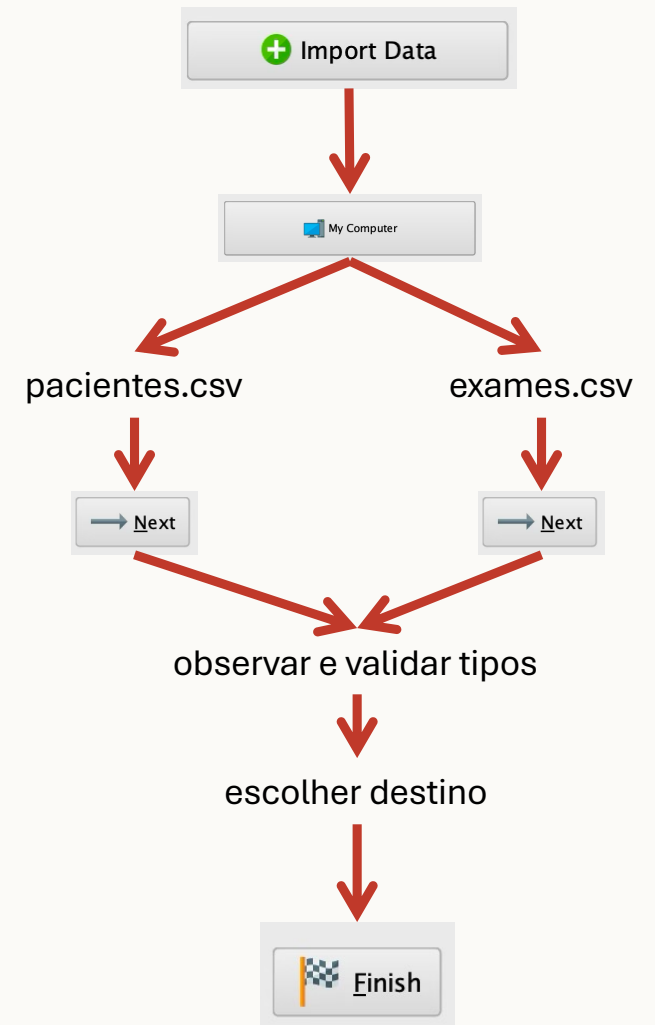
Caso Diabetes

- Carregar e inspecionar os dados
- Identificar problemas de qualidade
- Explorar estatísticas e distribuições
- Visualizar relações entre atributos

Compreensão dos Dados

Caso Diabetes

- **Carregar e inspecionar os dados**
- Identificar problemas de qualidade
- Explorar estatísticas e distribuições
- Visualizar relações entre atributos



Compreensão dos Dados

Caso Diabetes - pacientes.csv

id_paciente: identificador único do paciente

idade: idade do paciente, em anos

sexo: sexo do paciente

imc: índice de massa corporal

tensao_sistolica: tensão arterial sistólica, em mmHg

tensao_diastolica: tensão arterial diastólica, em mmHg

colesterol_total: colesterol total, em mg/dL

fumador: indica se o paciente é fumador

sedentarismo: indica se o paciente tem um estilo de vida sedentário

historico_familiar_diabetes: indica se há antecedentes familiares de diabetes

glicose_jejum: glicose em jejum, em mg/dL

diabetico: atributo-alvo; indica se o paciente tem diabetes

Compreensão dos Dados

Caso Diabetes - exames.csv

id_exame: identificador único do exame

id_paciente: chave que liga o exame ao paciente em *pacientes.csv*

data_visita: data em que o exame foi realizado

glicose_mg_dl: glicose medida na visita, em mg/dL

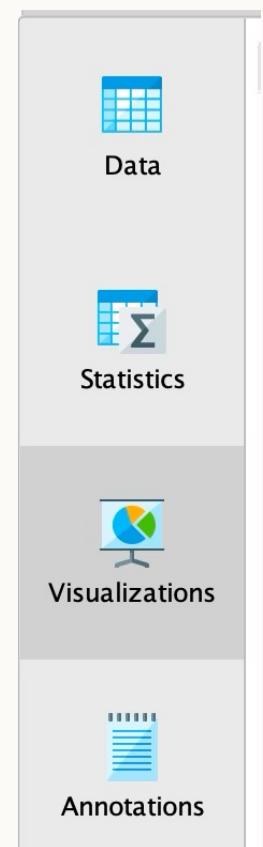
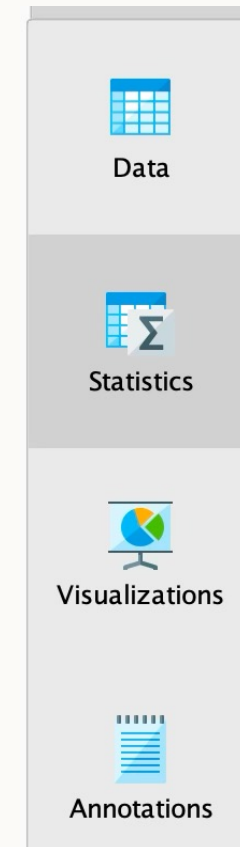
hba1c_percent: hemoglobina glicada, em percentagem; reflete o controlo da glicose nos últimos meses

creatinina_mg_dl: creatinina, em mg/dL; indicador da função renal

trigliceridos_mg_dl: triglicéridos, em mg/dL

Compreensão dos Dados: Caso Diabetes

- Carregar e inspecionar os dados
- **Identificar problemas de qualidade**
- **Explorar estatísticas e distribuições**
- **Visualizar relações entre atributos**



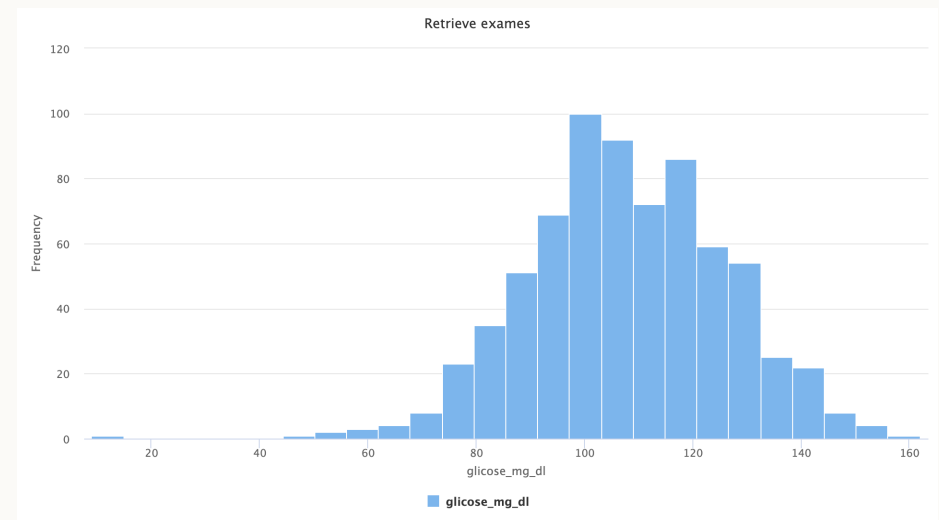
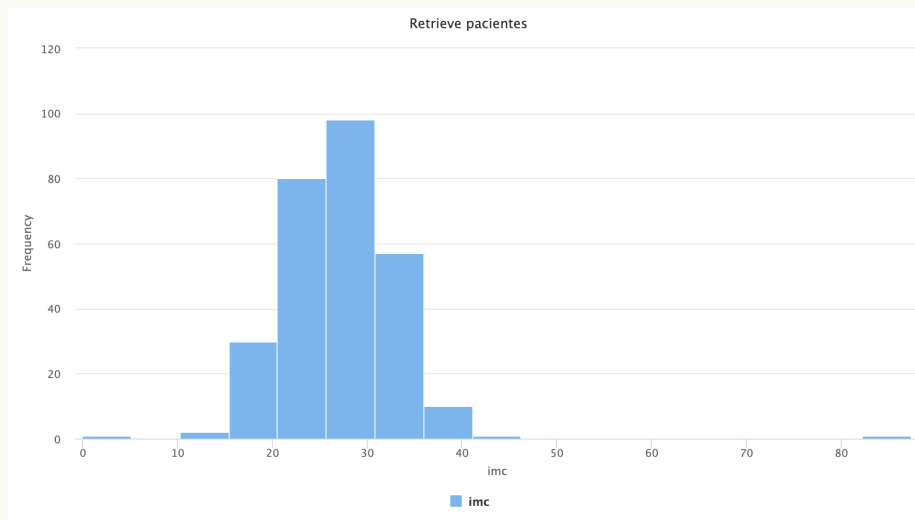
Metadados e Qualidade

Avaliar os metadados dos dados importados para os 2 datasets!

Name	Type	Missing	Statistics		
✓ id_paciente	Nominal	0	Least P0300 (1)	Most P0004 (2)	Values P0004 (2), P0051 (2), ...[298 more]
✓ idade	Integer	0	Min 18	Max 137	Average 53.586
✓ sexo	Nominal	0	Least Masculino (1)	Most F (150)	Values F (150), M (147), ...[2 more]
✓ imc	Real	22	Min 0	Max 87.400	Average 27.177
✓ tensao_sistolica	Real	0	Min 75	Max 280	Average 126.974
✓ tensao_diastolica	Real	15	Min 10	Max 117	Average 81.411
✓ colesterol_total	Real	30	Min 9	Max 294	Average 199.426
✓ fumador	Nominal	0	Least Nao (5)	Most nao (207)	Values nao (207), sim (90), ...[1 more]
✓ sedentarismo	Nominal	0	Least sim (124)	Most nao (178)	Values nao (178), sim (124)
✓ historico_familiar_diabetes	Nominal	9	Least sim (78)	Most nao (215)	Values nao (215), sim (78)
✓ glicose_jejum	Real	12	Min 63	Max 164	Average 107.634
✓ diabetico	Integer	0	Min 0	Max 1	Average 0.129

Visualizações – pacientes.csv

Visualizar **histogramas** dos diferentes atributos. Tentar perceber tendências



Visualizações – Classe ?

Gráfico de barras de uma potencial classe: Diabético

Plot 1 

Plot type

 Bar (Column) ▼

Value column

diabetico ▼

☒ Aggregate data

Group by

- ▼

Aggregation Function

Count ▼

Sorting

None ▼

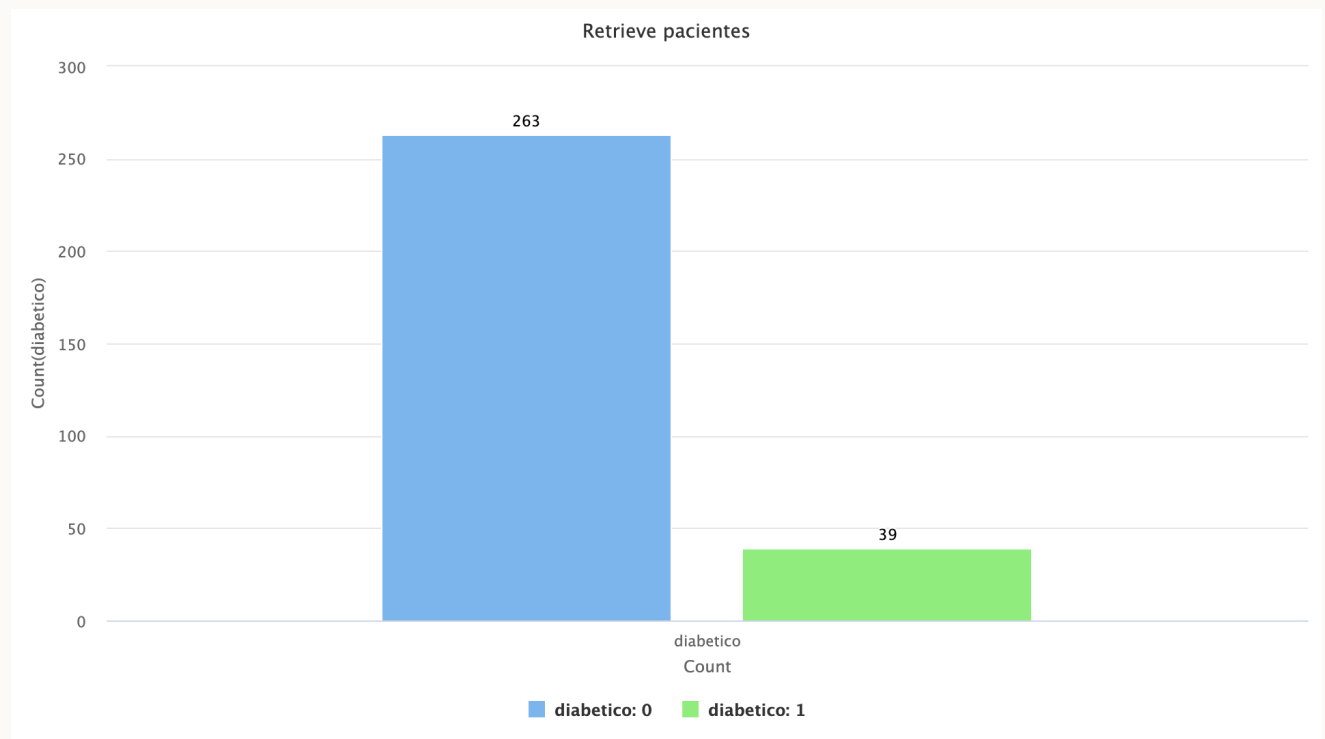
Color Group

diabetico ▼

Stacking

No stacking ▼

Plot style >>



Visualizações – Insights??

Criar um Scatter Plot com glicose_jejum no eixo X e imc no eixo Y, com cor por diabético.

Plot 1 

Plot type

 Scatter / Bubble ▼

X-Axis column

glicose_jejum ▼

Value column

imc ▼

Color

diabetico ▼

Size

– ▼

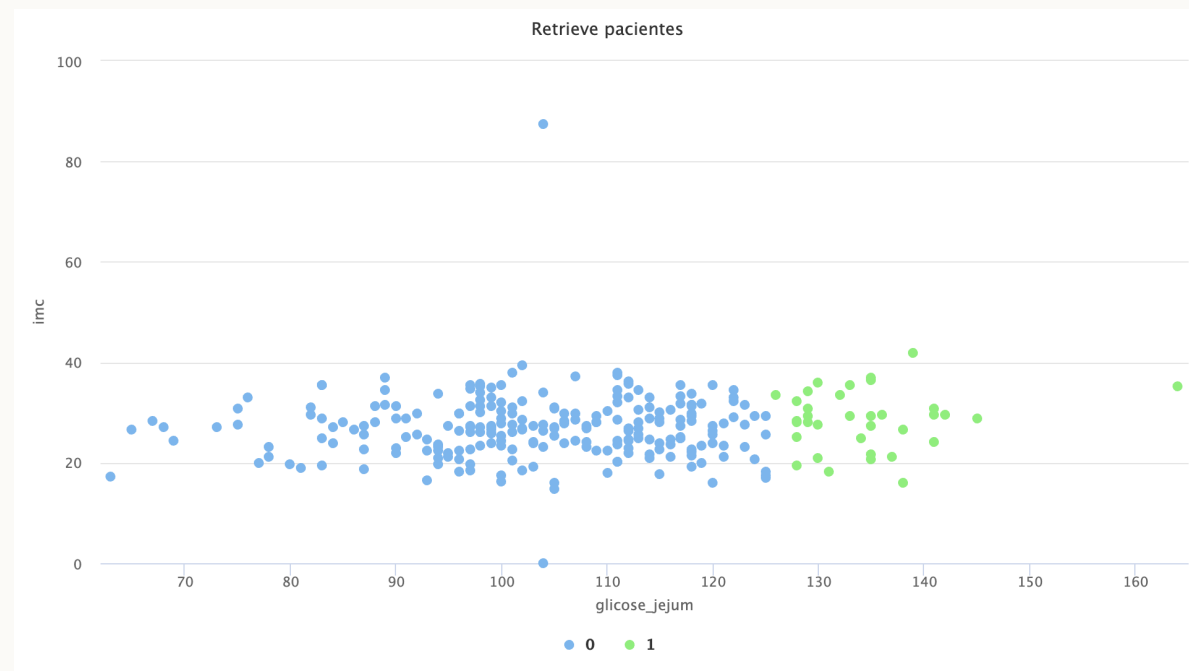
Jitter

Regression interpolation

None ▼

Plot style >>



Visualizações – Insights??

Criar um Scatter Plot com glicose_jejum no eixo X e idade no eixo Y, com cor por diabético.

Plot 1 

Plot type

 Scatter / Bubble ▼

X-Axis column

glicose_jejum ▼

Value column

idade ▼

Color

diabetico ▼

Size

- ▼

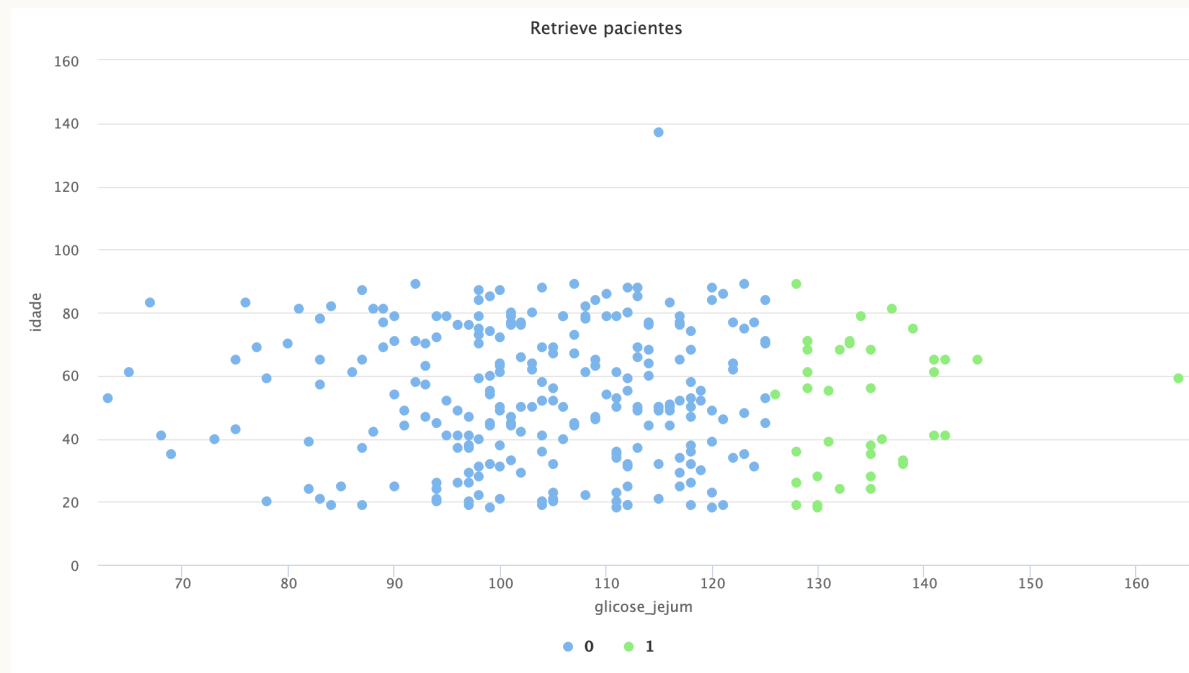
Jitter

 ☐

Regression interpolation

None ▼

Plot style >>



Exercício Conjunto

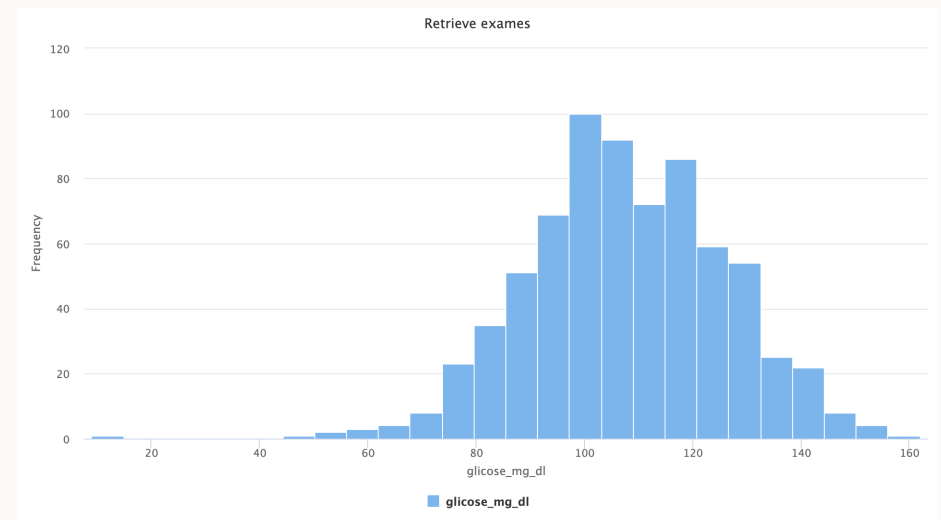
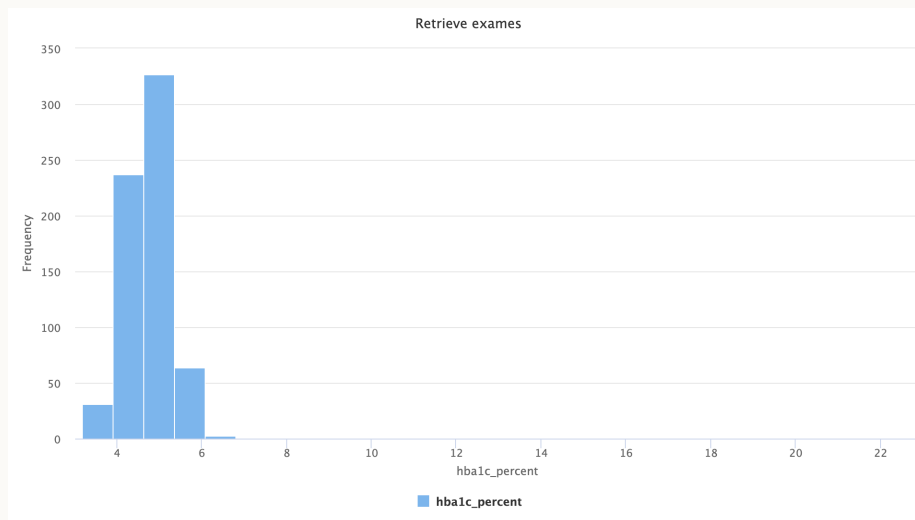
Compreensão dos Dados:

Os pacientes diabéticos aparecem mais concentrados em alguma zona do gráfico?



Visualizações – exames.csv

Criar um Histograma para hba1c_percent e outro para glicose_mg_dl.



Qualidade dos Dados

- Reparar que *pacientes.csv* tem 1 linha por paciente, mas *exames.csv* tem entre 1 e 4 linhas por paciente.
- Esta diferença de granularidade terá de ser resolvida no Data Preparation (ex.: Aggregate antes do Join).
- Verificar se existe algum id_paciente em *pacientes.csv* sem nenhuma linha correspondente em *exames.csv*.

Compreensão dos dados

- **Dois datasets, granularidades diferentes**
- **Valores em falta em ambos os ficheiros**
- **Forte desequilíbrio de classes em diabético**
- **Primeiras pistas visuais: glicose_jejum parece distinguir bem os dois grupos**

Síntese

✓ Qualquer projeto tem método: Waterfall / Ágil / CRISP-DM

✓ CRISP-DM : Confiável, Repetível e Flexível

✓ Data Mining: Engloba Várias áreas da ciência de dados

✓ Conhecemos o nosso exercício de exemplo

PRÓXIMA SESSÃO

Compreensão dos Dados

Transformar este enunciado num processo de data mining confiável e repetível.

Deixar a stack instalada e a funcionar!

Aprendizagem e Extração do Conhecimento

PL01 · CRISP – DM | Business Understanding | Data Understanding

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.