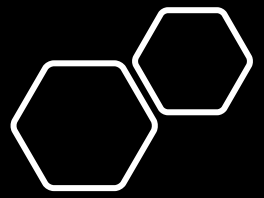
# Machine Learning no Spark



# Bibliotecas

- spark.mllib
- spark.ml

- ML baseado em RDD está descontinuado
- Implementações todas em DataFrames

Variáveis Independentes

Variável Dependente

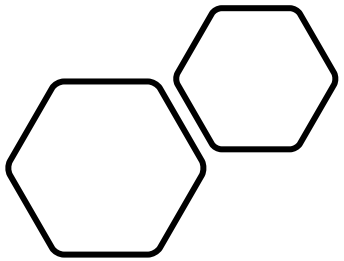
| CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|
| 619 | France | Female | 42 | 2 | 0 | 1 | 1 | 1 | 10134888 | 1 |
| 608 | Spain | Female | 41 | 1 | 8380786 | 1 | 0 | 1 | 11254258 | 0 |
| 502 | France | Female | 42 | 8 | 1596608 | 3 | 1 | 0 | 11393157 | 1 |
| 699 | France | Female | 39 | 1 | 0 | 2 | 0 | 0 | 9382663 | 0 |
| 850 | Spain | Female | 43 | 2 | 12551082 | 1 | 1 | 1 | 790841 | 0 |
| 645 | Spain | Male | 44 | 8 | 11375578 | 2 | 1 | 0 | 14975671 | 1 |
| 822 | France | Male | 50 | 7 | 0 | 2 | 1 | 1 | 100628 | 0 |
| 376 | Germany | Female | 29 | 4 | 11504674 | 4 | 1 | 0 | 11934688 | 1 |
| 501 | France | Male | 44 | 4 | 14205107 | 2 | 0 | 1 | 749405 | 0 |
| 684 | France | Male | 27 | 2 | 13460388 | 1 | 1 | 1 | 7172573 | 0 |
| 528 | France | Male | 31 | 6 | 10201672 | 2 | 0 | 0 | 8018112 | 0 |
| 497 | Spain | Male | 24 | 3 | 0 | 2 | 1 | 0 | 7639001 | 0 |
| 476 | France | Female | 34 | 10 | 0 | 2 | 1 | 0 | 2626098 | 0 |
| 549 | France | Female | 25 | 5 | 0 | 2 | 0 | 0 | 19085779 | 0 |
| 635 | Spain | Female | 35 | 7 | 0 | 2 | 1 | 1 | 6595165 | 0 |
| 616 | Germany | Male | 45 | 3 | 14312941 | 2 | 0 | 1 | 6432726 | 0 |
| 653 | Germany | Male | 58 | 1 | 13260288 | 1 | 1 | 0 | 509767 | 1 |
| 549 | Spain | Female | 24 | 9 | 0 | 2 | 1 | 1 | 1440641 | 0 |
| 587 | Spain | Male | 45 | 6 | 0 | 1 | 0 | 0 | 15868481 | 0 |
| 726 | France | Female | 24 | 6 | 0 | 2 | 1 | 1 | 5472403 | 0 |
| 732 | France | Male | 41 | 8 | 0 | 2 | 1 | 1 | 17088617 | 0 |
| 636 | Spain | Female | 32 | 8 | 0 | 2 | 1 | 0 | 13855546 | 0 |

PROF. FERNANDO AMARAL
www.datascientist.com.br
APACHE Spark

# Tradicionalmente

- Variáveis Independentes são colunas distintas
- Variável Dependente: outra coluna

# No Spark

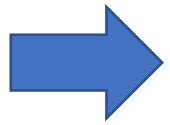


- Normalmente todas as variáveis independentes devem compor uma mesma coluna
- Cria-se um vetor único, que é adicionado em nova coluna no DataFrame

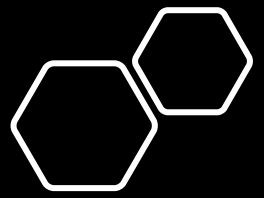
# One HotEncoding

- Machine Learning Suporta Apenas Números
- Atributos Categóricos devem ser Transformados

| EstadoCivil |
|---|
| Casado |
| Solteiro |
| Divorciado |
| Casado |
| Solteiro |
| Casado |
| Solteiro |
| Casado |
| Casado |
| Solteiro |
| Casado |
| Solteiro |
| Divorciado |
| Casado |
| Solteiro |
| Casado |
| Casado |
| Casado |
| Solteiro |



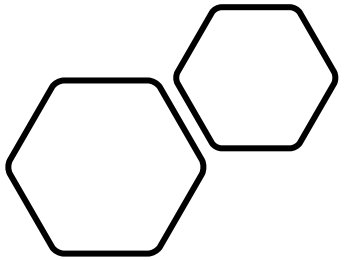
| Casado | Solteiro | Divorciado |
|---|---|---|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |



# One HotEncoding

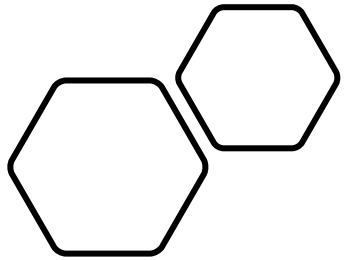
- Se o atributo tiver muitos valores, muitas colunas serão criadas
- Spark permite o uso de matriz esparsa
- Muitos valores zero que não são registrados

# Formulas no R



- R permite definir modelo através de fórmula
- [variável dependente] ~ [variável independentes]
- Ponto define todas os atributos – variável dependente
- Spark implemente Rformula
  - Aplica One HotEncoding e combina variáveis independentes em uma única coluna

# Pipelines



- Transformer: Transforma um DF em outro DF
- Estimator: Fit em DF para produzir um Transformer
- Pipeline: conecta Transformers e Estimators para Produzir modelo
- Parâmetros: Transformers e Estimators compartilham uma Api para definir parâmetros