

Data Science
Academy

Data Science
Academy

Formação Desenvolvedor Microsoft para Data Science e Inteligência Artificial

Power BI Avançado para Análise de Dados



Capítulo 10

Machine Learning





Conteúdo

- Tipos de Análises
- Azure Machine Learning Studio
- Treinamento de um Modelo de Machine Learning
- Avaliação de um Modelo de Machine Learning
 - Métricas de Avaliação
- Deploy de um Modelo de Machine Learning
- Obtenção de Predições
 - via POSTMAN, Excel e Python



Conteúdo

- Integração do Power BI com Azure Machine Learning Studio utilizando linguagem “M”
- Atualizações de Dados no Power BI Service
- Sobre os recursos “AutoML” e “Insights da IA” disponíveis no Power BI
- Visual “Influenciadores-Chave”



Alinhando expectativas

- Aprenderemos a treinar modelos de Machine Learning e a realizar previsões a partir deste modelo.
- Não será foco deste capítulo compreender como funcionam os algoritmos de Machine Learning e como o processo de treinamento é realizado. Estes assuntos são discutidos em detalhes no curso “[Machine Learning](#)” da Formação Cientista de Dados.





Tipos de Análises

- Descritiva
- Diagnóstica
- Preditiva
- Prescritiva



Análise Descritiva

- Analisa dados históricos com o objetivo de responder perguntas no passado:
 - Qual foi o faturamento no último ano?



Análise Diagnóstica

- Analisa dados históricos para compreender o motivo que levou a um determinado resultado:
 - Por que o faturamento no último ano foi de 2 milhões de reais?
- Através de recursos de drill down, hierarquias, análises de cenários, gráficos com comparações ao longo do tempo, relatórios (dashboards) podem auxiliar o usuário a encontrar a resposta:
 - O faturamento foi de 2 milhões no último ano, 30% menor do que o ano anterior devido ao aumento dos valores de insumos e do fechamento de lojas no País ABC devido a piora de sua economia.




Análise Preditiva

- Analisa dados históricos para responder perguntas no futuro:
 - Qual será o faturamento no próximo ano?
 - Quem será o melhor cliente?
 - Quanto venderei no Brasil?
 - As perdas serão inferiores a 10%?



Análise Preditiva

- Qual a probabilidade do paciente morrer no próximo mês?
 - Variáveis preditoras:
 - Sexo
 - Idade
 - Raça
 - Pressão Arterial, Glicemia, Colesterol
 - Realiza atividades físicas?
 - Fuma?
 - Variável-alvo:
 - Morrerá no próximo ano?  São necessários os dados históricos!



Análise Prescritiva

- Realiza predições baseadas em ações que podem ser tomadas.
 - O faturamento no próximo ano será de 3 milhões se no próximo mês abrimos 3 lojas no Canadá e 2 nos EUA.
- Além de algoritmos de Machine Learning, pode incorporar regras de negócio e outras técnicas, além de utilizar-se de dados históricos, dados em tempo real, outras previsões, etc.

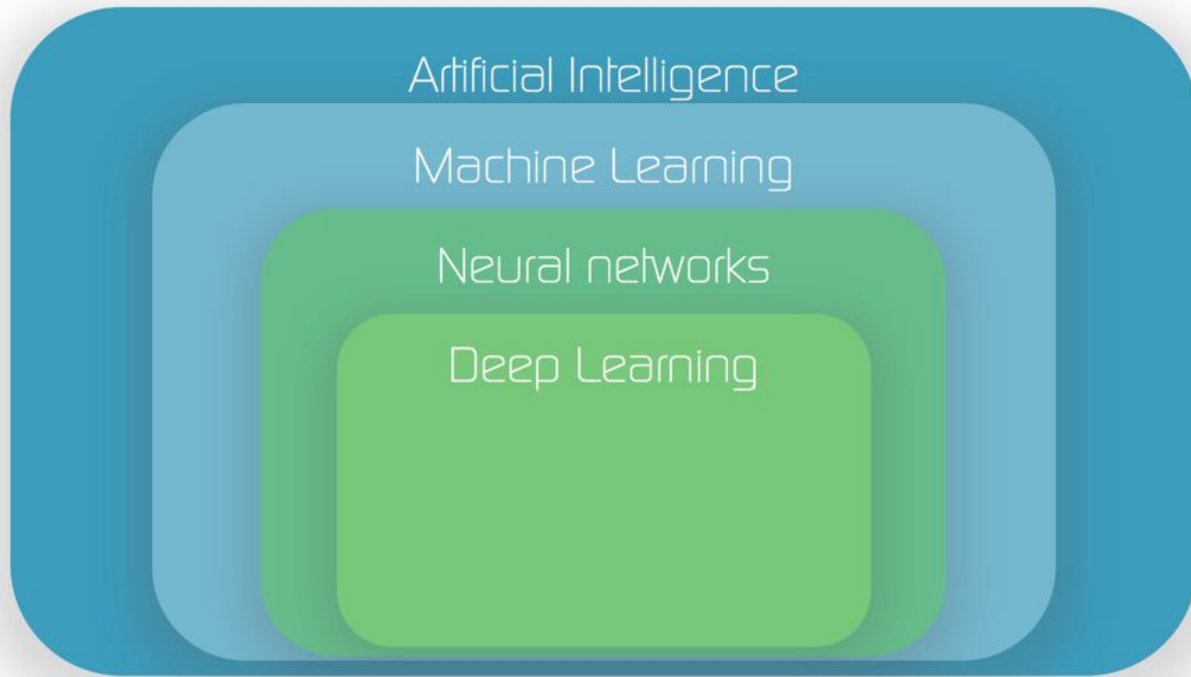


Tipos de Análises

- Descritiva
 - O que aconteceu?
- Diagnóstica
 - Por que aconteceu?
- Preditiva
 - O que irá acontecer?
- Prescritiva
 - O que acontecerá se eu tomar determinada ação?
 - Quais ações devo tomar para que determinada situação aconteça?



Inteligência Artificial e Machine Learning





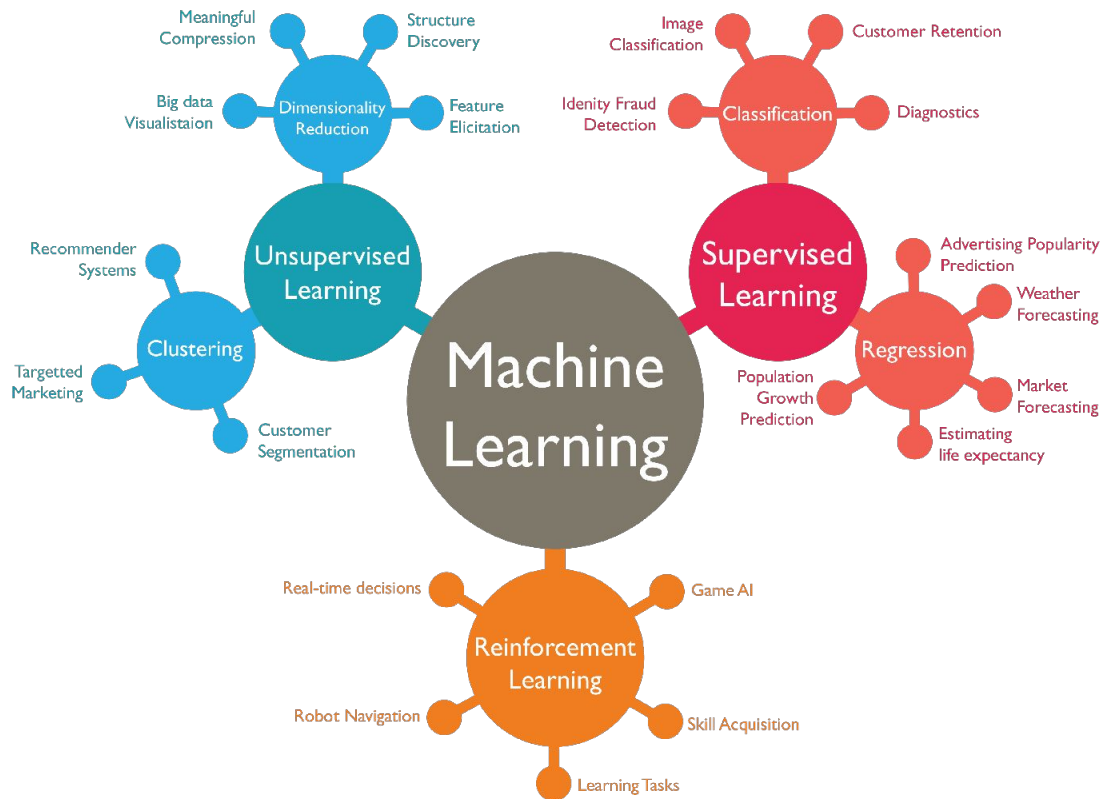
Machine Learning

“A **aprendizagem automática** ou **aprendizado de máquina** é um sub-campo da **inteligência artificial** dedicado ao desenvolvimento de algoritmos e técnicas que permitem ao computador aprender *sem serem explicitamente programados para isto*” (Wikipedia).





Algoritmos de Machine Learning





Regressão

- Quando a saída do modelo - o resultado da predição - é um número.
- Exemplos: peso, altura, valor faturado, etc.





Classificação

- Quando a saída do modelo - o resultado da predição - é uma classe.
- Exemplos: Paciente: doente/não doente; qualidade do material (ruim, regular, bom), etc.





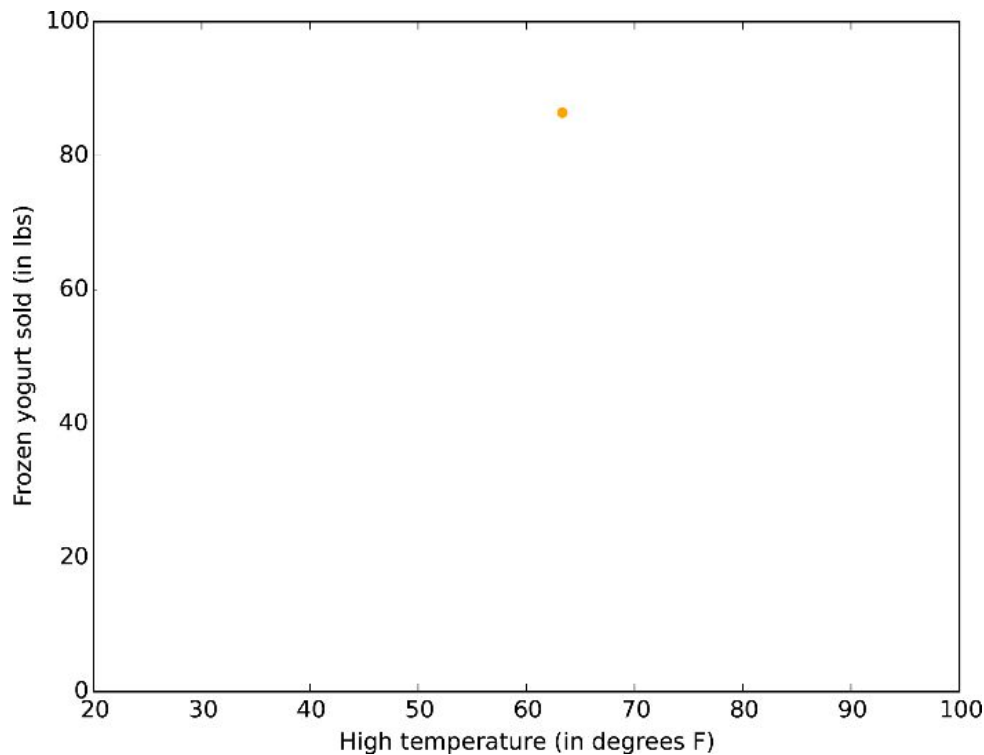
Detecção de anomalias

- Geralmente aplicado quando as classes de um dataset são “desbalanceadas”, por exemplo detecção de fraudes.
- Como a maioria das transações são consideradas “normais”, pode ser difícil compreender o que é uma transação “fraudulenta”, desta forma pode-se compreender o que é uma transação “normal” e classificar como “fraudulenta” os casos que forem significativamente diferentes de “normais”.





Regressão

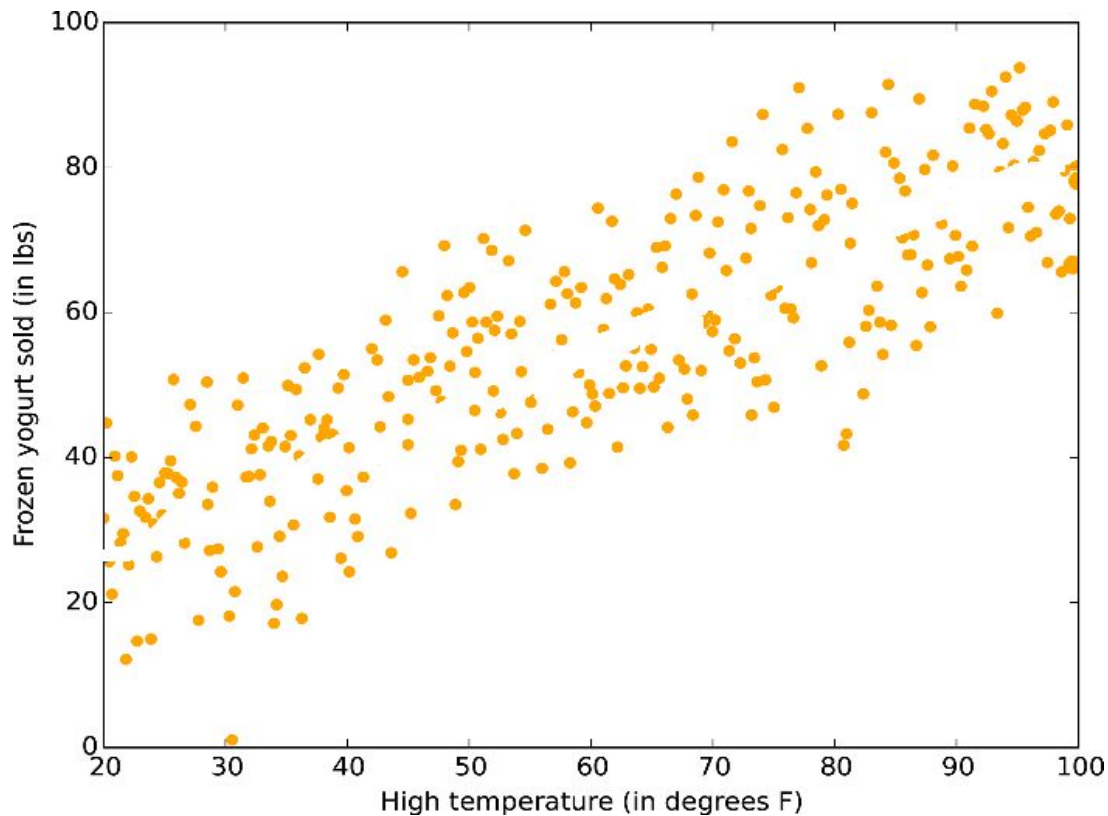


<https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice>





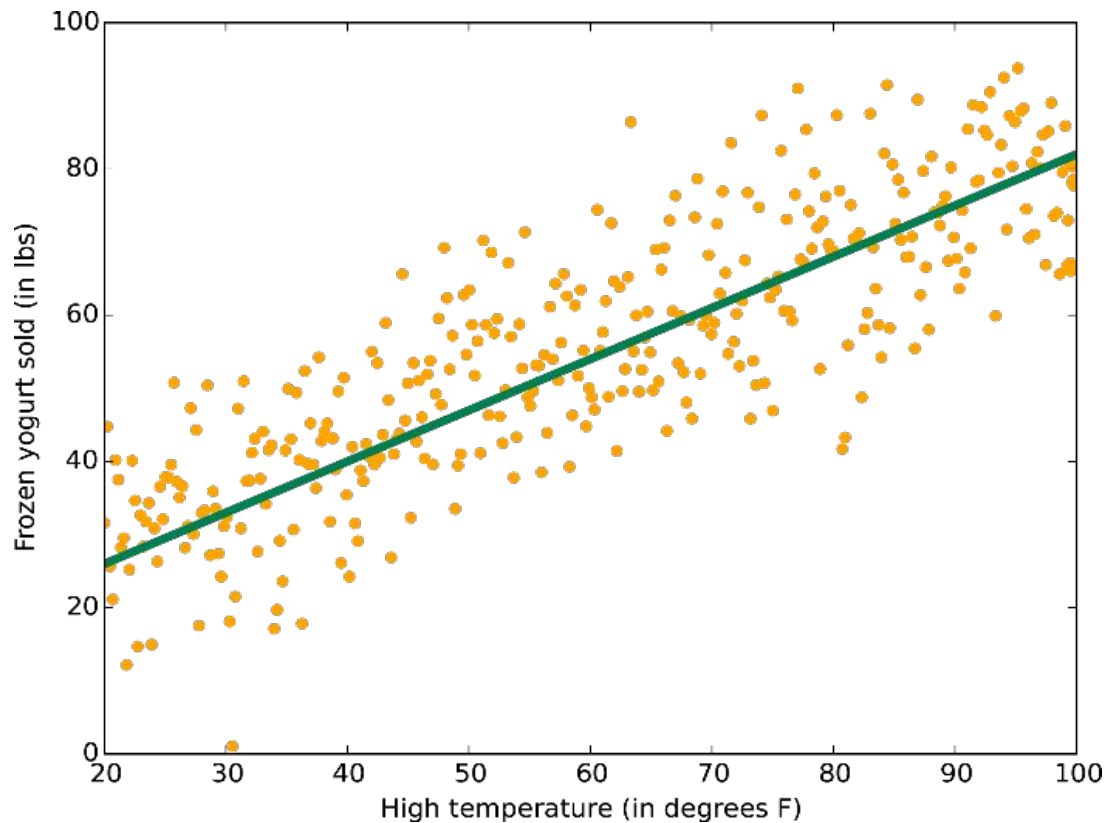
Regressão





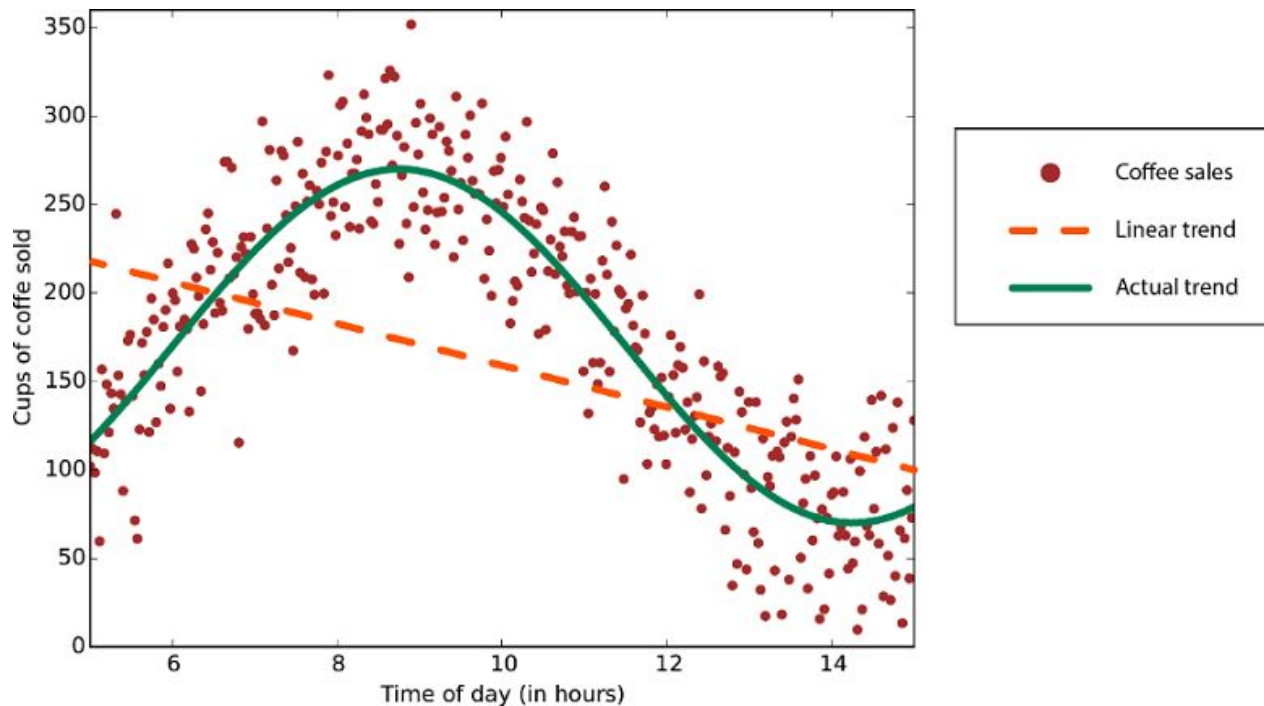
Regressão

Regressão Linear $y = ax + b$



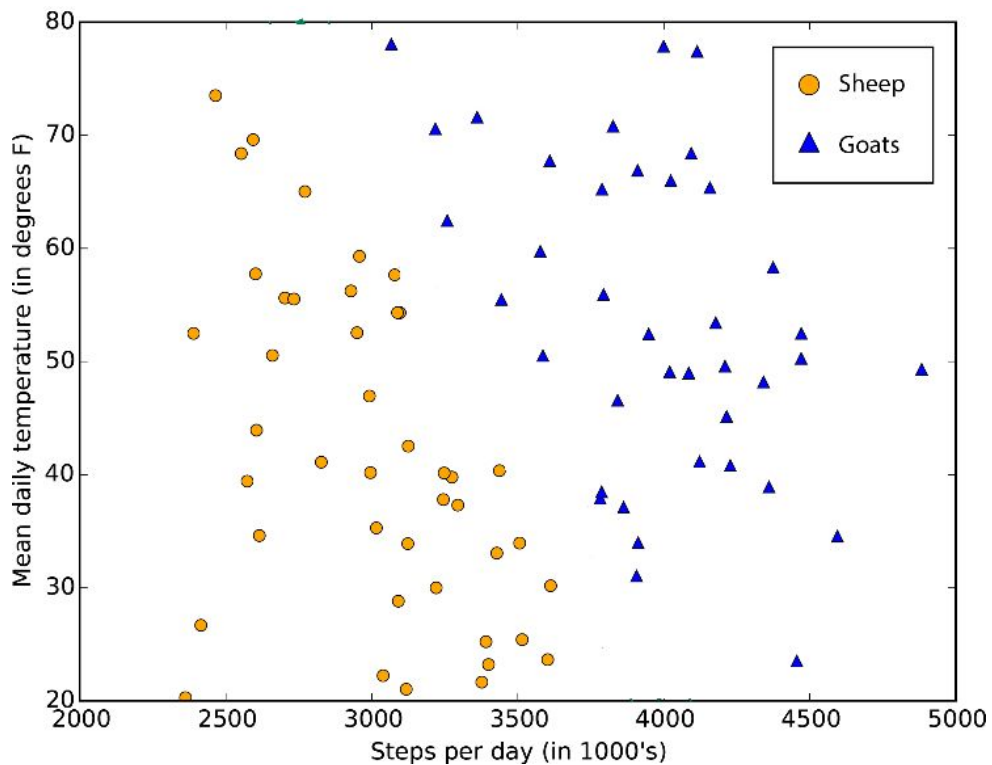


Regressão

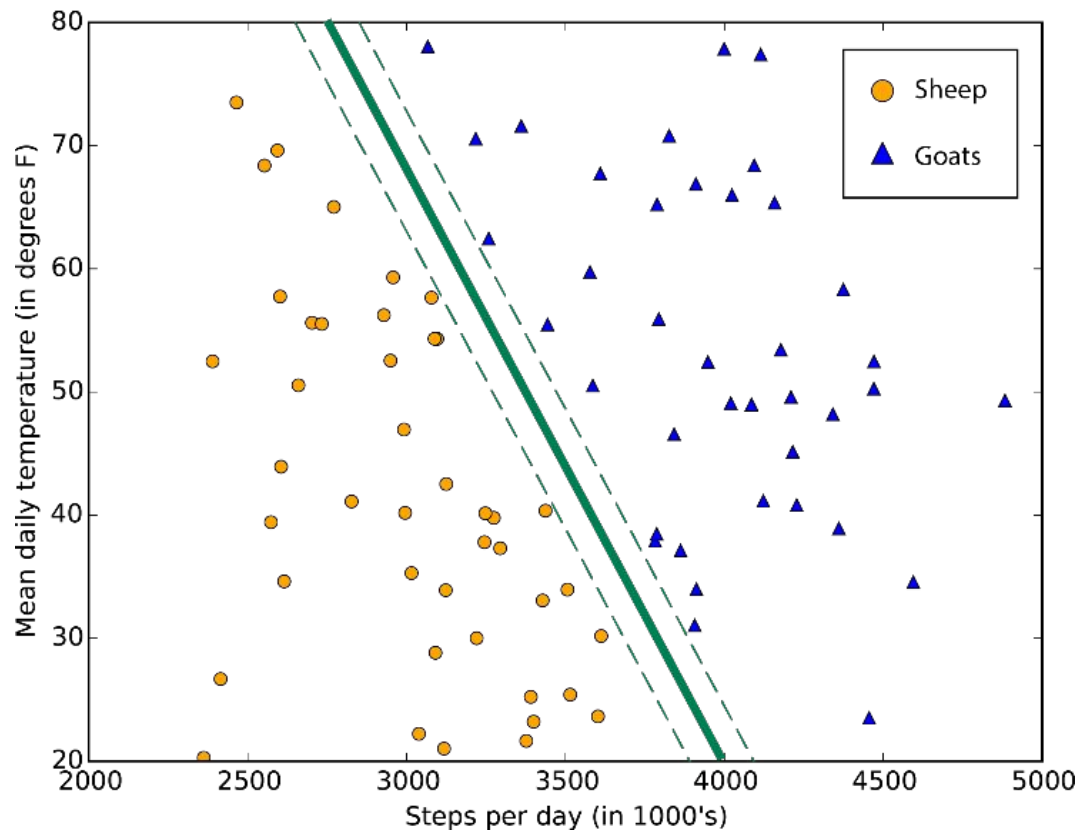




Classificação

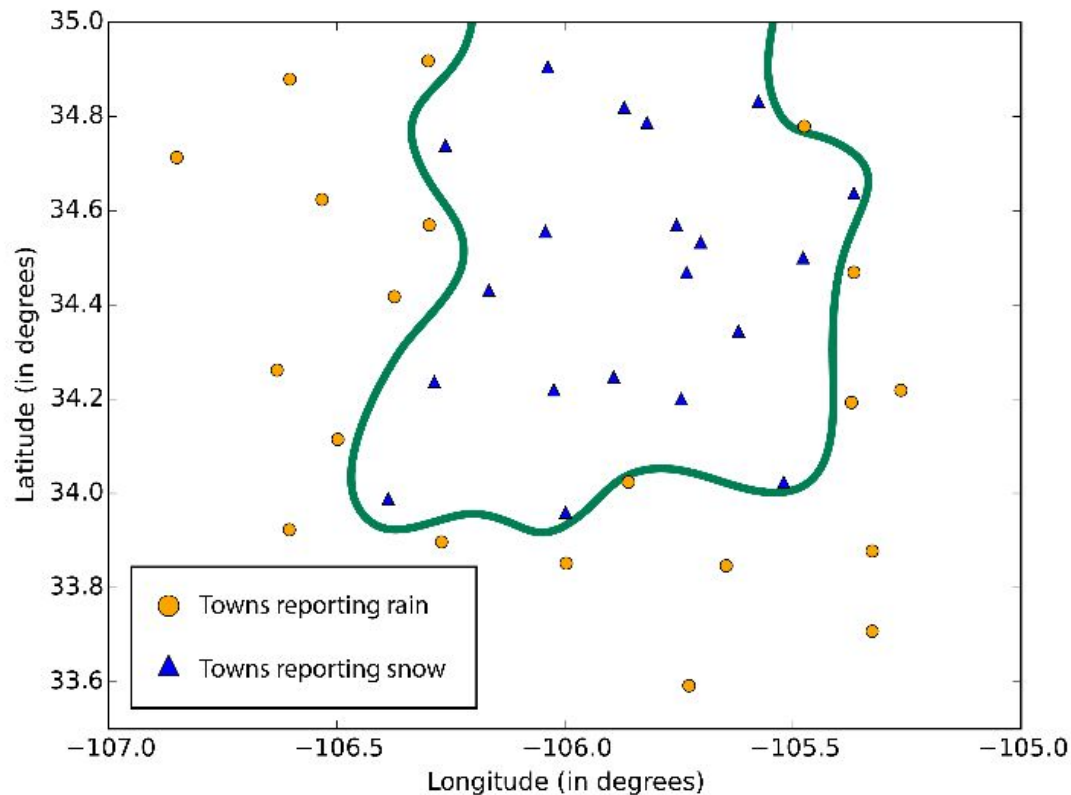


Classificação



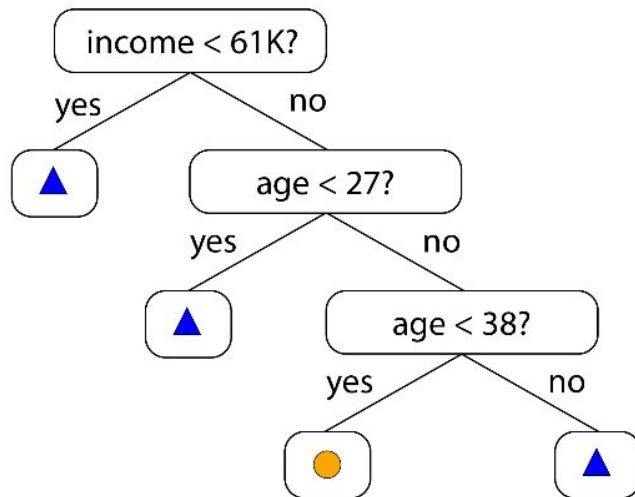
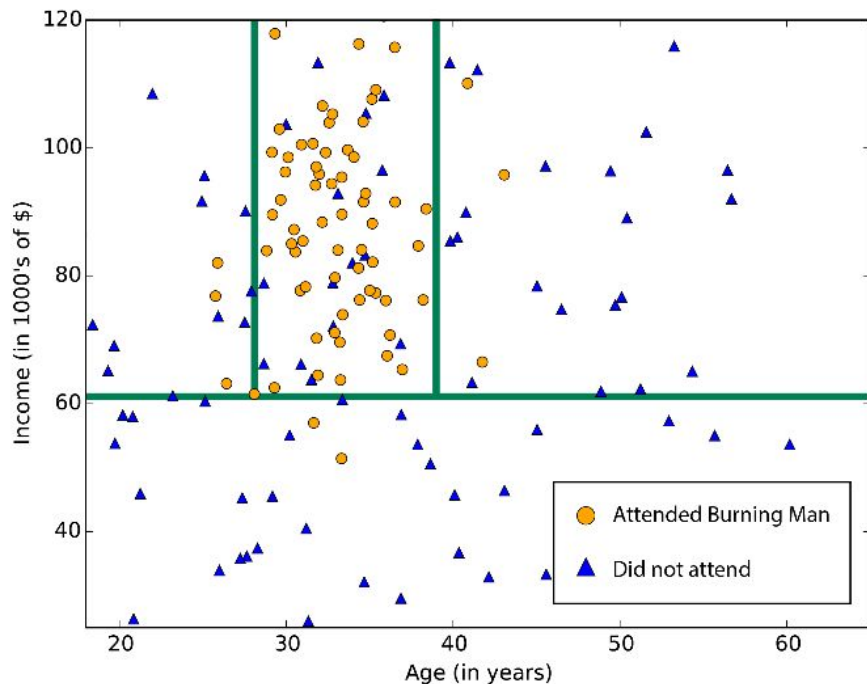


Classificação



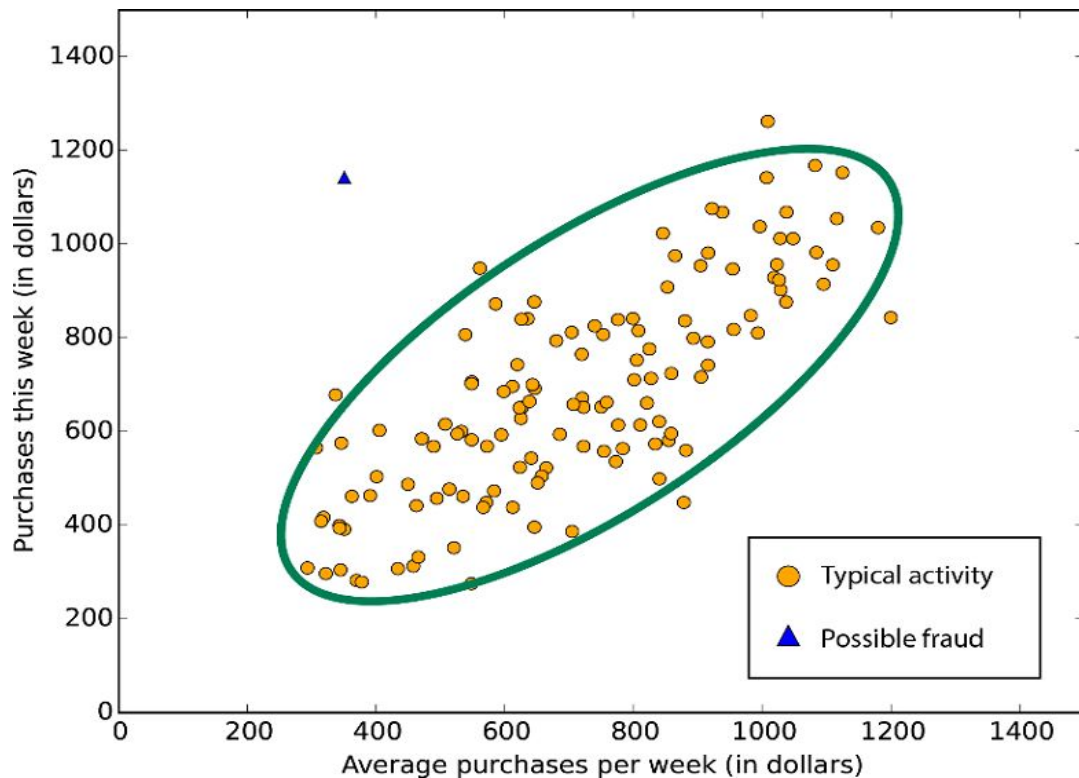


Classificação





Detecção de Anomalias





Machine Learning Workflow

1. Compreender o problema
2. Preparar os dados
 - a. Extração, Carga, Transformação (ETL)
3. Treinar o modelo
4. Avaliar o modelo
5. Implantar o modelo





Aprendizagem Supervisionada

- **Treinamento:** O algoritmo é treinado a partir de um conjunto de dados rotulados - que possui as variáveis de entrada e a **variável-alvo (a variável que estamos tentando prever)**, gerando um **modelo**.
- **Inferência:** O modelo recebe novos dados (variáveis de entrada) e realiza a inferência, dando como saída a predição para aquele conjunto de dados de entrada.





Métricas para Avaliação de Modelos de Machine Learning





Confusion Matrix (Matriz de Confusão)

n = 165	Previsto: Não	Previsto: Sim
Observado: Não	50	10
Observado: Sim	5	100

n = número de observações

Considerando “X” classes possíveis de saída, a matriz de confusão terá dimensões X por X





Confusion Matrix (Matriz de Confusão)

n = 165	Previsto: Não	Previsto: Sim
Observado: Não	50	10
Observado: Sim	5	100





Confusion Matrix (Matriz de Confusão)

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





Acurácia

Acurácia: acertos / tentativas

$$\text{Acurácia} = (TP + TN) / n = 150 / 165 = 0.91 = 91\%$$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





Taxa de Erro

Taxa de Erro: erros / tentativas

Taxa de Erro: $(FP + FN) / n = 15 / 165 = 0.09 = 9\%$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





Acurácia em datasets desbalanceados

Somente 2% da população tem uma determinada doença e esta é a mesma proporção da nossa amostra.

O modelo pode “entender” que a forma mais fácil de acertar um resultado é classificar todas as instâncias como **negativo**.

Considere um dataset com 100 instâncias. Qual a acurácia do modelo se nós classificarmos todas as instâncias como **negativo**?





Acurácia em datasets desbalanceados

n = 100	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 98	FP = 0	98
Observado: Sim	FN = 2	TP = 0	2
	100	0	





Precisão

Precisão: Das instâncias classificadas como positivas, quantas eram positivas

$$TP / (TP + FP) = 100 / (100 + 10) = 0.909$$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





Sensibilidade (ou Recall)

Sensibilidade: Das instâncias positivas, quanto eram positivas?

$$TP / (TP + FN) = 100 / (100 + 5) = 0.952$$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





Especificidade

Especificidade: Das instâncias não-positivas, quantas eram não-positivas?

$$TN / (TN + FP) = 50 / (50 + 10) = 0.83$$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	





F1-Score

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$2 \cdot ((0.909 \cdot 0.952) / (0.909 + 0.952))$$

$$2 \cdot (0.865 / 1.861)$$

$$2 \cdot 0.464$$

$$\mathbf{F1 = 0.928}$$

n = 165	Previsto: Não	Previsto: Sim	
Observado: Não	TN = 50	FP = 10	60
Observado: Sim	FN = 5	TP = 100	105
	55	110	



Obrigado.

Tenha uma excelente jornada de aprendizagem!



Data Science
Academy