



BigML

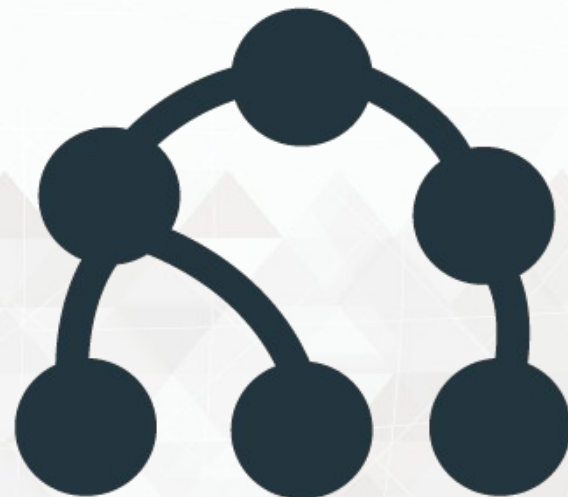
Árvores de Decisão

2023

<lapti>

Um modelo ideal para ML

- Grande poder representacional
 - › Ajustando uma linha é um exemplo de baixo
 - › Deep neural networks é um exemplo de alto
- Facilidade de uso
 - › Fácil configuração – relativamente poucos parâmetros
 - › Fácil interpretação – como as decisões são tomadas?
 - › Fácil para se colocar em produção
- Habilidade para trabalhar com dados de mundo real
 - › Tipos variados: numéricos, categóricos, textos, etc
 - › Tratar valores inexistentes (missing values)
 - › Resistente a outliers
- E existem muitas possibilidades a escolher, claro...



Decision Trees

- Classification and Regression Trees (CART)

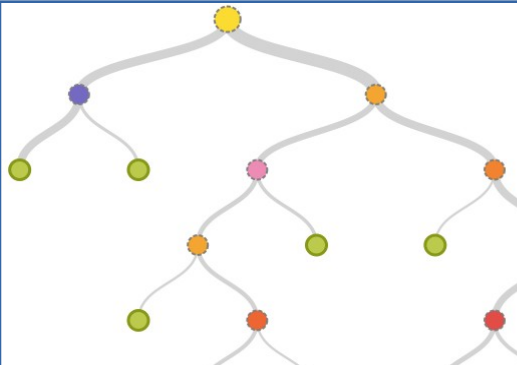
- › Proposto por Leo Breiman
- › BigML usa uma extensão do algoritmo
 - Adaptado ao formato de dataset usado

BigML usa mtree (memory)
e stree (streaming)

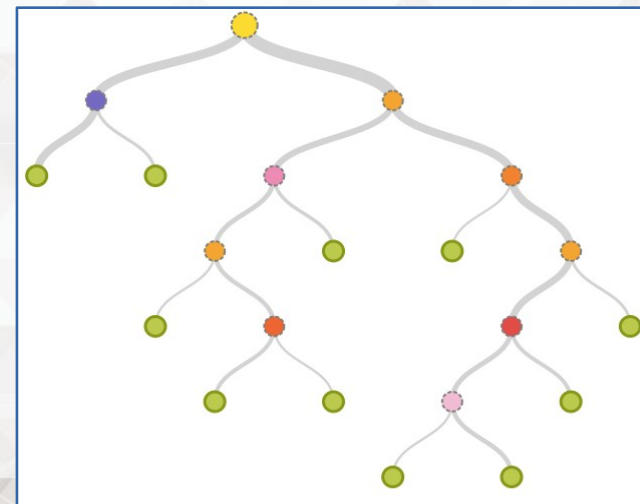
- Funcionamento

- › Split (quebra) de dados em partições
 - Cada partição maximiza o ganho de informação (classificação) ou minimiza o Mean Square Error (regressão)
 - Cada partição é associada a um predicado
 - › Baseado em um campo (ex.: saldo < 1000)
- › Processo de partição é recursivo
 - Formando uma hierarquia de partições

Decision Trees

- Scoring e splitting
 - › Para cada nó, seleciona o melhor split para cada campo, e então seleciona o melhor campo da lista
 - Pruning
 - › Poda acontece quando um novo ramo não aumenta a confiança ou diminui o erro
 - › Pode ser configurado
 - Importância de campos
 - Confiança e probabilidade
 - › Confiança (confidence) penaliza mais um baixo número de instâncias
 - Medida pessimista
 - Wilson score formula
- 
- $$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\frac{\hat{p} + \frac{1}{2n}z^2 - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{1}{n}z^2}$$



Decision Trees

Minutes Used	Last Month Bill	Support Calls	Website Visits	Churn?
104	\$103.60	0	0	No
124	\$56.33	1	0	No
56	\$214.60	2	0	Yes
2410	\$305.60	0	5	No
536	\$145.70	0	0	No
234	\$122.09	0	1	No
201	\$185.76	1	7	Yes
111	\$83.60	3	2	No

Decision Trees

- Website Visits > 0

Minutes Used	Last Month Bill	Support Calls	Website Visits	Churn?
104	\$103.60	0	0	No
124	\$56.33	1	0	No
56	\$214.60	2	0	Yes
2410	\$305.60	0	5	No
536	\$145.70	0	0	No
234	\$122.09	0	1	No
201	\$185.76	1	7	Yes
111	\$83.60	3	2	No

Decision Trees

- Minutes Used > 200

Minutes Used	Last Month Bill	Support Calls	Website Visits	Churn?
104	\$103.60	0	0	No
124	\$56.33	1	0	No
56	\$214.60	2	0	Yes
2410	\$305.60	0	5	No
536	\$145.70	0	0	No
234	\$122.09	0	1	No
201	\$185.76	1	7	Yes
111	\$83.60	3	2	No

Decision Trees

- Last Bill > \$180

Minutes Used	Last Month Bill	Support Calls	Website Visits	Churn?
104	\$103.60	0	0	No
124	\$56.33	1	0	No
56	\$214.60	2	0	Yes
2410	\$305.60	0	5	No
536	\$145.70	0	0	No
234	\$122.09	0	1	No
201	\$185.76	1	7	Yes
111	\$83.60	3	2	No

Decision Trees

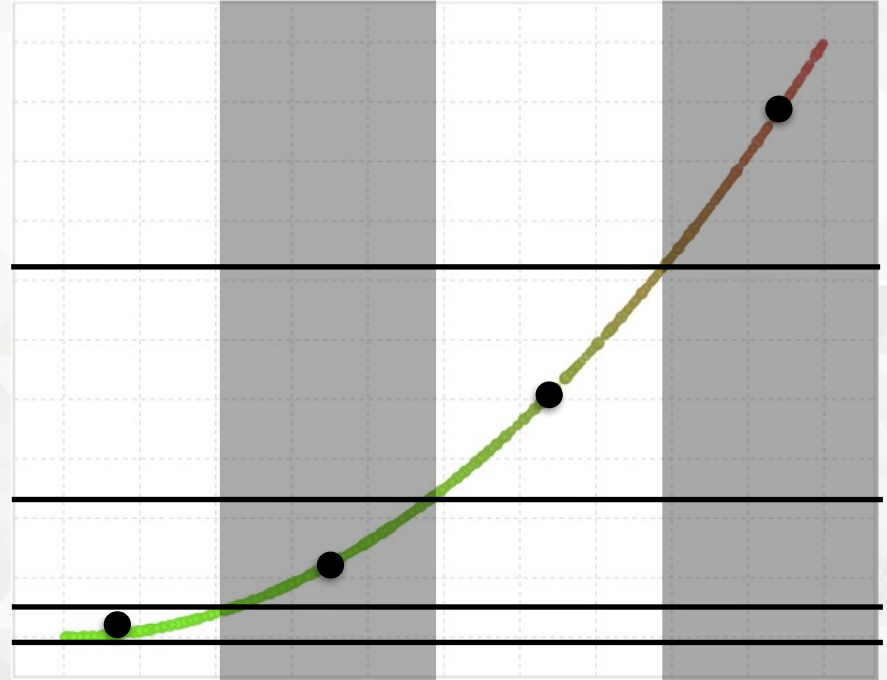
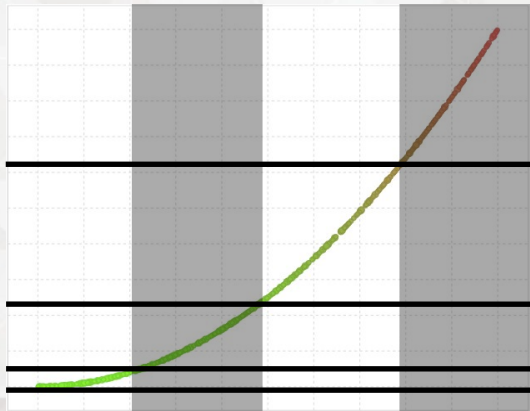
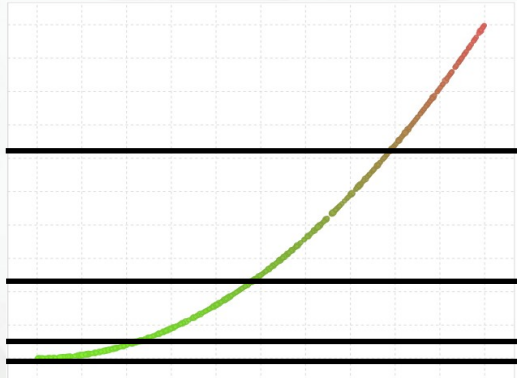
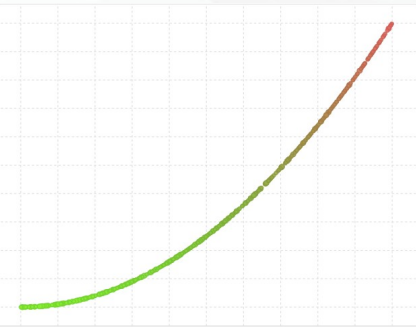
- Last Bill > \$180 e Support Calls > 0

Minutes Used	Last Month Bill	Support Calls	Website Visits	Churn?
104	\$103.60	0	0	No
124	\$56.33	1	0	No
56	\$214.60	2	0	Yes
2410	\$305.60	0	5	No
536	\$145.70	0	0	No
234	\$122.09	0	1	No
201	\$185.76	1	7	Yes
111	\$83.60	3	2	No

Porque Decision Trees

- Funciona para classificação e regressão

Regressão em DT



Porque Decision Trees

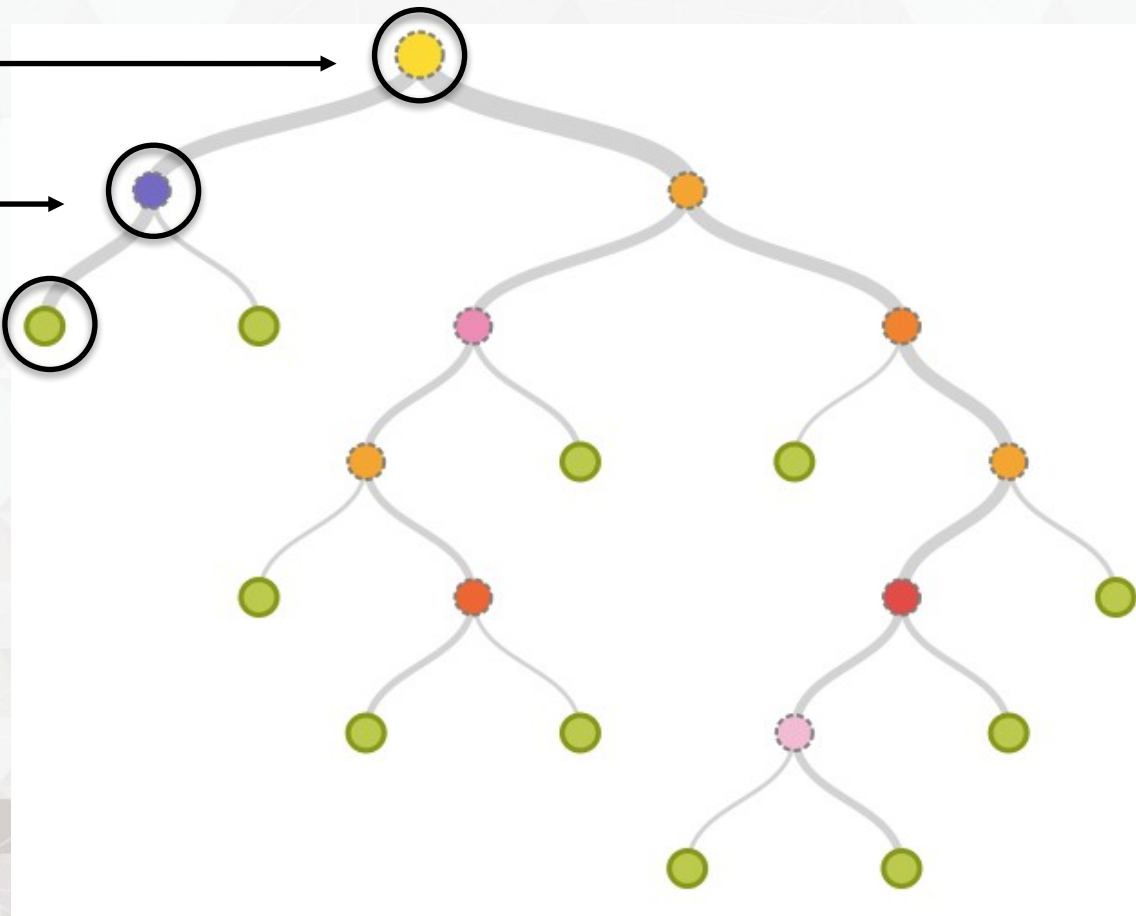
- Funciona para classificação e regressão
- Entendimento simples: splits são baseados nas features e nos seus valores
- Leve e muito rápida em tempo de predição

Predições em DT

Question 1:

Question 2:

Prediction



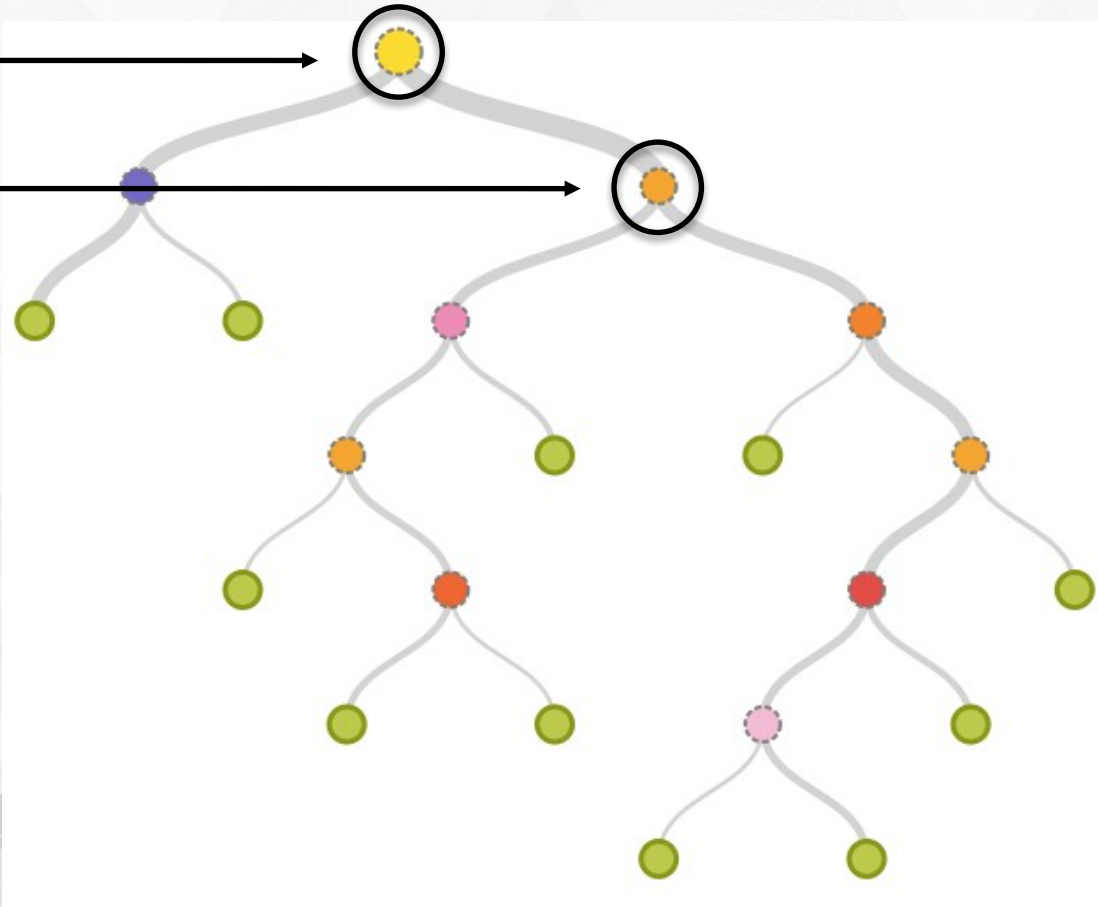
Porque Decision Trees

- Funciona para classificação e regressão
- Entendimento simples: splits são baseados nas features e nos seus valores
- Leve e muito rápida em tempo de predição
- Dados podem estar desorganizados
 - › Features inúteis são automaticamente ignoradas pelo algoritmo
 - › Trabalha com dados desnormalizados e não nivelados
 - › Trabalha com dados “missing” no treinamento e predição

Treinamento com missing

Loan Amount?

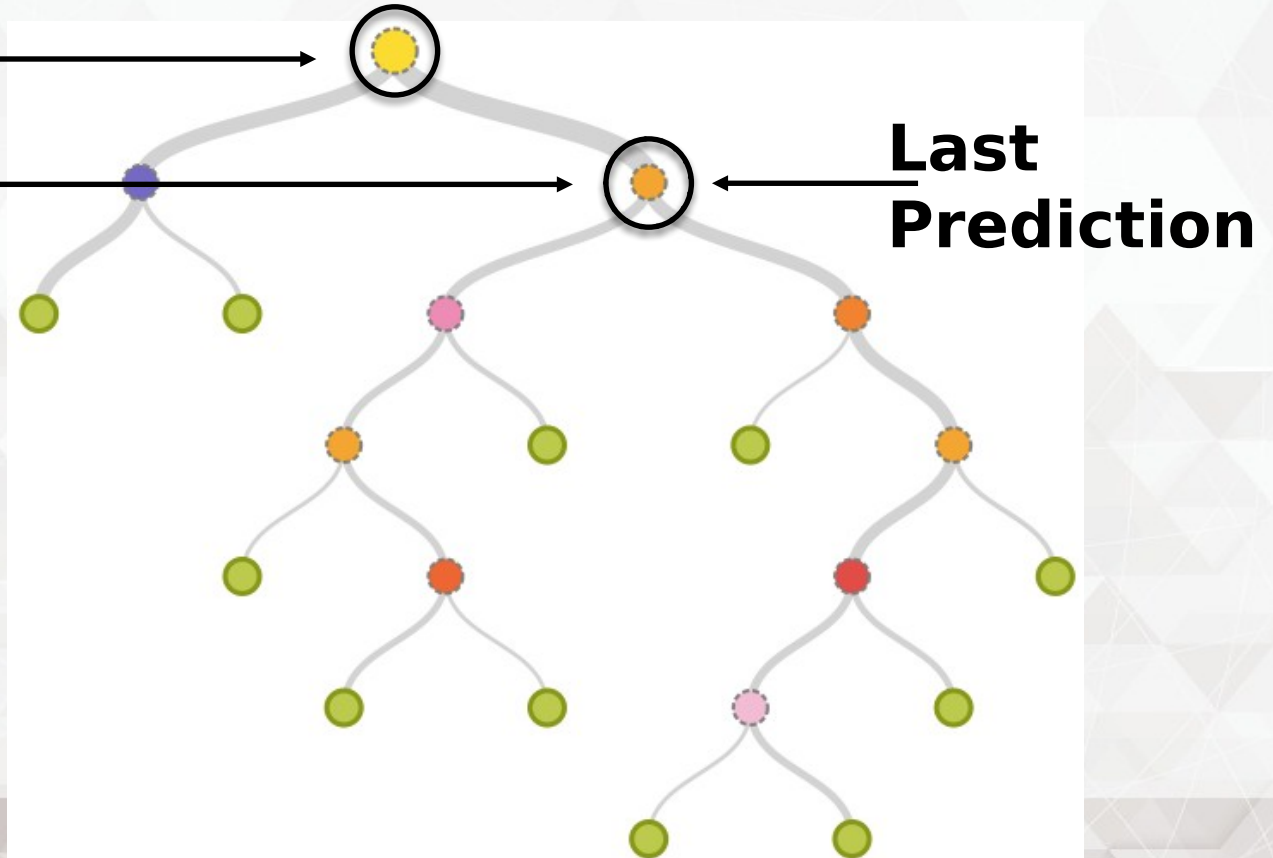
Reason Missing?



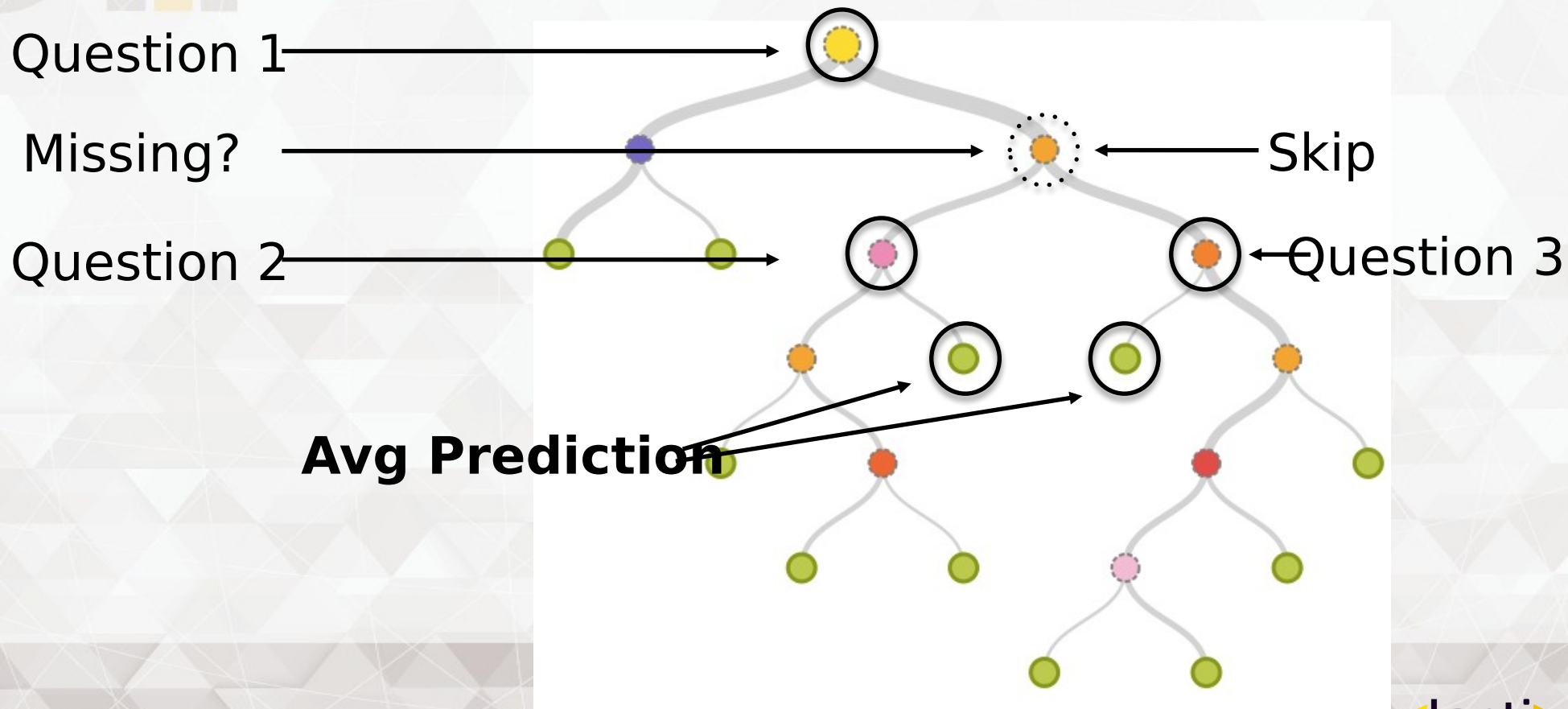
Predição com missing

Question 1 →

Missing? →



Predição com missing



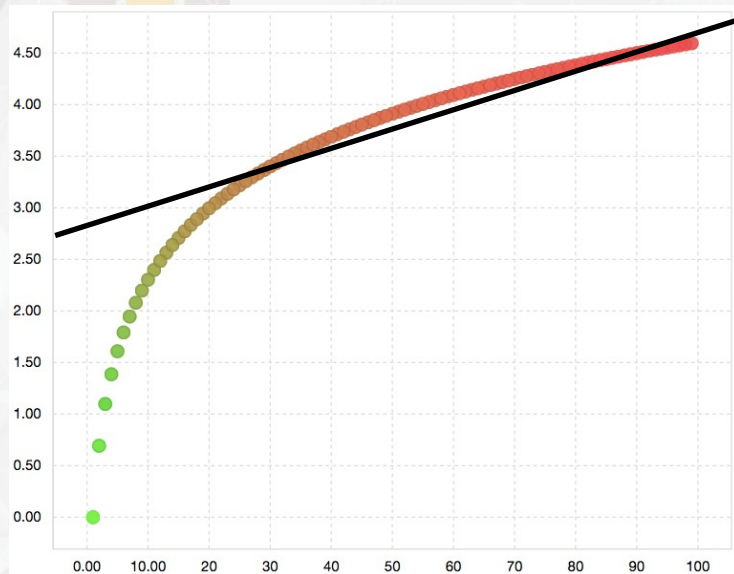
Porque Decision Trees

- Funciona para classificação e regressão
- Entendimento simples: splits são baseados nas features e nos seus valores
- Leve e muito rápida em tempo de predição
- Dados podem estar desorganizados
 - › Features inúteis são automaticamente ignoradas pelo algoritmo
 - › Trabalha com dados desnormalizados e não nivelados
 - › Trabalha com dados “missing” no treinamento e predição
 - › Resiliente com outliers
- Alto poder representacional
- Trabalha facilmente com diversos tipos de dados

Porque NÃO Decision Trees

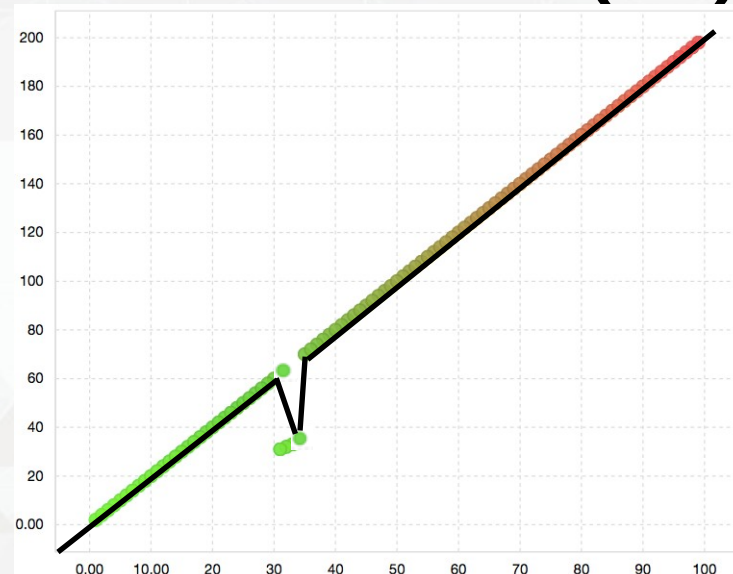
- Tem uma mínima tendência a over-fitting

Problemas em aprendizado (fit)



Under-fitting

- Modelo nunca converge
- Não captura as tendências dos dados
- Recomendação: trocar algoritmo ou features



Over-fitting

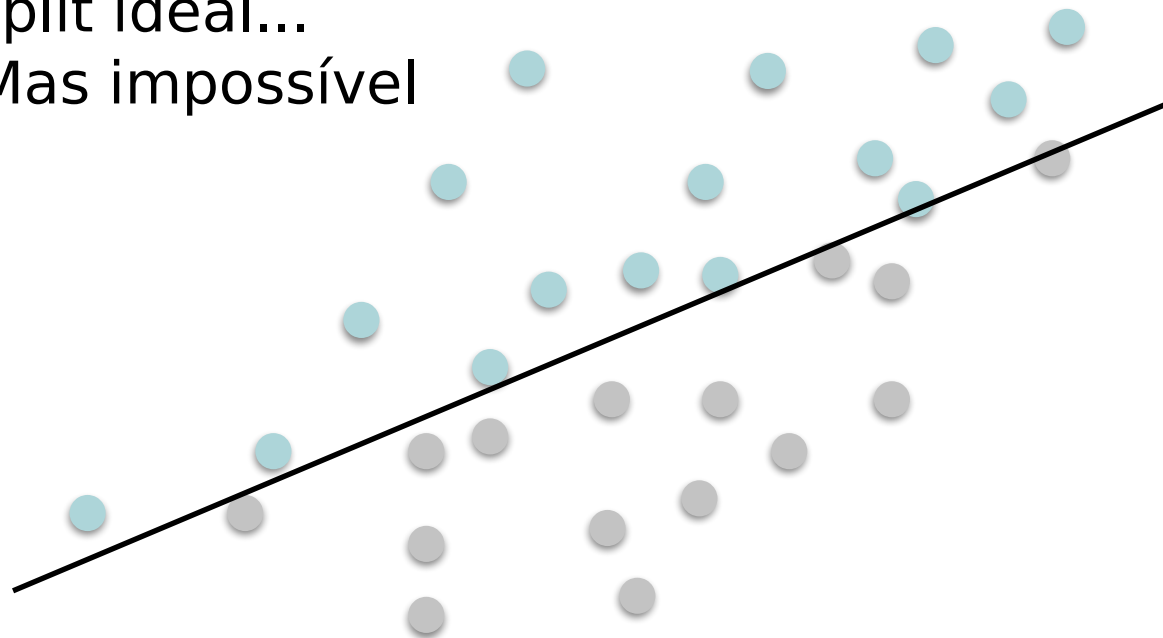
- Modelo concorda demais com os dados
- Captura ruído ou outliers nos dados
- Recomendação: trocar algoritmo ou filtrar outliers

Porque NÃO Decision Trees

- Tem uma mínima tendência a over-fitting
 - ↳ Mas isso pode ser resolvido com ensembles
 - Random forests
- O processo de splitting cria fronteiras de decisão que são perpendiculares aos eixos das features

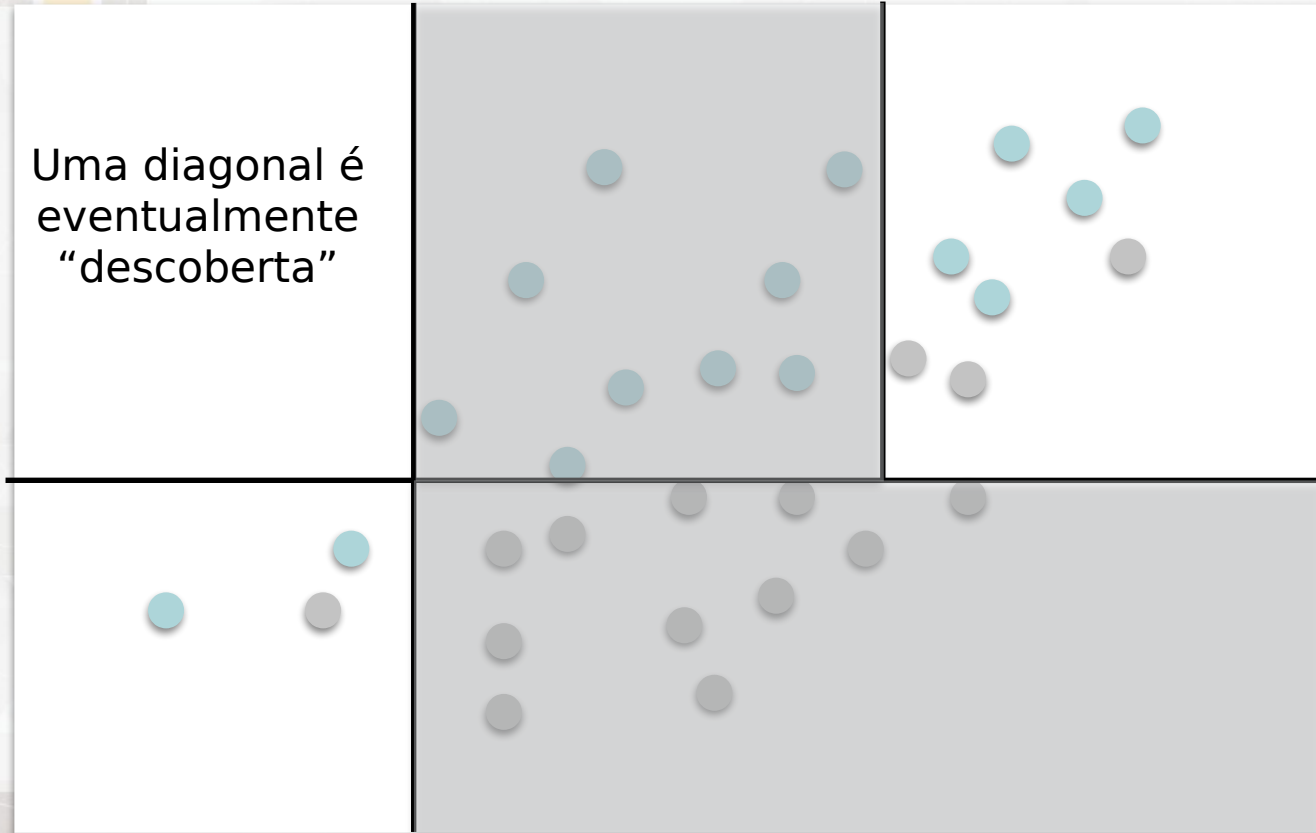
Splits paralelos aos eixos

Split ideal...
Mas impossível



Splits paralelos aos eixos

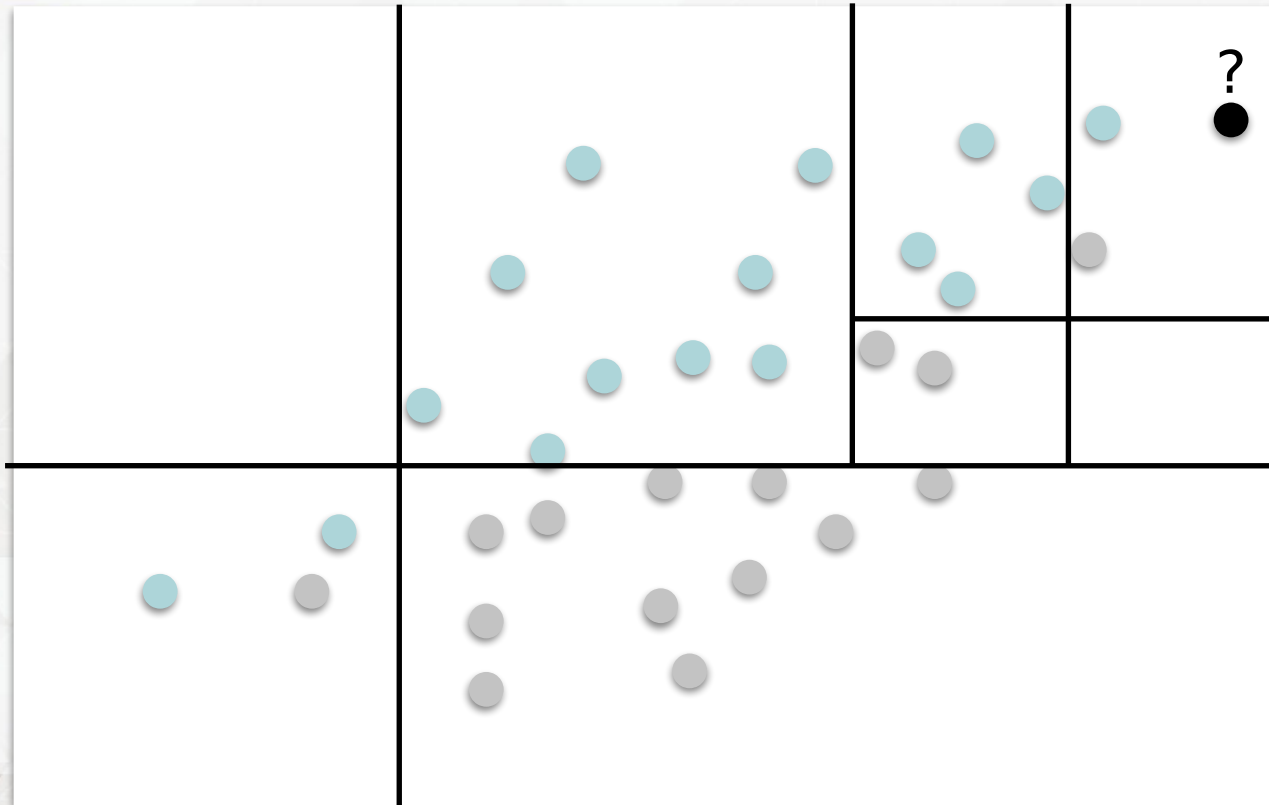
Uma diagonal é eventualmente “descoberta”



Porque NÃO Decision Trees

- Tem uma mínima tendência a over-fitting
 - › Mas isso pode ser resolvido com ensembles
 - Random forests
- O processo de splitting cria fronteiras de decisão que são perpendiculares aos eixos das features
 - › Acrescente mais dados!
- Predições fora dos dados de treinamento podem ser problemáticas

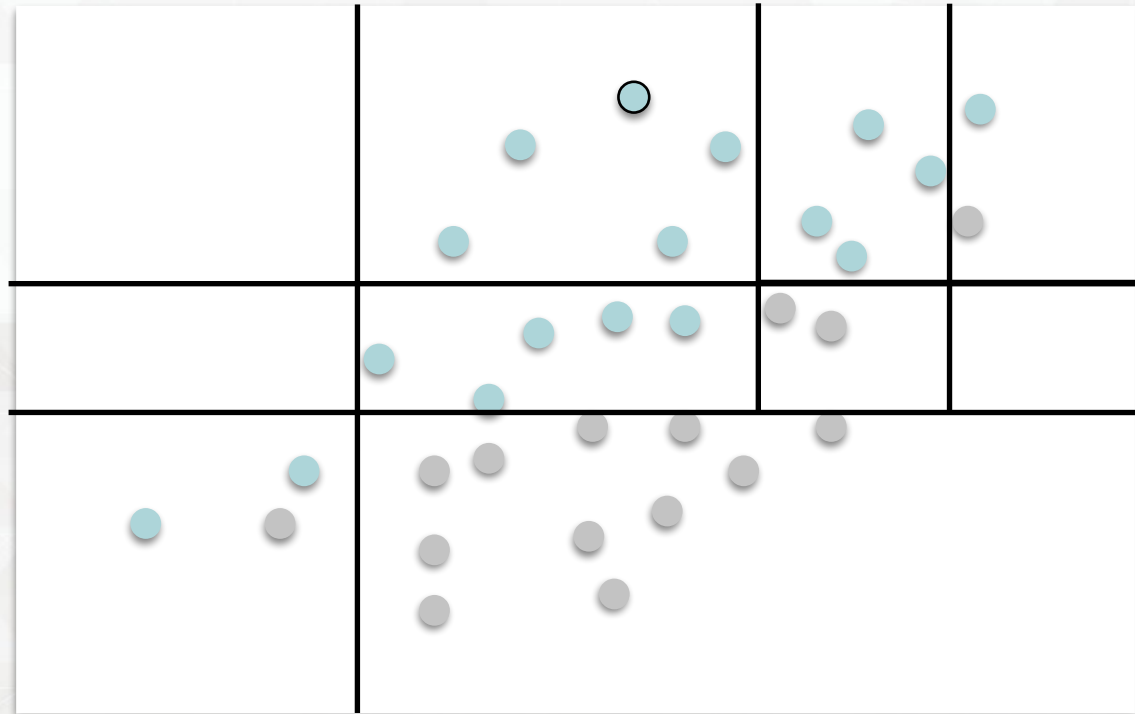
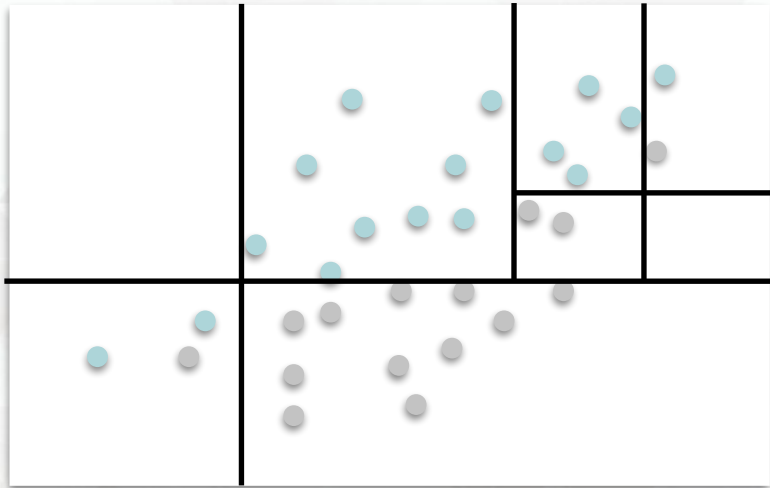
Outlier prediction



Porque NÃO Decision Trees

- Tem uma mínima tendência a over-fitting
 - › Mas isso pode ser resolvido com ensembles
 - Random forests
- O processo de splitting cria fronteiras de decisão que são perpendiculares aos eixos das features
 - › Acrescente mais dados!
- Predições fora dos dados de treinamento podem ser problemáticas
 - › Pode ser detectado com testes de “model competence”
- Pode ser sensível a pequenas alterações nos dados de treinamento

Outlier prediction



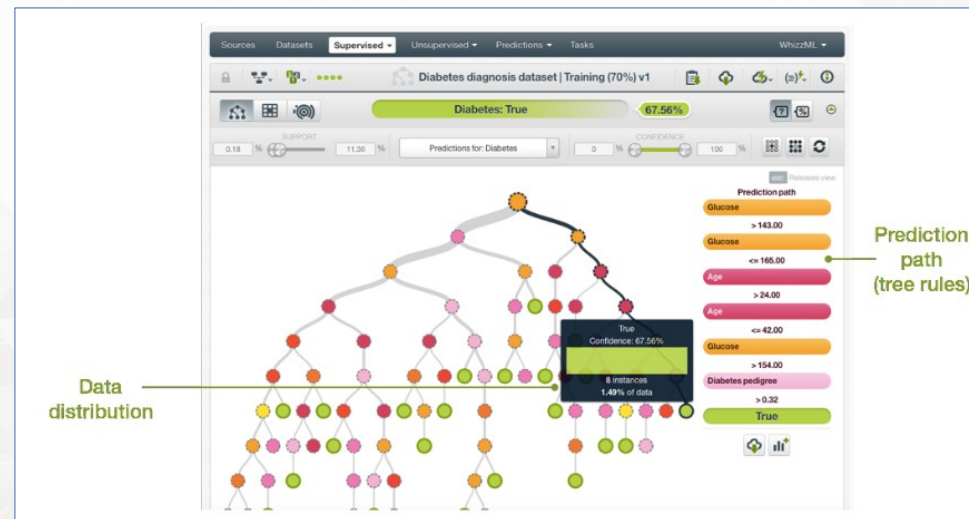
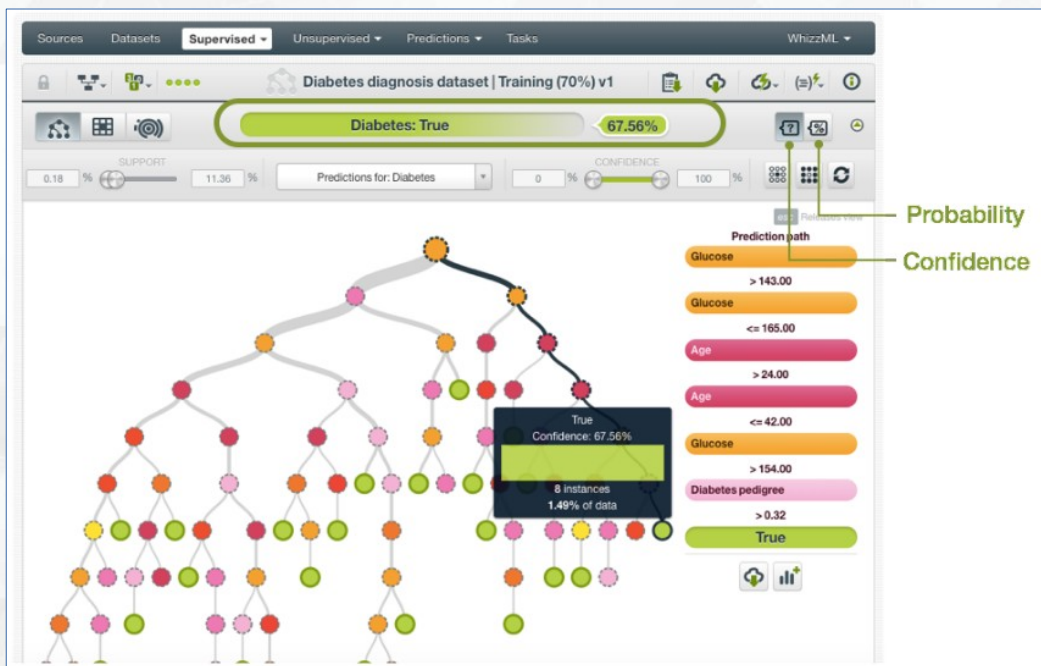
Configuração

- Objective field
- Otimizações automáticas
- Parâmetros do modelo

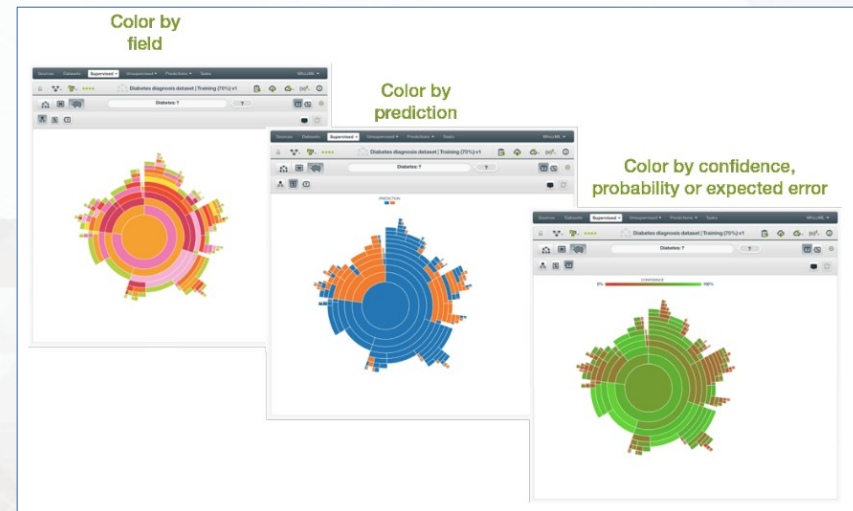
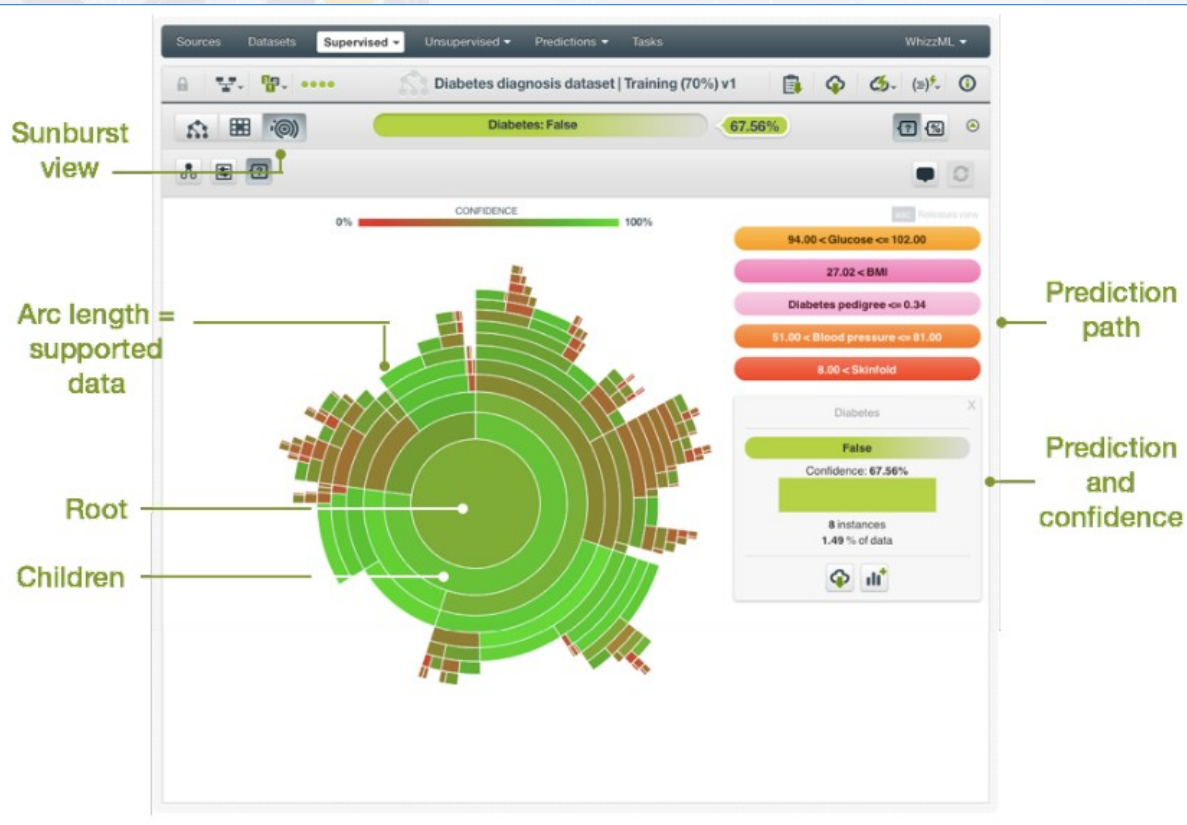
Configuração

- Pruning
 - } Smart
 - Poda mais nós com menos de 1% de instâncias
 - } Statistical
 - Considera todos os nós para poda
 - } No statistical
 - Desativa poda
- Missing splits
- Limite de nós
 - } Default de 512
- Peso de campos
 - } Balance objective
 - Balanceado automaticamente baseado no número de instâncias
 - } Manual objective
 - Objective weights
 - } Weight field
- Ordenação e shuffling
 - } Split é aplicado em linhas com ordem aleatória
- API Request Preview

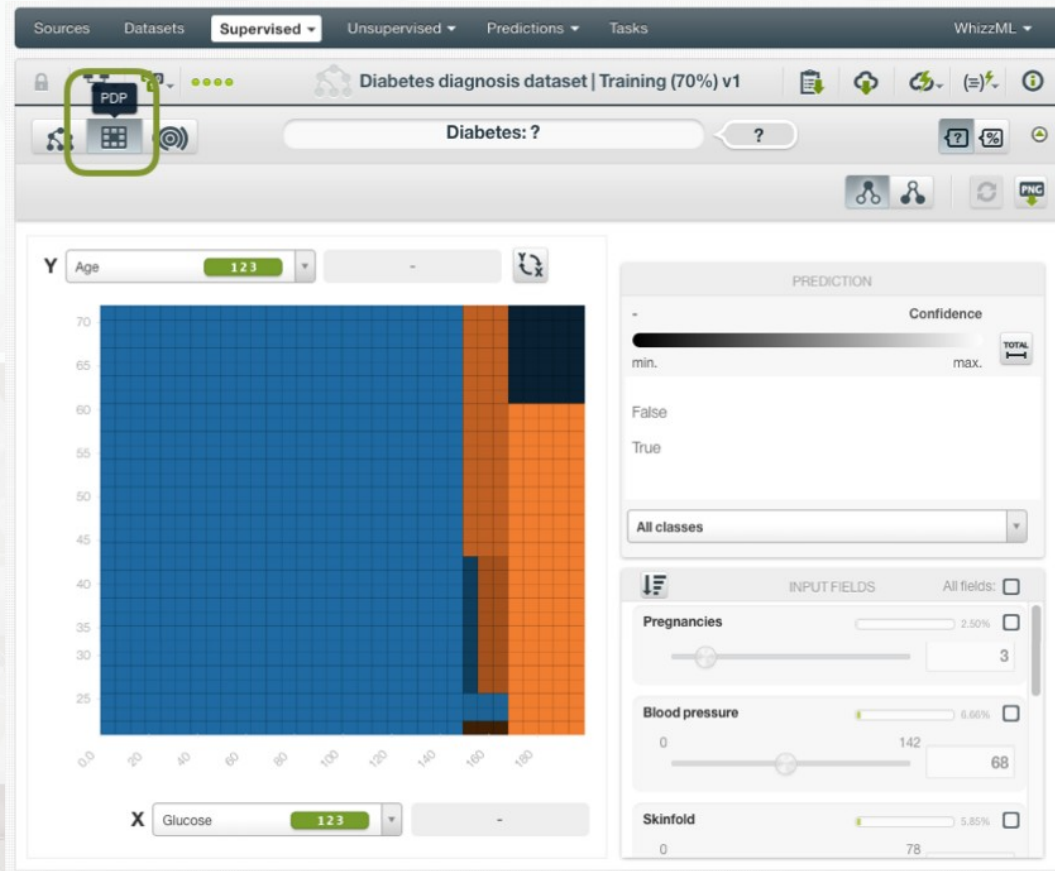
Visualização



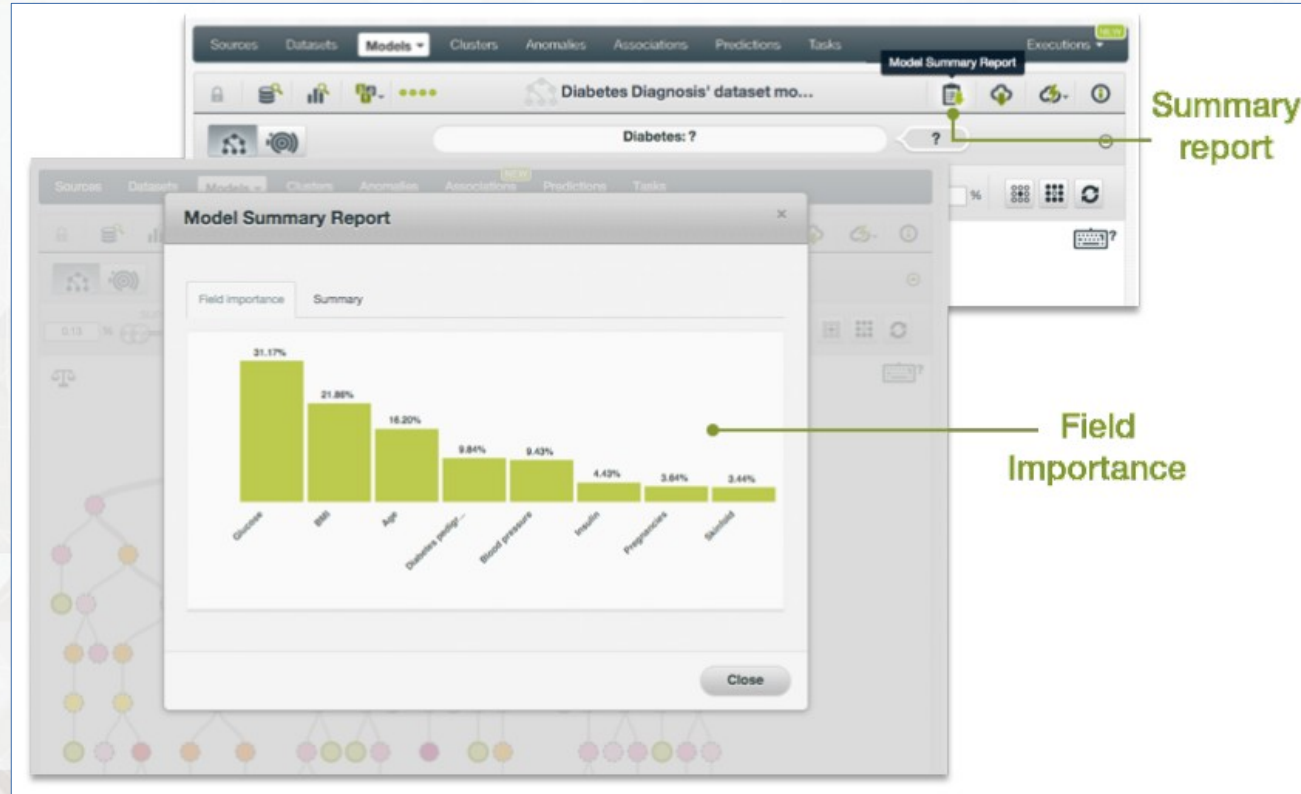
Visualização



Visualização



Summary Report



Predições

The screenshot displays a web-based machine learning interface. At the top, a navigation bar includes tabs for Sources, Datasets, Models (selected), Clusters, Anomalies, Associations, Predictions, and Tasks. A 'Scripts' button with a 'NEW' badge is on the right. Below the navigation bar, the 'Models' section is active, showing a table of models. The first model is 'iris dataset's model', which is highlighted. A context menu is open over this model, listing several actions: PREDICT BY QUESTION, PREDICT, CREATE BATCH PREDICTION, EVALUATE, VIEW DETAILS, DELETE MODEL, and MOVE TO... The table columns include Name, Type, Objective, a date (2d 6h), a size (4.5 KB), and three numerical values (0, 0, 0). The table also shows pagination controls (1 to 1 of 1 models) and a search icon.

Name	Type	Objective				
iris dataset's model		species	2d 6h	4.5 KB	0	0

The background features a repeating geometric pattern of triangles in shades of gray. At the top, there is a horizontal band with a yellow and gray geometric design, including a stylized 'U' shape. The word 'Obrigado' is centered in a large, black, sans-serif font.

Obrigado

leandro@utfpr.edu.br

<http://lapti.ct.utfpr.edu.br>

<lapti>