Kazuo Teramoto.
Research & Development

Machine learning além do buzzword.

idwall

Quem sou eu?

Faço parte do *Squad* de *Docs & Biometrics* da idwall. Junto do *squad* faço pesquisa e implementação na área de **visão computacional** e **processamento de texto**. Desenvolvo soluções para os nossos produtos de ***validação de documentos*** e **biometria facial**.

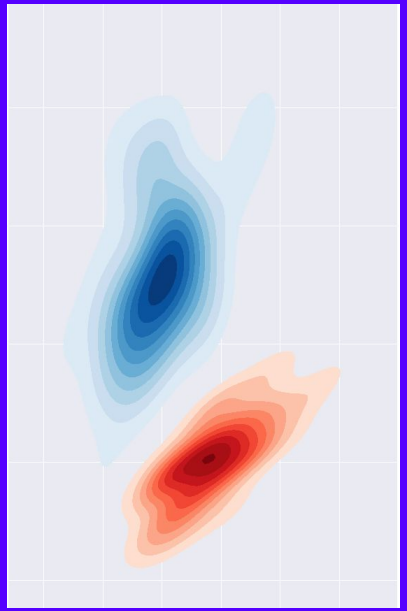
O que eu fiz antes:

- **Bacharel** em Física pela USP;
- **Mestre** em Física pela USP;
- **Doutor** em Física pela USP;
- Fiz doutorado dentro de **física teórica** e **matemática,** realizando pesquisa nas áreas de **computação quântica**, **estados topológicos da matéria** e **álgebras de Hopf**.

idwall

O que é ML?

idwall

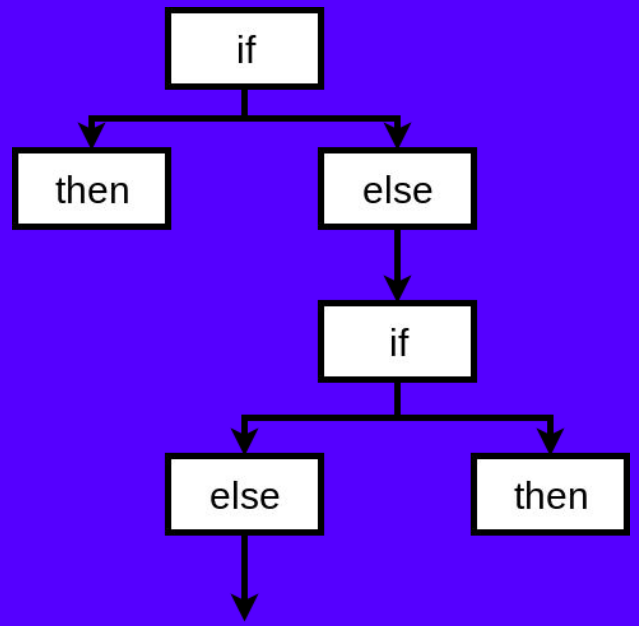


É um algoritmo de inferência.

Novos resultados de antigos.

- Conjunto (dataset) de **resultados conhecidos**;
- Dados são pares **entrada - saída;**
- Obtemos predições sobre **novos valores**.

idwall



Não explicitamente programado.

Se adapta ao dados.

- **ML** difere de uma árvore de decisões (*if-else*).
- Importância das variáveis é obtida **observando os dados**.

idwall

O que é preciso?

1

Conhecimento sobre o problema.

Entender o problema e o domínio. O problema pode ser resolvido? Como?

2

Dados.

Exemplos, muitos exemplos. O conjunto de dados precisa ser fiel ao real e diverso o suficiente.

3

Treinar o modelo.

O conjunto de dados é utilizado para treinar um modelo. Diversas métricas são obtidas e inferência pode ser realizada em novos dados.

4

Validar.

As predições do modelo fazem sentido? As métricas condizem com o observado? O que pode ser melhorado.

idwall

Automação e insights.

Tarefas manuais que precisam ser automatizadas.

É o principal caso de uso de ML.

- O problema já é conhecido;
- Muitos exemplos já existem;
- Modelos parciais já trazem resultados;
- Pode ser usado paralelamente.

Novos insights sobre os dados.

O processo de desenvolver e treinar um modelo pode trazer insights sobre os dados.

Novas relações entre variáveis, podem ser utilizadas pela melhorar a performance. Buscar cenários poucos exploradas.

idwall

Ferramentas.

Keras.

Um framework para Deep Learning. Suporta diversas bibliotecas, por exemplo TensorFlow

scikit-learn.

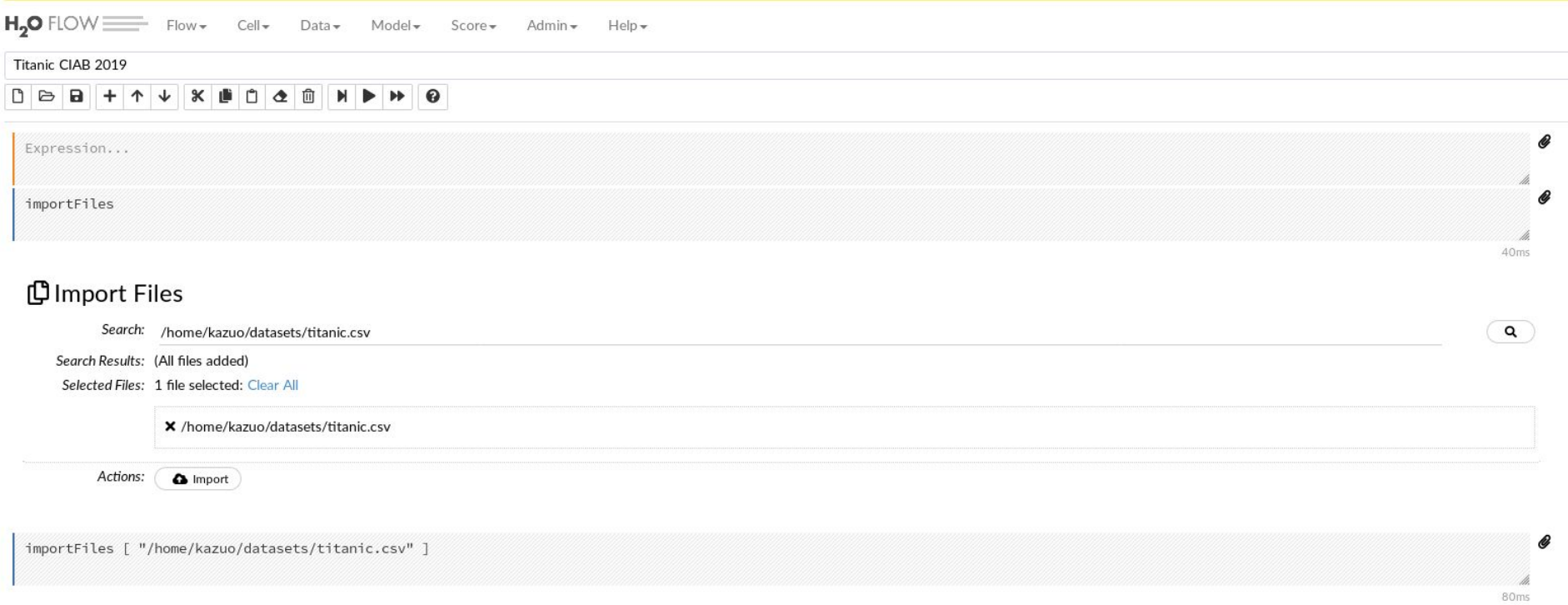
Um grande conjunto de ferramentas para ML. Extensa documentação e muitos exemplos.

H2O.

Ferramenta integrada para treinamento e visualização. Completa e user-friendly.

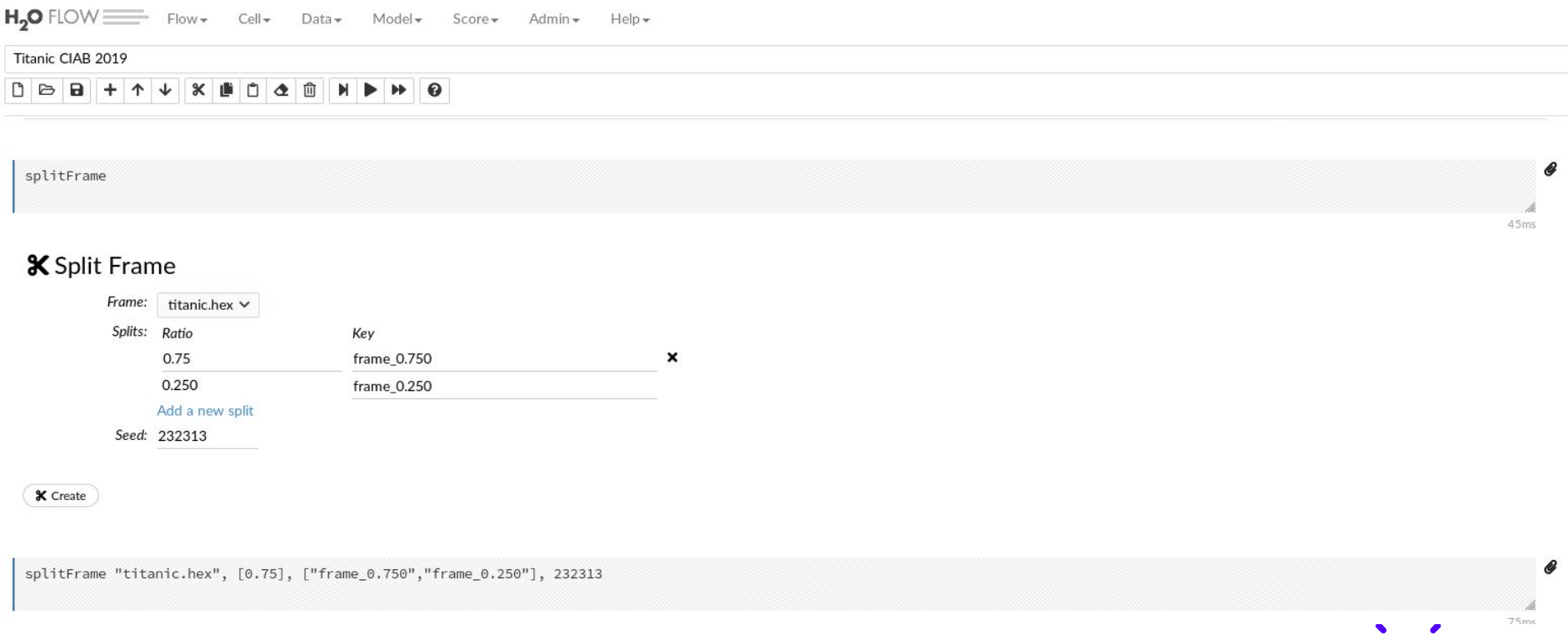
idwall

H2O.


H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Titanic CIAB 2019
Expression...
importFiles
40ms
Import Files
Search: /home/kazuo/datasets/titanic.csv
Search Results: (All files added)
Selected Files: 1 file selected: Clear All
/home/kazuo/datasets/titanic.csv
Actions: Import
importFiles ["/home/kazuo/datasets/titanic.csv"]
80ms

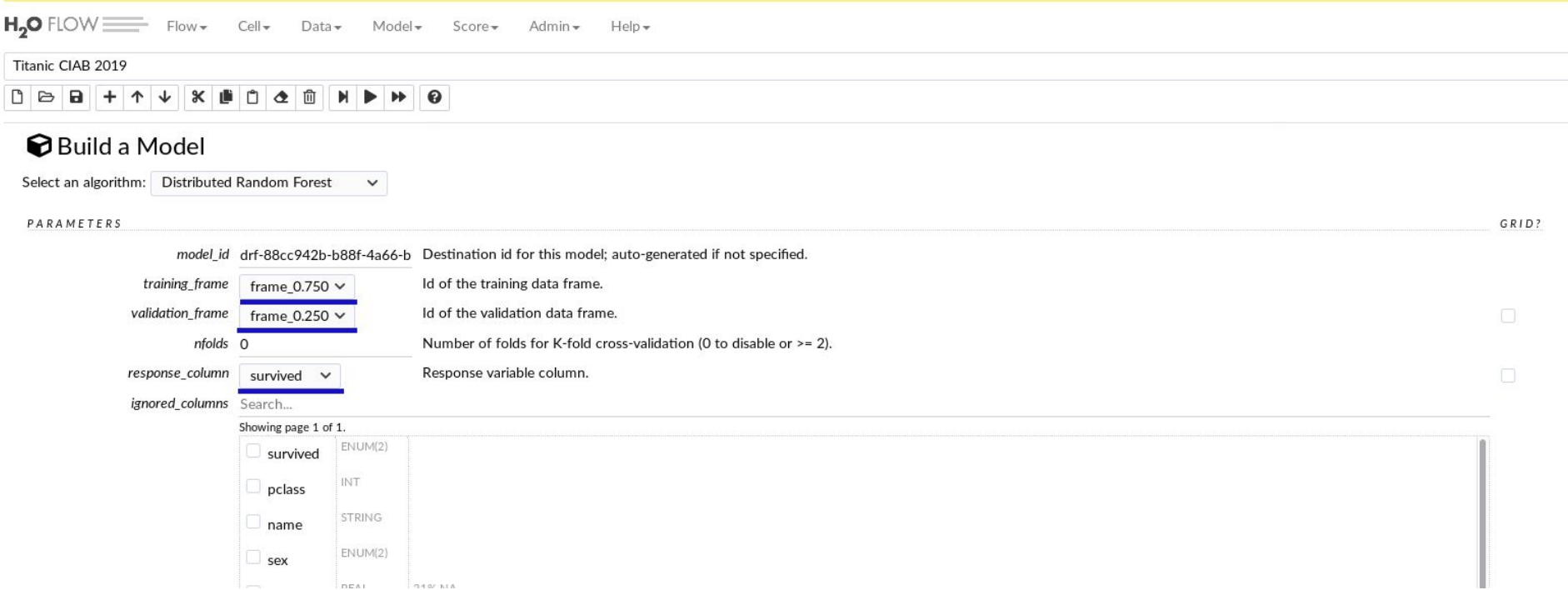
idwall

H2O.


H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Titanic CIAB 2019
splitFrame
45ms
Split Frame
Frame:
titanic.hex
Splits:
Ratio
Key
0.75
frame_0.750
0.250
frame_0.250
Add a new split
Seed:
232313
Create
splitFrame "titanic.hex", [0.75], ["frame_0.750","frame_0.250"], 232313

idwall

H2O.


H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Titanic CIAB 2019
Build a Model
Select an algorithm: Distributed Random Forest
PARAMETERS
GRID?
model_id
drf-88cc942b-b88f-4a66-b
Destination id for this model; auto-generated if not specified.
training_frame
frame_0.750
Id of the training data frame.
validation_frame
frame_0.250
Id of the validation data frame.
nfolds
0
Number of folds for K-fold cross-validation (0 to disable or >= 2).
response_column
survived
Response variable column.
ignored_columns
Search...
Showing page 1 of 1.
survived
ENUM(2)
pclass
INT
name
STRING
sex
ENUM(2)

idwall

H2O.

H2O FLOW Flow Cell Data Model Score Admin Help

Titanic CIAB 2019

Model

Model ID: drf-88cc942b-b88f-4a66-bcb7-7efa57afb0e7

Algorithm: Distributed Random Forest

Actions: Refresh | Predict... | Download POJO | Download Model Deployment Package (MOJO) | Export | Inspect | Delete | Download Gen Model

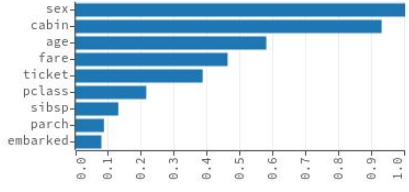
▸ MODEL PARAMETERS

▸ SCORING HISTORY - LOGLOSS

▸ ROC CURVE - TRAINING METRICS , AUC = 0.876057

▸ ROC CURVE - VALIDATION METRICS , AUC = 0.922145

▾ VARIABLE IMPORTANCES



▸ TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

▾ VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	0	1	Error	Rate	Precision
0	105	5	0.0455	5 / 110	0.83
1	21	57	0.2692	21 / 78	0.92
Total	126	62	0.1383	26 / 188	
Recall	0.95	0.73			

idwall

Obrigado!