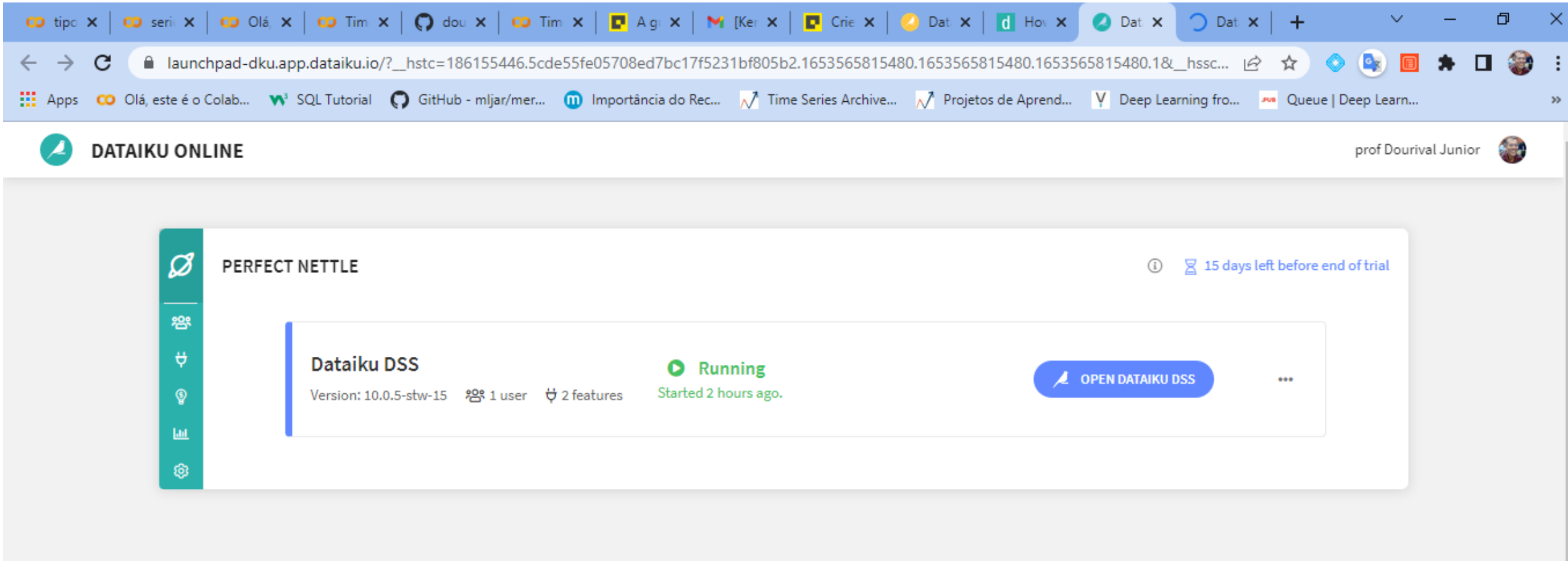
Neste estudo estarei mostrando como utilizar o framework Dataiku: Data Science Studio.

Eu fiz uma aplicação com o banco de dados do Loan_Prediction_Problem e usei um técnicas de ETL e ML.

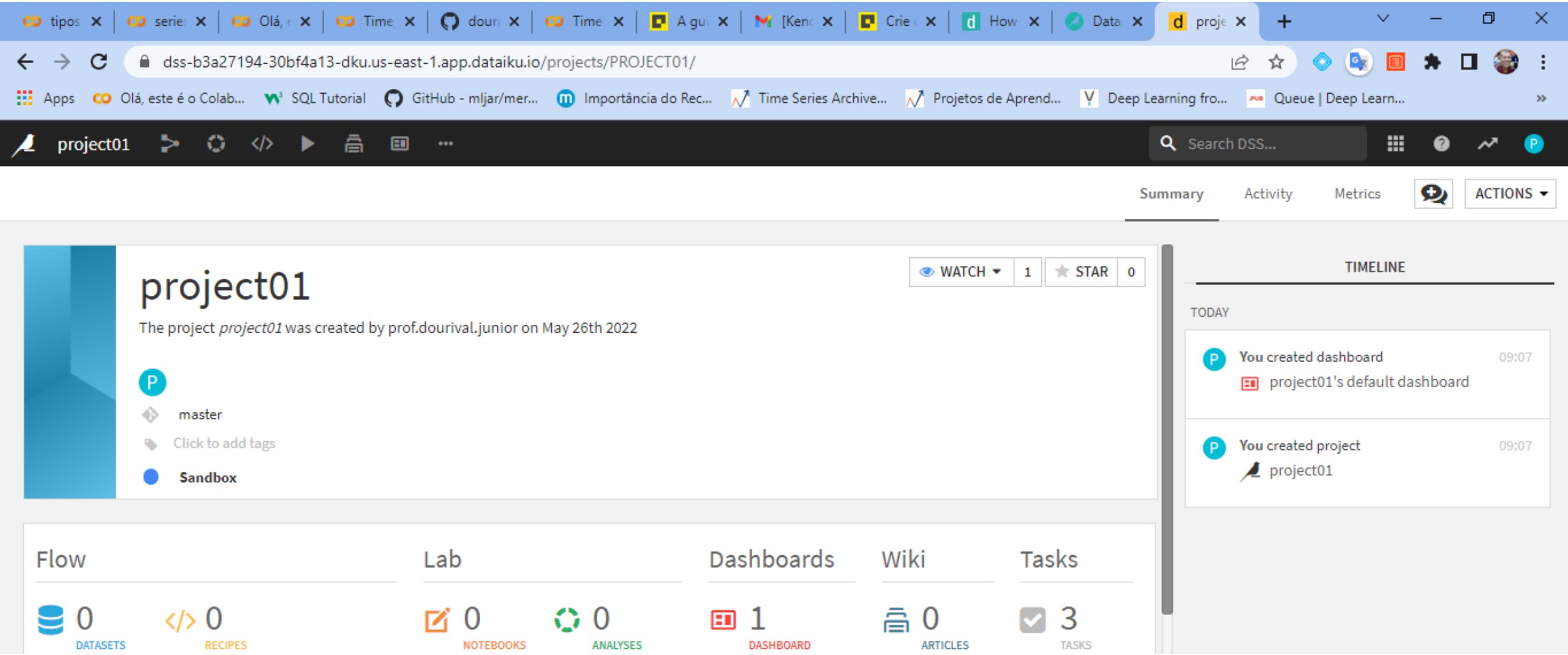
1º Passo:

Abra uma conta no site dataiku.com.

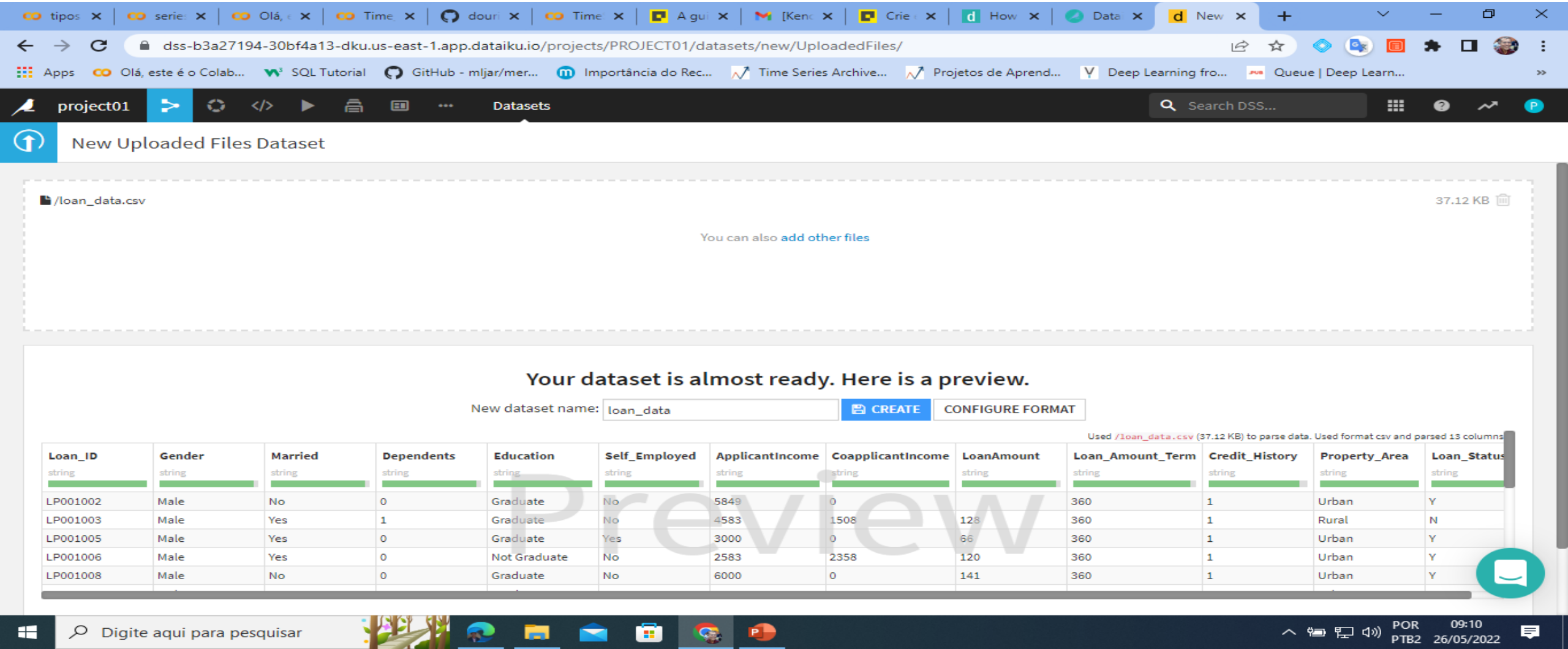
Preencha os dados e abra o workbench.



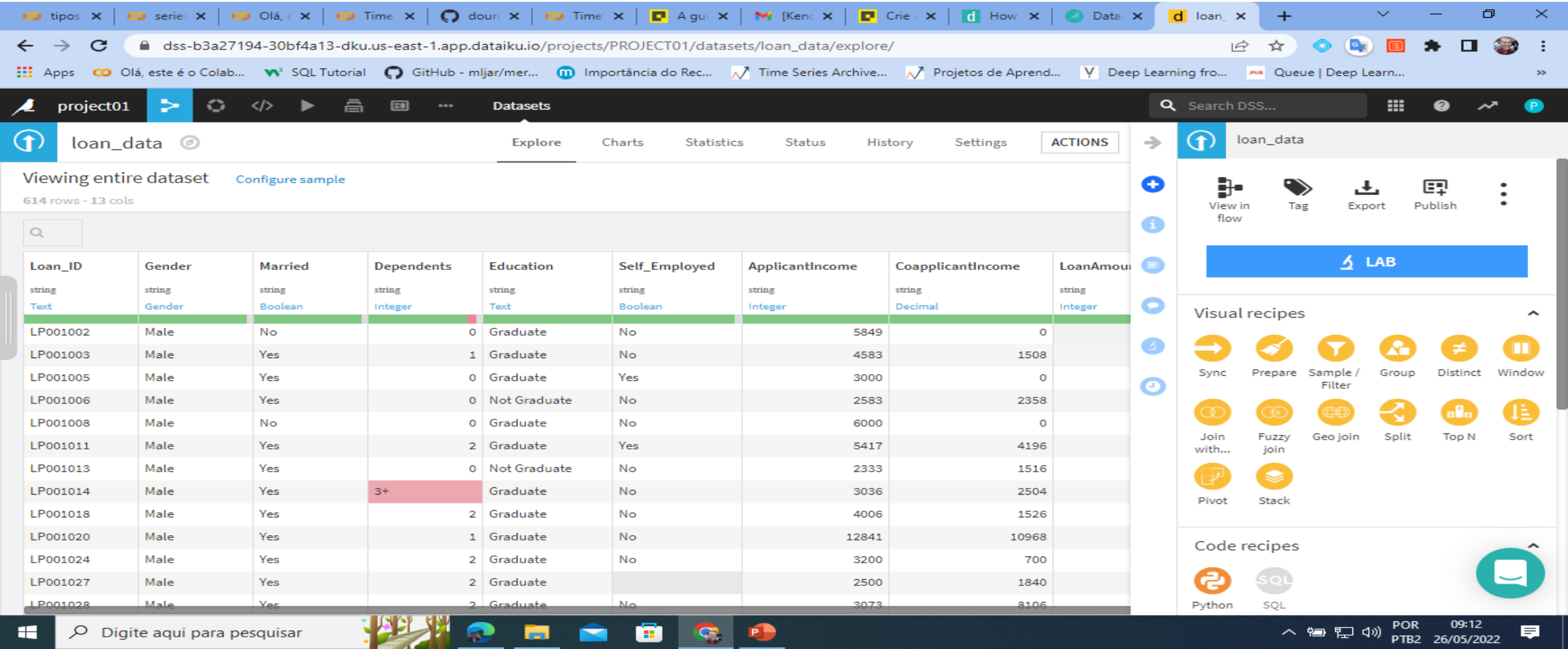
Aqui está o workbench onde criei o projeto.



2º Passo:
Carregue o seu dataset

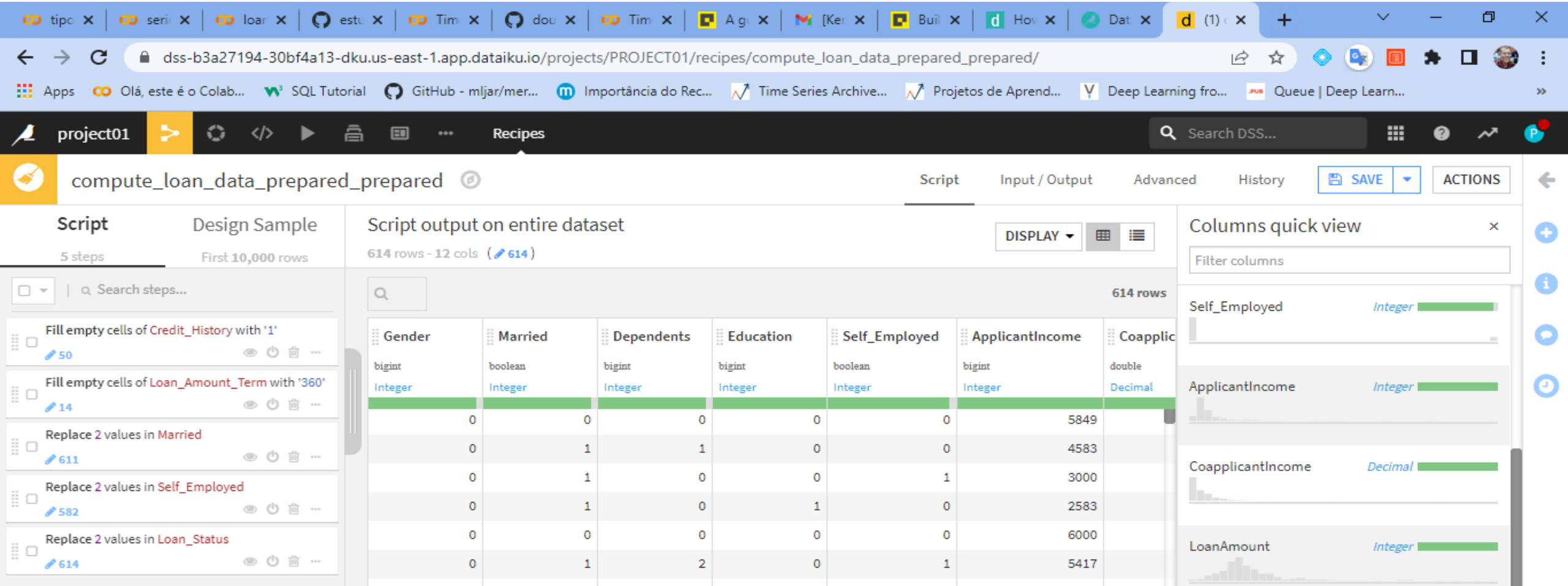


3º Passo:
Comece a realizar transformações nos seus dados clicando no ícone prepare.

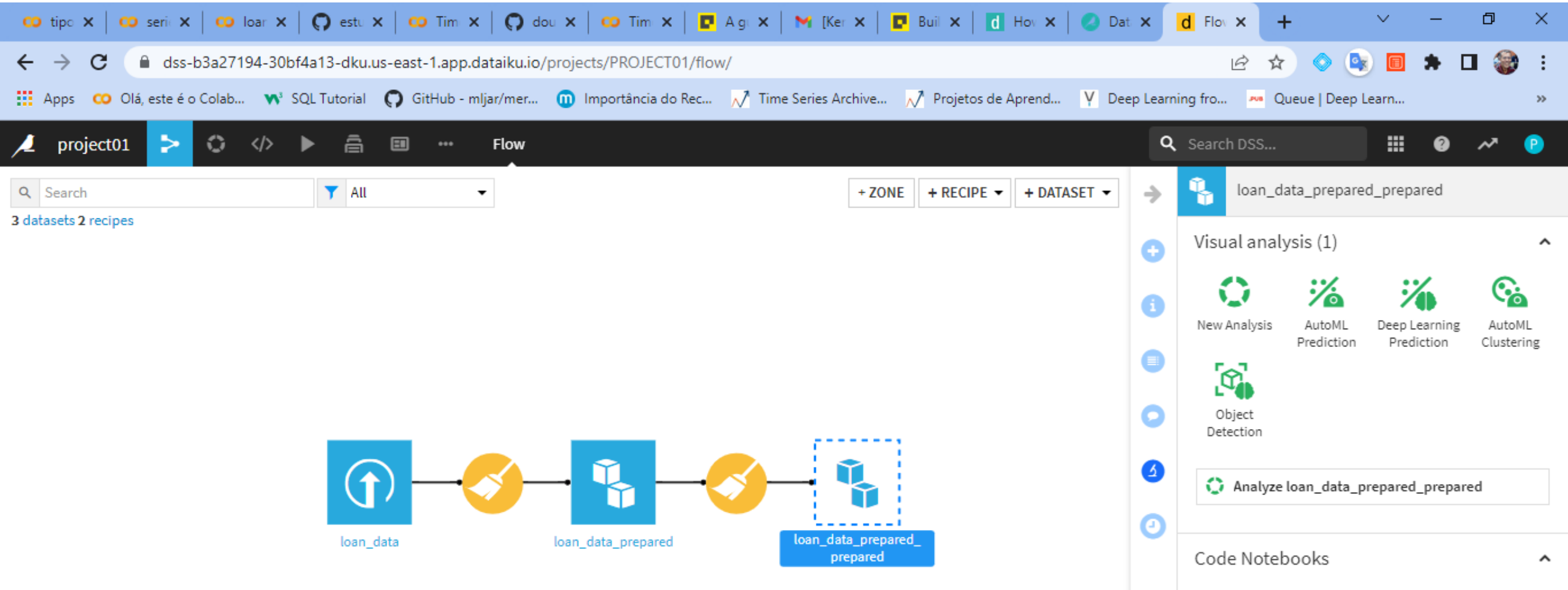


4º Passo:

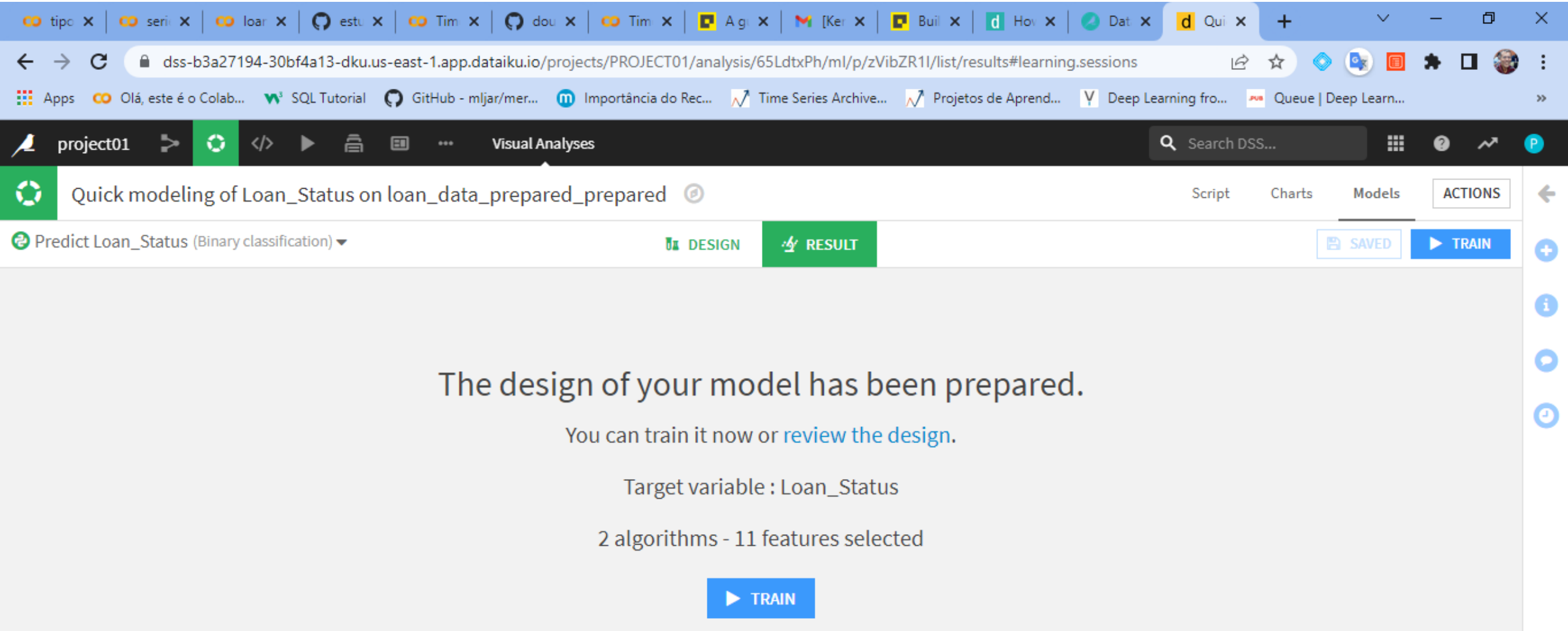
Realize todas as operações nas colunas. Por exemplo transformei na coluna Genger M por 0 e F por 1.



5º Passo:
Vá na aba run para colocar seu projeto em ML.
Carregue o seu arquivo tratado.



6º Passo:
Defina a variável alvo. Depois escolha quais modelos de ML você deseja aplicar no seu dataset.



7º Passo:
Rode o software e espere pelos resultados.
Eu usei vários métodos para análise.

Quick modeling of Loan_Status on loan_data_prepared_prepared

Script | Charts | Models | ACTIONS

Predict Loan_Status (Binary classification) | DESIGN | RESULT | SAVED | TRAIN

ACTIONS | Search... | Filter | Metric: Accuracy | SESSIONS | MODELS | TABLE

Previously trained

- [x] SESSION 2
  - [x] Random forest (s2) 0.808
  - [x] Logistic Regression (s2) 0.814
  - [x] Decision Tree (s2) 0.779
  - [x] Extra trees (s2) 0.808
  - [x] Artificial Neural Networ... 0.814

SESSION 2

Started Today at 11:05 , ended Today at 11:05 | *5 models* | *12 / 12 Features*

- Artificial Neural Network (s2) (no xval.) 0.814
- Logistic Regression (s2) 0.814
- Random forest (s2) 0.808
- Extra trees (s2) (no xval.) 0.808
- Decision Tree (s2) (no xval.) 0.779

Random forest (s2) 0.808

Done 2 minutes ago (2022-05-26 11:05:24) | Diagnostics (2)

Trees	100
Depth	6
Min samples	1
Size of hyperparameter search	2

Most important variables

Train set	442 rows
Test set	172 rows
Train time	about 7 seconds

Algumas considerações:

- Não fiz a normalização dos dados.
- Não apliquei a técnica de log nas variáveis que apresentavam um distribuição ‘skewed’.
- Não ajustei os hiper-parâmetros nos modelos de ML.
- O objetivo deste estudo é apresentar a ferramenta, suas funcionalidades.