

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.

The screenshot displays the Dataiku Online web interface. At the top, a browser tab bar shows multiple open tabs, including 'tipc', 'seri', 'Olá', 'Tim', 'dou', 'Tim', 'A g', '[Ker', 'Crie', 'Dat', 'Hov', 'Dat', and 'Dat'. The address bar shows the URL 'launchpad-dku.app.dataiku.io/?\_\_hstc=186155446.5cde55fe05708ed7bc17f5231bf805b2.1653565815480.1653565815480.1653565815480.1&\_\_hssc...'. Below the browser bar, the Dataiku Online header includes the logo, the text 'DATAIKU ONLINE', and the user profile 'prof Dourival Junior'. The main content area features a sidebar with icons for workspace management. The central workspace is titled 'PERFECT NETTLE' and includes a trial status indicator: '15 days left before end of trial'. A card within the workspace displays the 'Dataiku DSS' instance, which is currently 'Running'. Additional details for the instance include 'Version: 10.0.5-stw-15', '1 user', and '2 features'. It also notes that the instance 'Started 2 hours ago.' and provides a blue button labeled 'OPEN DATAIKU DSS'.

Aqui está o workbench onde criei o projeto.

The screenshot shows the Dataiku DSS interface for a project named 'project01'. The browser address bar shows the URL: `dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/`. The project page includes a header with tabs for Summary, Activity, and Metrics. The main content area displays the project name 'project01' and a description: 'The project *project01* was created by prof.dourival.junior on May 26th 2022'. Below this, there are buttons for 'WATCH' (1), 'STAR' (0), and 'Sandbox'. The bottom section shows a summary of project components: 0 Datasets, 0 Recipes, 0 Notebooks, 0 Analyses, 1 Dashboard, 0 Articles, and 3 Tasks. On the right, a 'TIMELINE' section shows recent activity, including 'You created dashboard' and 'You created project'.

project01

The project *project01* was created by prof.dourival.junior on May 26th 2022

WATCH 1 STAR 0

Flow

0 DATASETS

0 RECIPES

Lab

0 NOTEBOOKS

0 ANALYSES

Dashboards

1 DASHBOARD

Wiki

0 ARTICLES

Tasks

3 TASKS

TIMELINE

TODAY

You created dashboard 09:07

project01's default dashboard

You created project 09:07

project01

# 2º Passo: Carregue o seu dataset

tipos x | serie x | Olá, x | Time x | dour x | Time x | A gui x | [Kenc x | Crie x | d How x | Data x | d New x +

dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/datasets/new/UploadedFiles/

Apps Olá, este é o Colab... SQL Tutorial GitHub - mljar/mer... Importância do Rec... Time Series Archive... Projetos de Aprend... Deep Learning fro... Queue | Deep Learn...

project01 Datasets Search DSS...

New Uploaded Files Dataset

/loan\_data.csv 37.12 KB

You can also [add other files](#)

**Your dataset is almost ready. Here is a preview.**

New dataset name:  [CREATE](#) [CONFIGURE FORMAT](#)

Used /loan\_data.csv (37.12 KB) to parse data. Used format csv and parsed 13 columns

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
string	string	string	string	string	string	string	string	string	string	string	string	string
LP001002	Male	No	0	Graduate	No	5849	0	360	1	Urban	Y	
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y

Windows Digite aqui para pesquisar 09:10 26/05/2022

### 3º Passo:

Comece a realizar transformações nos seus dados clicando no ícone prepare.

The screenshot displays the Dataiku DSS interface for a project named 'project01'. The 'Datasets' tab is active, showing the 'loan\_data' dataset in the 'Explore' view. The dataset is described as having 614 rows and 13 columns. The table below shows the first 15 rows of data, with columns for Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, ApplicantIncome, CoapplicantIncome, and LoanAmount. The 'Visual recipes' sidebar on the right offers various data transformation options, including Sync, Prepare, Sample/Filter, Group, Distinct, Window, Join, Fuzzy join, Geo join, Split, Top N, Sort, Pivot, and Stack. The bottom of the image shows a Windows taskbar with a search bar and several application icons, along with a system clock indicating 09:12 on 26/05/2022.

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount
string Text	string Gender	string Boolean	string Integer	string Text	string Boolean	string Integer	string Decimal	string Integer
LP001002	Male	No	0	Graduate	No	5849	0	
LP001003	Male	Yes	1	Graduate	No	4583	1508	
LP001005	Male	Yes	0	Graduate	Yes	3000	0	
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	
LP001008	Male	No	0	Graduate	No	6000	0	
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	
LP001014	Male	Yes	3+	Graduate	No	3036	2504	
LP001018	Male	Yes	2	Graduate	No	4006	1526	
LP001020	Male	Yes	1	Graduate	No	12841	10968	
LP001024	Male	Yes	2	Graduate	No	3200	700	
LP001027	Male	Yes	2	Graduate		2500	1840	
LP001028	Male	Yes	2	Graduate	No	3073	8106	

## 4º Passo:

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

The screenshot displays the Dataiku web interface. The browser address bar shows the URL: `dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/recipes/compute_loan_data_prepared_prepared/`. The interface includes a top navigation bar with various icons and a search bar. Below this, the 'project01' tab is active, showing a list of recipes. The 'compute\_loan\_data\_prepared\_prepared' recipe is selected, and its 'Script' tab is open. The script is titled 'Script output on entire dataset' and shows 614 rows and 12 columns. The script contains five steps: 1. Fill empty cells of Credit\_History with '1' (50 rows affected), 2. Fill empty cells of Loan\_Amount\_Term with '360' (14 rows affected), 3. Replace 2 values in Married (611 rows affected), 4. Replace 2 values in Self\_Employed (582 rows affected), and 5. Replace 2 values in Loan\_Status (614 rows affected). The 'Script output on entire dataset' table shows the following columns: Gender, Married, Dependents, Education, Self\_Employed, ApplicantIncome, and CoapplicantIncome. The data is as follows:

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	0	0	0	0	5849	
0	1	1	0	0	4583	
0	1	0	0	1	3000	
0	1	0	1	0	2583	
0	0	0	0	0	6000	
0	1	2	0	1	5417	

On the right side of the interface, the 'Columns quick view' panel is open, showing a list of columns with their data types and distributions. The columns listed are: Self\_Employed (Integer), ApplicantIncome (Integer), CoapplicantIncome (Decimal), and LoanAmount (Integer).

5º Passo:

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

The screenshot displays the Dataiku DSS web interface. The browser address bar shows the URL: `dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/flow/`. The interface is divided into a top navigation bar, a main workspace, and a right-hand sidebar.

**Top Navigation Bar:** Includes tabs for various projects (e.g., `tipc`, `seri`, `loar`) and a search bar labeled "Search DSS...".

**Main Workspace:** Shows a project flow diagram with the following steps:

- `loan_data` (represented by a blue square icon with a circular arrow)
- An orange circular icon with a brush, representing a transformation step.
- `loan_data_prepared` (represented by a blue square icon with three cubes)
- Another orange circular icon with a brush.
- `loan_data_prepared_prepared` (represented by a blue square icon with three cubes, highlighted with a dashed blue border)

Below the flow diagram, there are filters for "All" and buttons for "+ ZONE", "+ RECIPE", and "+ DATASET".

**Right-Hand Sidebar:** Contains a section titled "Visual analysis (1)" with several analysis options:

- New Analysis
- AutoML Prediction
- Deep Learning Prediction
- AutoML Clustering
- Object Detection

Below these options, there is a button labeled "Analyze loan\_data\_prepared\_prepared". At the bottom of the sidebar, there is a section for "Code Notebooks".

6º Passo:

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

The screenshot displays the Dataiku DSS web interface. The browser's address bar shows the URL: `dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/analysis/65LdtXPh/ml/p/zVibZR1l/list/results#learning.sessions`. The interface includes a top navigation bar with a search field labeled "Search DSS...". Below this, the project name "project01" is visible. The main workspace shows a workflow titled "Quick modeling of Loan\_Status on loan\_data\_prepared\_prepared". The workflow is currently in the "DESIGN" phase, with a "RESULT" tab also available. The target variable is set to "Loan\_Status", and 2 algorithms with 11 features selected are shown. A large blue "TRAIN" button is prominently displayed at the bottom of the workspace. The right sidebar contains various icons for navigation and actions.

project01

Visual Analyses

Search DSS...

Quick modeling of Loan\_Status on loan\_data\_prepared\_prepared

Predict Loan\_Status (Binary classification)

DESIGN RESULT

SAVED TRAIN

The design of your model has been prepared.

You can train it now or [review the design](#).

Target variable : Loan\_Status

2 algorithms - 11 features selected

TRAIN



7º Passo:

Rode o software e espere pelos resultados.

Eu usei vários métodos para análise.

tipc x seri x loar x estu x Tim x dou x Tim x A g x [Ker x Buil x d Hov x Dat x d (1) x

← → ↺

dss-b3a27194-30bf4a13-dku.us-east-1.app.dataiku.io/projects/PROJECT01/analysis/65LdtxPh/ml/p/zVibZR11/list/results

🔗 ⭐ ⚙️ 👤

📱 📄 ⚙️

Apps Olá, este é o Colab... SQL Tutorial GitHub - mljar/mer... Importância do Rec... Time Series Archive... Projetos de Aprend... Deep Learning fro... Queue | Deep Learn...

🦅 project01

🔄

</>

▶

🖨

📄

⋮

Visual Analyses

🔍 Search DSS...

🗪 ? 📈 P

🔄

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

📄 SESSION 2

📄 Random forest (s2) 0.808 ☆

📄 Logistic Regression (s2) 🏆 0.814 ☆

📄 Decision Tree (s2) 0.779 ☆

📄 Extra trees (s2) 0.808 ☆

📄 Artificial Neural Networ... 🏆 0.814 ☆

SESSION 2

Started Today at 11:05 , ended Today at 11:05

5 models

12 / 12 Features ▾

Accuracy score

0.805

0.800

0.795

0.790

0s

5s

10s

15s

20s

25s

30s

Time (s)

● 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

Credit\_History

ApplicantIncome

LoanAmount

CoapplicantIncome

Property\_Area

Train set

442 rows

Test set

172 rows

Train time

about 7 seconds

🔍 Digite aqui para pesquisar

🌳 🏠 📧 📅 🌐 📄

🔊 🖨 📶

POR 11:07  
PTB2 26/05/2022

🗨

# 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.

