Universidade Federal
de Ouro Preto

Trabalho da Disciplina:
Redes Neurais e Aprendizado em Profundidade

Predição na Base de Dados SpaceShip Titanic do Kaggle

Halliday Gauss Costa dos Santos
William Gomes de Lima Natividade

Ouro Preto (Outubro-2022)

Organização do Trabalho

01 Introdução

02 Motivação e Objetivo

03 Trabalhos Relacionados

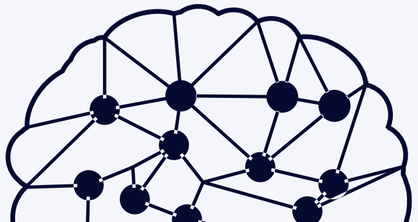
04 Base de Dados

05 Metodologia

06 Resultados

07 Conclusão

08 Referências



Introdução - Competições

- **Competições são de grande importância no aprendizado por trazer aos competidores a curiosidade e o desejo para resolver um problema.**

- **Isso se mantém mesmo nos desafios voltados para a área da computação, como, por exemplo, predições sobre uma base de dados.**

Introdução - Kaggle

- **O Kaggle é uma plataforma online de ciência de dados que hospeda competições, além de abrigar uma grande comunidade de competidores e cientistas de dados de diversas origens.**

- **Cada competidor pode participar das competições em aberto, submeter seus resultados na plataforma, e em seguida verificar o desempenho obtido e a classificação no ranking da competição.**

Competitions

Your Work

Search competitions

Filters

Getting Started

Results

Recently Launched

Spaceship Titanic
Predict which passengers are transported to an alternate dimension
Getting Started · 2104 Teams · Ongoing
Knowledge

Store Sales - Time Series Forecasting
Use machine learning to predict grocery sales
Getting Started · 583 Teams · Ongoing
Knowledge

I'm Something of a Painter Myself
Use GANs to create art - will you be the next Monet?
Getting Started · Code Competition · 55 Teams · Ongoing
Prizes

Contradictory, My Dear Watson
Detecting contradiction and entailment in multilingual text using TPUs
Getting Started · Code Competition · 41 Teams · Ongoing
Prizes

Petals to the Metal - Flower Classification on TPU
Getting Started with TPUs on Kaggle!
Getting Started · Code Competition · 93 Teams · Ongoing
Knowledge

Connect X
Connect your checkers in a row before your opponent!
Getting Started · Simulation Competition · 207 Teams · Ongoing
Knowledge

GettingStarted Prediction Competition

Spaceship Titanic

Predict which passengers are transported to an alternate dimension

Kaggle · 2,104 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team

New Notebo

Search notebooks

All Your Work Shared With You Bookmarks

Spaceship Titanic - Step by Step
Updated 4h ago
Score: 0.8071 · 3 comments · Spaceship Titanic +1

Spaceship Titanic Data Science Solutions
Updated 1h ago
21 comments · Spaceship Titanic

Spaceship Titanic - 0.80+ XGBoost/GridSearchCV
Updated 19h ago
Score: 0.8043 · 6 comments · Spaceship Titanic

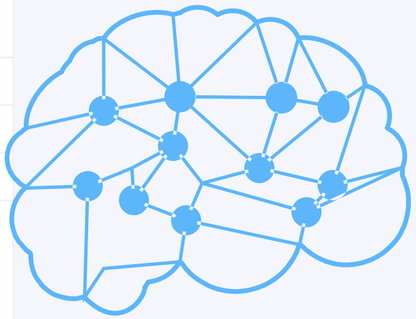
Overview Data Code Discussion Leaderboard Rules Team

My Submissions

Submit Predictions

This leaderboard is calculated with all of the test data.

#	Team	Members	Score	Entries	Last	Code	Join
1	CADang		0.87304	55	5d		
2	Setec Astronomy		0.86626	136	12d		
3	Involutional Neural Network		0.84264	159	6d		
4	Karl Cini		0.82253	18	23d		
5	Rizky Anugrah Ananto		0.81692	4	1mo		
6	A Xiao		0.81692	1	21d		
7	Konstantin Dobkes		0.81552	35	15h		
8	sejoong.kim		0.81505	2	1mo		
9	Mustafa Kel		0.81505	14	14d		
10	Ryan Li		0.81482	8	2mo		



Spaceship Titanic - Step by Step

Notebook Data Logs Comments (3)

5 Copy & Edit 3

Competition Notebook
Spaceship Titanic

Run
546.2s

Public Score
0.80476

Version 2 of 2

Spaceship Titanic

The task is to predict whether a passenger was transported to an alternate dimension during the Spaceship Titanic's collision with the spacetime anomaly. You're given a set of personal records recovered from the ship's damaged computer system.

For this challenge I took some inspiration from the following Notekooks:

- https://www.kaggle.com/code/sakshamparashar/simple-approach-to-spaceship-titanic-a-z
- https://www.kaggle.com/code/yijiecao/titanic-prediction-ensemble-accuracy-0-83

We start with loading some basic libraries and configuring some settings:

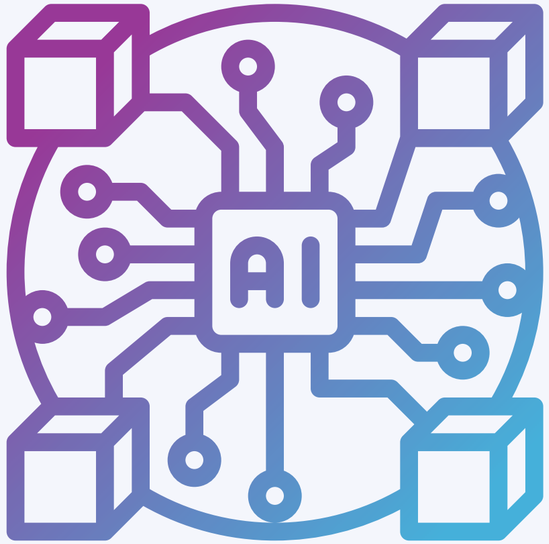
In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid")
```

Table of Contents

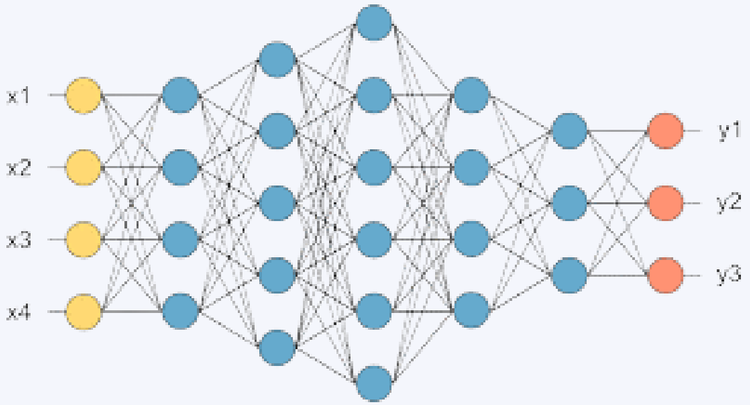
Introdução - Algoritmos

Para resolver esses problemas, de classificação ou de regressão, por exemplo, os cientistas de dados utilizam de diversas técnicas e algoritmos (Casper, 2020).

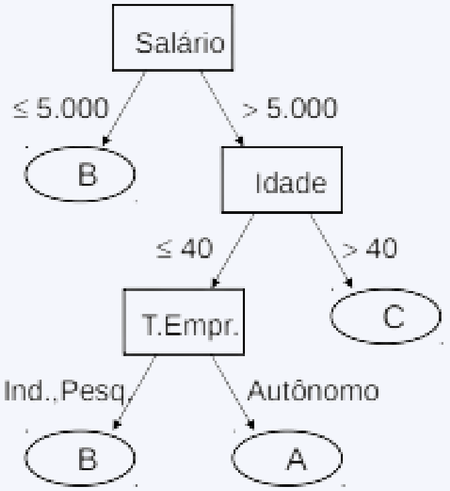


Algoritmos de Aprendizado

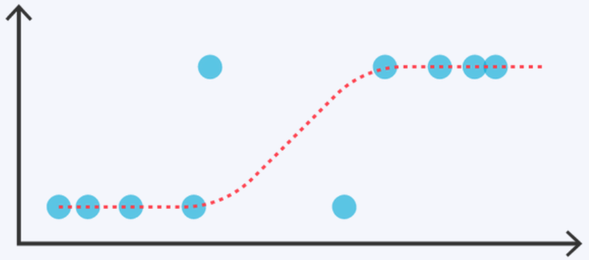
MLP



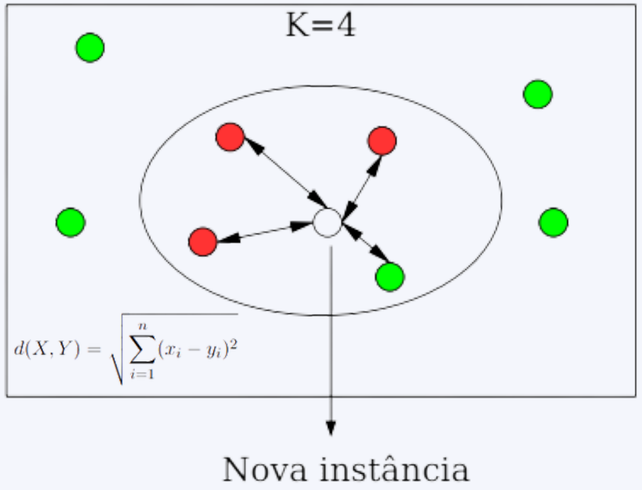
Árvore de Decisão



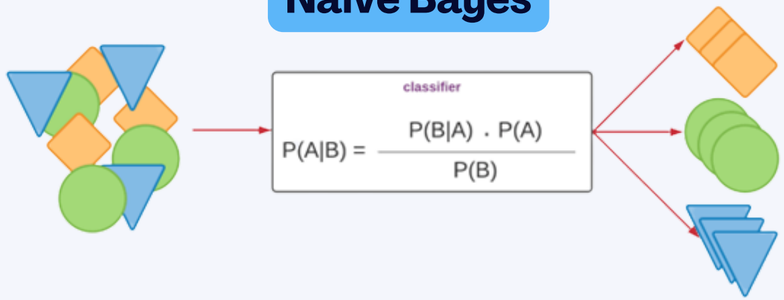
Regressão Logística



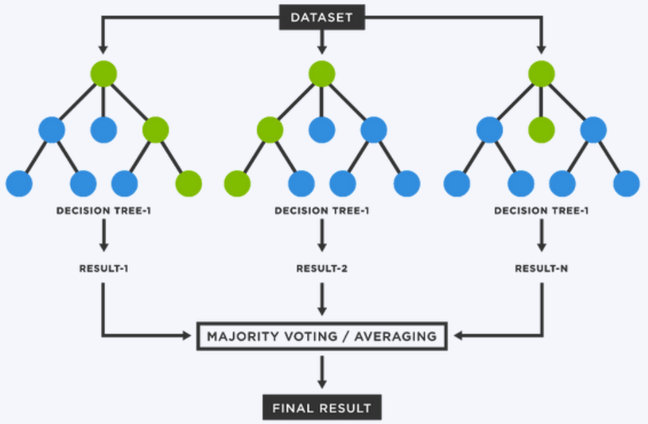
KNN



Naive Bayes



Random Forest



Introdução - Qualidade dos Dados

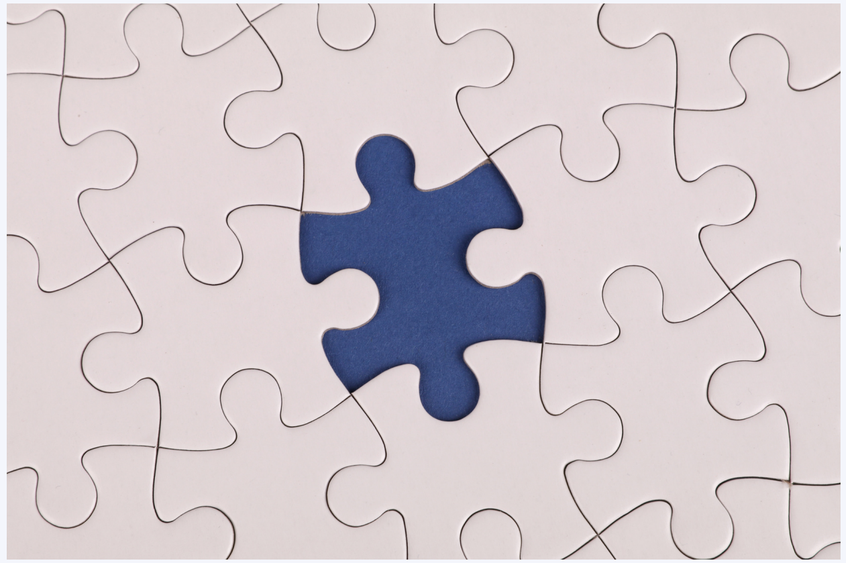
- **A qualidade dos dados impactam diretamente no desempenho dos algoritmos e é uma das principais preocupações no Aprendizado de Máquina.**

- **O tratamento de valores ausentes deve ser cuidadosamente planejado, caso contrário, distorções podem ser introduzidas no conhecimento e o desempenho dos classificadores pode ser prejudicado (Gustavo, 2003).**

- **A qualidade do conhecimento depende da qualidade dos dados.**

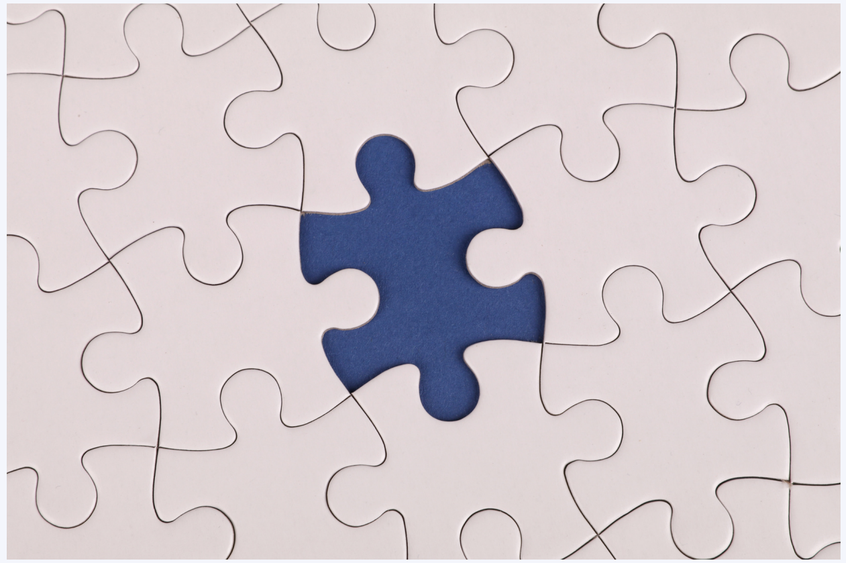
Introdução - Lidando com Dados Ausentes

- **Dados contínuos: média, moda ou métodos de regressão.**
- **Não costumam funcionar bem com a presença de atributos categóricos ou dos dois juntos. (Ramérez, 2011)**

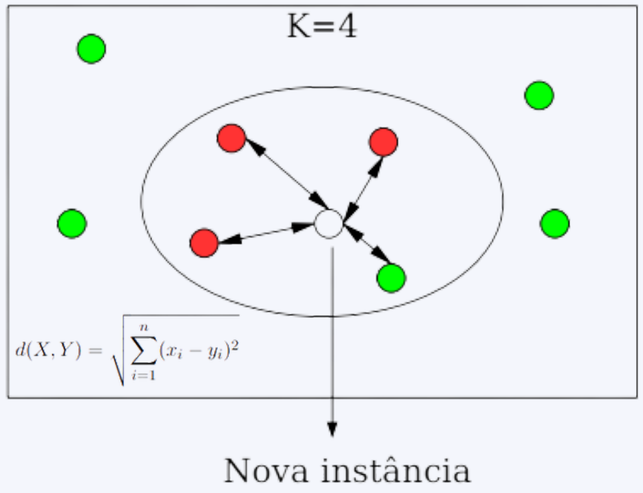


Introdução - Lidando com Dados Ausentes

Existem bons algoritmos que são utilizados para a imputação de um valor ausente no conjunto de dados, dentre eles o KNNImputer , o qual utiliza uma abordagem local (Armina, 2017).

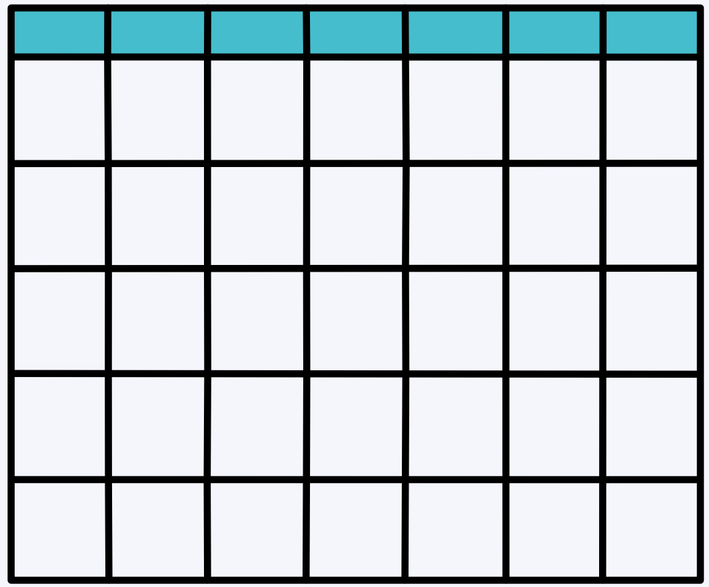


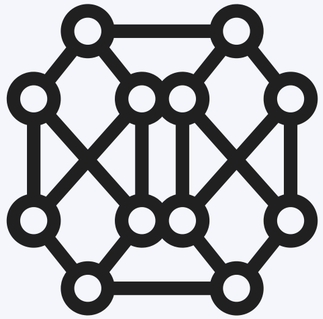
KNN



Introdução - Redes Neurais em dados tabulares

- **Redes Neurais Profundas (DNN) são boas para resolução de problemas envolvendo textos, imagens e audios.**
- **Não possuem bons resultados em problemas de classificação envolvendo dados tabulares comparado as diversas técnicas atuais de aprendizado.**

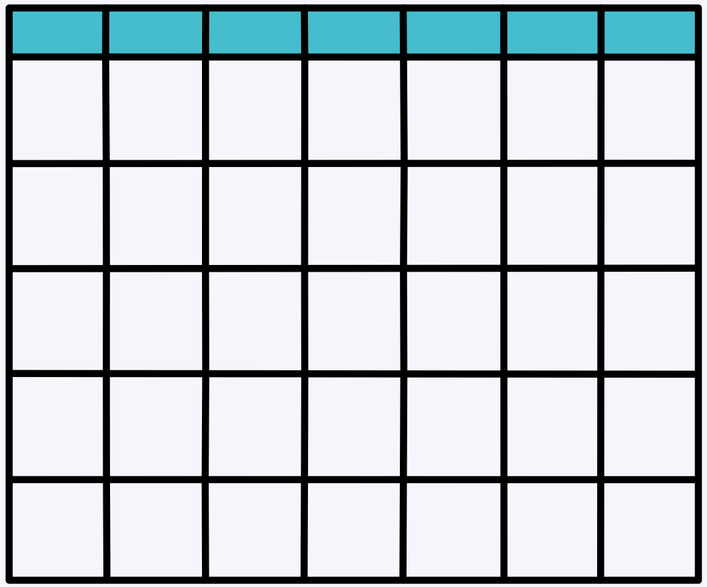


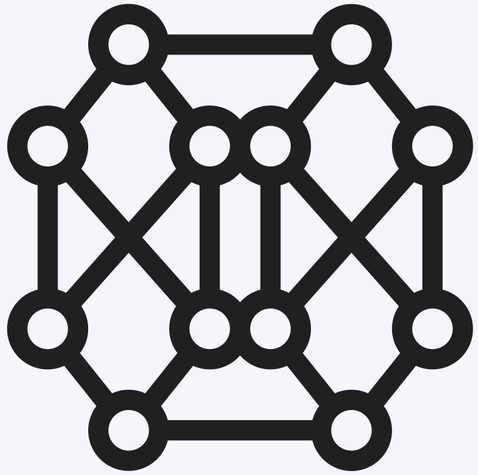


Introdução - TabNet

DNN TabNet. Características:

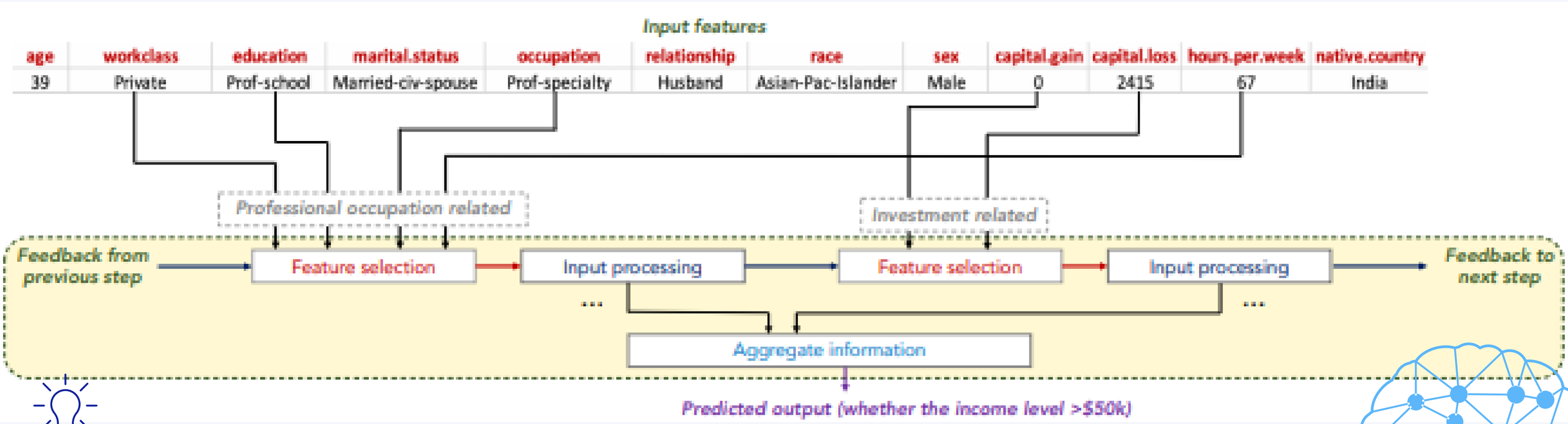
- **Flexibilidade: é possível inserir dados tabulares brutos sem qualquer pré-processamento e treinar utilizando otimização baseada em gradiente descendente.**





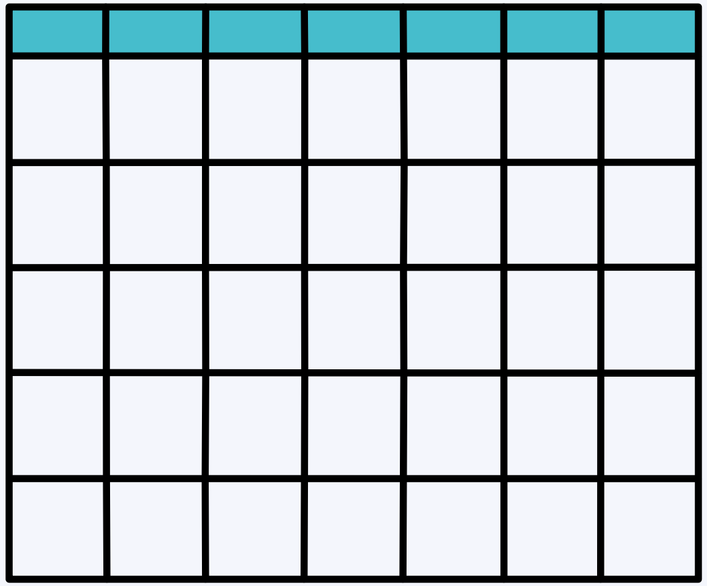
Introdução - TabNet

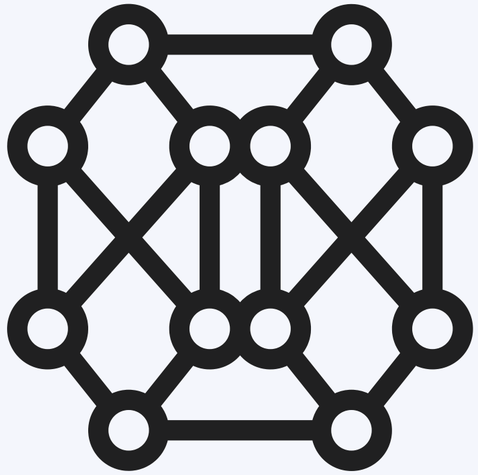
Capacidade de Aprendizado: o TabNet usa atenção sequencial para escolher quais características serão utilizadas em cada etapa de decisão, permitindo interpretabilidade e melhor aprendizado ao destacar as características mais relevantes.



Introdução - TabNet

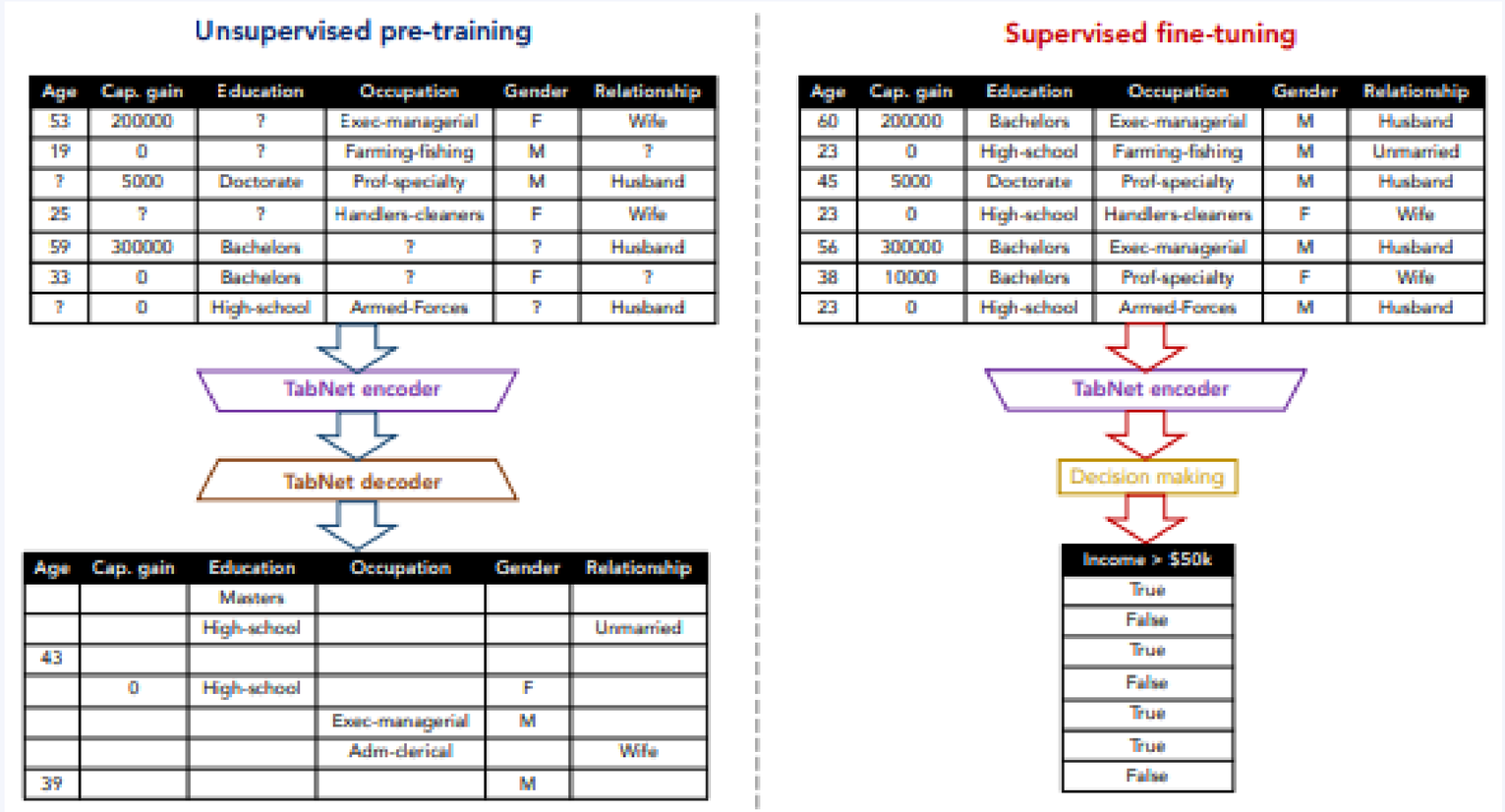
Desempenho: supera ou está no mesmo nível de outros modelos de aprendizado tabular em vários conjuntos de dados para problemas de classificação e regressão de diferentes domínios.





Introdução - TabNet

Pré-treinamento não supervisionado: é possível utilizar pré-treinamento para aprender relações, prever os valores mascarados, aprender os pesos e usar para uma tarefa supervisionada. O uso desse recurso aumenta muito o desempenho da tarefa proposta.

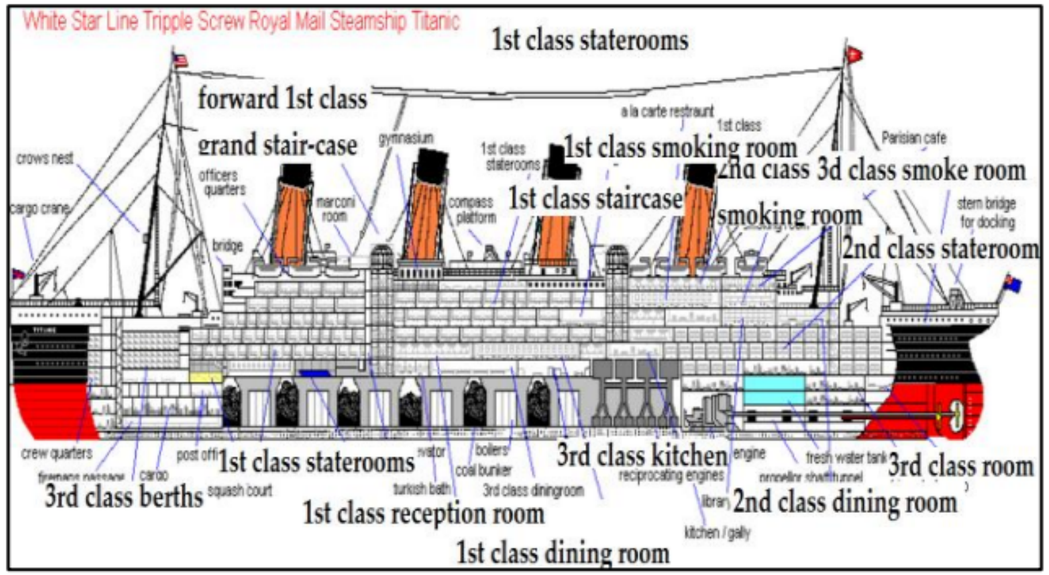


Motivação e Objetivos

- **Predição sobre a base de dados do Kaggle chamada: SpaceShip Titanic.**
- **Utilizar a DDN TabNet.**
- **Técnicas de mineração de dados para: descobrimento de informações e transformações dos dados.**
- **Pré-processamento com ênfase no uso do KNNImputer no preenchimento dos valores nulos**

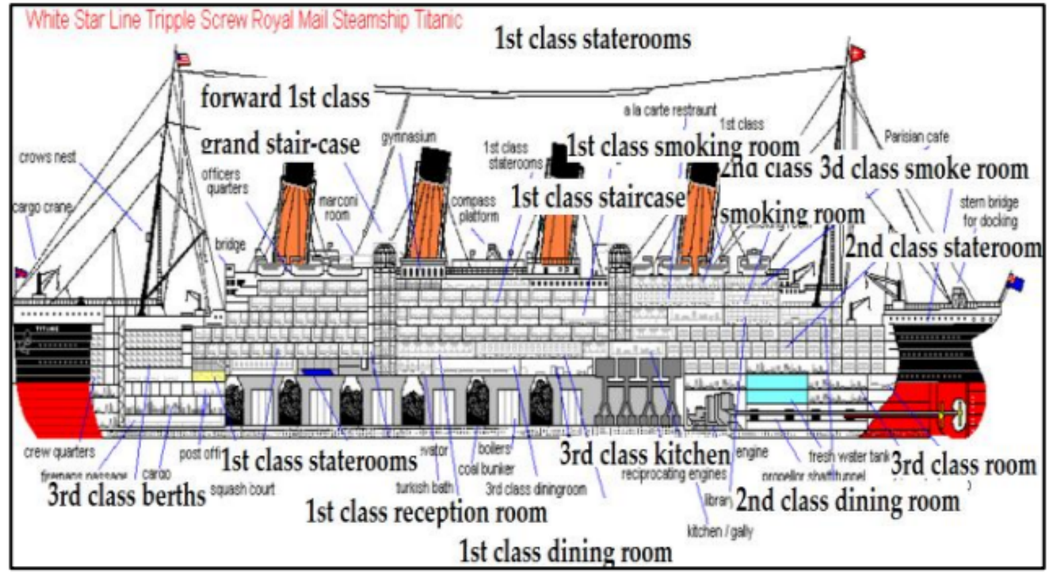
Trabalhos Relacionados - Kothari Yash (2018)

- **Kothari Yash. ”Predicting the Survivors of the Titanic -Kaggle, Machine Learning From Disaster”. Academia.edu, 2018.**
- **Base de dados Titanic (Kaggle).**
- **Informações sobre cada passageiro presente no Titanic, incluindo 11 variáveis descritivas associadas a 891 passageiros no conjunto de treino e 418 no conjunto de teste.**
- **Visão Geral, correlações.**



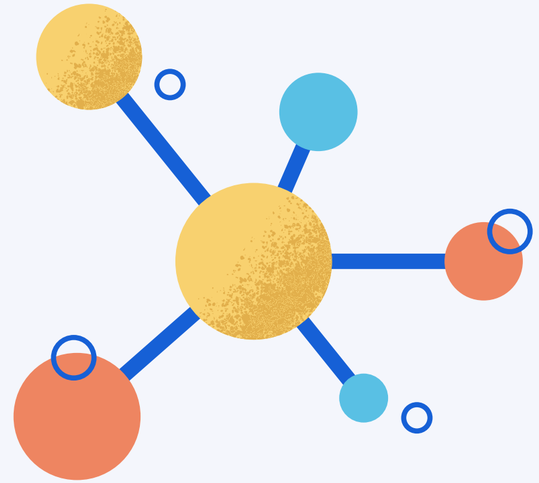
Trabalhos Relacionados - Kothari Yash (2018)

- **Objetivo: prever quais passageiros sobrevivem e quais morrem.**
- **Algoritmos: Naive Bayes, Árvore de Decisão e Random Forest.**
- **Naive Bayes 92,52%.**



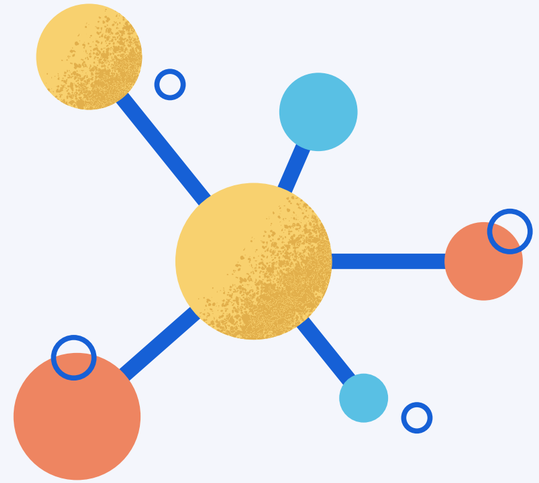
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **Mechanisms of Action (MoA) Prediction do Kaggle.**
- **Prever mecanismos de ação (MoA), ou seja, a atividade biológica de uma molécula.**
- **Modelos selecionados por meio de um concurso público.**
- **Acurácia: valor médio da função de perda logarítmica aplicada a cada par de anotações de droga-MoA.**



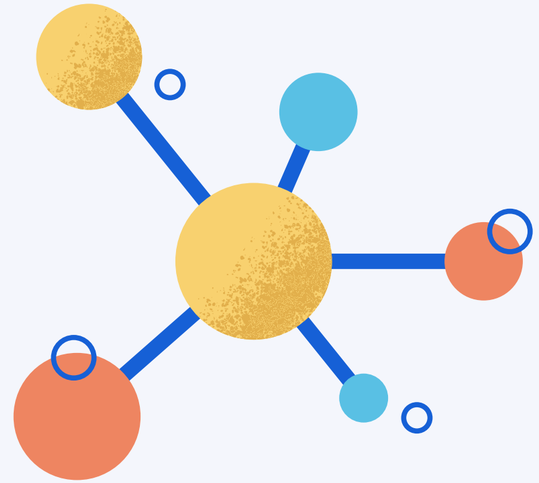
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **Pré-Processamento:**
  - **Seleção de atributos.**
  - **Quantile Transformer.**
  - **PCA.**
  - **Novos atributos: soma, média, desvio padrão, curtose e assimetria sobre as características de expressão gênica e viabilidade celular.**



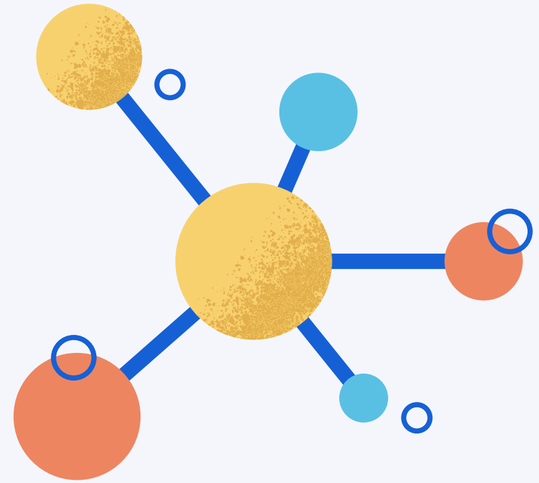
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **Validação:**
  - **Estratificação Multilabel.**
  - **Cada conjunto não usado para o treinamento é utilizado para avaliar a validação cruzada: log-loss(entropia cruzada).**



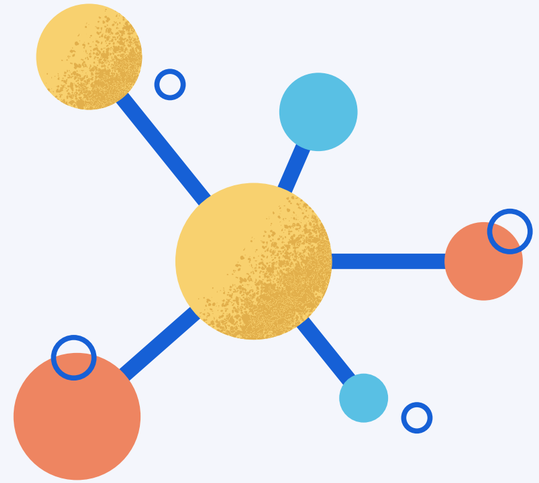
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **Modelos:**
  - **NN.**
  - **DNN TabNet.**
  - **ResNet.**
  - **Ensemble.**



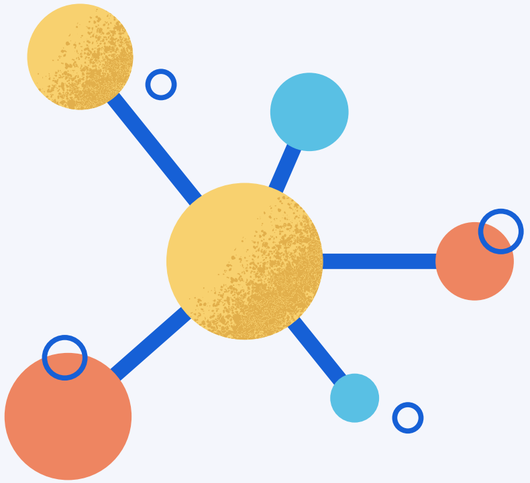
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **NN:**
  - **Uma camada de entrada, uma camada oculta com 1000 neurônios e uma camada de saída.**
  - **Inserção de um dropout após cada camada.**
  - **Função de ativação LeakyReLu**
  - **Adam, log-loss.**



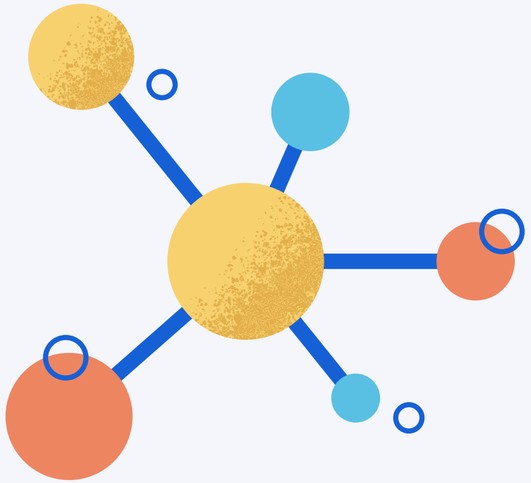
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **DNN TabNet:**
  - **Largura de 32 para a camada de previsão de decisão e largura de 32 para a incorporação de atenção para cada máscara**
  - **1 na etapa da arquitetura, gama de 1.3, otimizador Adam, lr 2e-2 e um decaimento de peso de 1e-5.**
  - **Coeficiente de perda de esparsidade de 0, entmax como função de mascaramento.**
  - **Batch_size de 1024, virtual_batch_size de 128, 200 épocas e paciência de 50 épocas.**



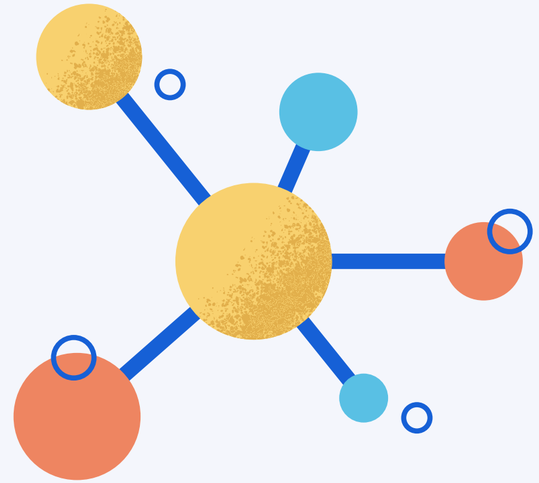
Trabalhos Relacionados - Lombardi Alessandro (2020)

- **ResNet:**
  - **Primeiramente, prever alvos não pontuados. Adam, 50 épocas, binary cross entropy, lr de 1e-3, batch_size de 128.**
  - **Escalonador de taxa de aprendizado ReduceLROnPlateau com um fator de 0.5 e paciência de 4 épocas.**
  - **Transferência de aprendizado para outra ResNet treinada da mesma forma que a NN.**



Trabalhos Relacionados - Lombardi Alessandro (2020)

- **Resultados:**
  - **Na validação: NN, Tabnet e ResNet (0.0159, 0.0150 e 0.0147).**
  - **No Kaggle: 0.01626 (NN), 0.01633 (TabNet), 0.01644 (ResNet) , 0.1610 (Ensemble).**
  - **Ensemble em 72° lugar no Kaggle.**



Base de Dados (Spaceship Titanic) - O Problema

No ano de 2912, foi recebido uma transmissão de quatro anos-luz de distância. A nave espacial Titanic com quase 13.000 passageiros a bordo, partiu em uma viagem inaugural transportando emigrantes do nosso sistema solar para três exoplanetas recém-habitáveis orbitando estrelas próximas.

Base de Dados (Spaceship Titanic) - Descrição dos Dados

Treino: 8693 registros e 14 atributos. Teste : 4277 registros e 13 atributos.

PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True
0005_01	Earth	False	F/0/P	PSO J318.5-22	44.0	False	0.0	483.0	0.0	291.0	0.0	Sandie Hinetthews	True
0006_01	Earth	False	F/2/S	TRAPPIST-1e	26.0	False	42.0	1539.0	3.0	0.0	0.0	Billex Jacostaffey	True
0006_02	Earth	True	G/0/S	TRAPPIST-1e	28.0	False	0.0	0.0	0.0	0.0		Candra Jacostaffey	True
0007_01	Earth	False	F/3/S	TRAPPIST-1e	35.0	False	0.0	785.0	17.0	216.0	0.0	Andona Beston	True
0008_01	Europa	True	B/1/P	55 Cancri e	14.0	False	0.0	0.0	0.0	0.0	0.0	Erraiam Flatic	True

Base de Dados (Spaceship Titanic) - Valores Ausentes

Treino: 2087 registros. Teste : 996 registros.

PassengerId	0
HomePlanet	201
CryoSleep	217
Cabin	199
Destination	182
Age	179
VIP	203
RoomService	181
FoodCourt	183
ShoppingMall	208
Spa	183
VRDeck	188
Name	200
Transported	0

PassengerId	0
HomePlanet	87
CryoSleep	93
Cabin	100
Destination	92
Age	91
VIP	93
RoomService	82
FoodCourt	106
ShoppingMall	98
Spa	101
VRDeck	80
Name	94

Metodologia - Transformações

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
3	0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
4	0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8688	9276_01	Europa	False	A/98/P	55 Cancri e	41.0	True	0.0	6819.0	0.0	1643.0	74.0	Gravior Noxnuther	False
8689	9278_01	Earth	True	G/1499/S	PSO J318.5-22	18.0	False	0.0	0.0	0.0	0.0	0.0	Kurta Mondalley	False
8690	9279_01	Earth	False	G/1500/S	TRAPPIST-1e	26.0	False	0.0	0.0	1872.0	1.0	0.0	Fayey Connon	True
8691	9280_01	Europa	False	E/608/S	55 Cancri e	32.0	False	0.0	1049.0	0.0	353.0	3235.0	Celeon Hontichre	False
8692	9280_02	Europa	False	E/608/S	TRAPPIST-1e	44.0	False	126.0	4688.0	0.0	0.0	12.0	Propsh Hontichre	True

8693 rows × 14 columns

Metodologia - Transformações

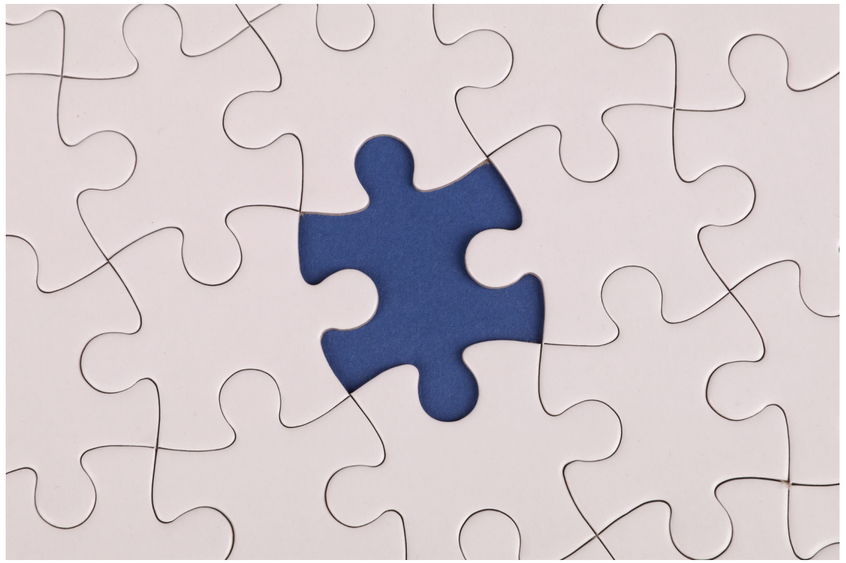
Alguns valores nulos foram preenchidos.

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Transported	Group	Cabin_Num	Total_Spent	...	VIP_False	VIP_True	PositionGroup_1	PositionGroup_2
0	39.0	0.0	0.0	0.0	0.0	0.0	0	1	0	0.0	...	1.0	0.0	1.0	0.0
1	24.0	109.0	9.0	25.0	549.0	44.0	1	2	0	736.0	...	1.0	0.0	1.0	0.0
2	58.0	43.0	3576.0	0.0	6715.0	49.0	0	3	0	10383.0	...	0.0	1.0	1.0	0.0
3	33.0	0.0	1283.0	371.0	3329.0	193.0	0	3	0	5176.0	...	1.0	0.0	0.0	1.0
4	16.0	303.0	70.0	151.0	565.0	2.0	1	4	1	1091.0	...	1.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8688	41.0	0.0	6819.0	0.0	1643.0	74.0	0	9276	98	8536.0	...	0.0	1.0	1.0	0.0
8689	18.0	0.0	0.0	0.0	0.0	0.0	0	9278	1499	0.0	...	1.0	0.0	1.0	0.0
8690	26.0	0.0	0.0	1872.0	1.0	0.0	1	9279	1500	1873.0	...	1.0	0.0	1.0	0.0
8691	32.0	0.0	1049.0	0.0	353.0	3235.0	0	9280	608	4637.0	...	1.0	0.0	1.0	0.0
8692	44.0	126.0	4688.0	0.0	0.0	12.0	1	9280	608	4826.0	...	1.0	0.0	0.0	1.0

8693 rows × 38 columns

Metodologia - Transformações

- **Preenchimento de valores nulos: KNNImputer com o valor de K igual a 5.**
- **Normalização MinMax.**
- **Não deu certo:**
  - **Combinações entre: MinMax, StandardScaler e RobustScaler**
  - **PCA e SMOTE (50.4% e 49.6 %).**



Metodologia - Classificação

- **Ensemble de 5 TabNet's pré-treinadas (melhor que que sem pré-treinamento)**
- **Ajustados de formas distintas e treinados sobre porções diferentes ou não da base de dados**

Metodologia - Classificação - Parâmetros

Um modelo não supervisionado.

```
unsupervised_model_no_preproc = TabNetPretrainer(
    optimizer_fn=torch.optim.Adam,
    optimizer_params=dict(lr=2e-2),
    mask_type='entmax',
    )

# Treinando a TabNet não supervisionada
unsupervised_model_no_preproc.fit(
    train_x,
    eval_set=[validation_x],
    max_epochs=1000 , patience=100,
    batch_size=256, virtual_batch_size=128,
    num_workers=0,
    drop_last=False,
    pretraining_ratio=0.8,
)
```

Metodologia - Classificação - Parâmetros

Um modelo supervisionado.

```
modelo_1 = TabNetClassifier(optimizer_fn=torch.optim.Adam,
                        optimizer_params=dict(lr=2e-2),
                        scheduler_params={"step_size":10,
                                          "gamma":0.9},
                        scheduler_fn=torch.optim.lr_scheduler.StepLR,
                        mask_type='entmax'
                       )
# Treinando a TabNet utilizando os pesos na etapa de Aprendizado Não
modelo_1.fit(
    train_x, train_y,
    eval_set=[(train_x, train_y), (validation_x, validation_y)],
    eval_name=['train', 'valid'],
    eval_metric=['auc', 'accuracy'],
    max_epochs=1000 ,patience=100,
    batch_size=256, virtual_batch_size=128,
    num_workers=0,
    weights=1,
    drop_last=False,
    from_unsupervised=loaded_pretrain
)
```

Metodologia - Classificação - Parâmetros

Parâmetros Variáveis:

N° Modelo	Treino - Validação (%)	Modelo Não Supervisionado	Modelo Supervisionado
1	70% - 30%	lr=0.02	lr=0.02
2	70% - 30%	lr=0.01	lr=0.01
3	80% - 20%	lr=0.02	lr=0.02
4	80% - 20%	lr=0.01	lr=0.01
5	67% - 33%	lr=0.01	lr=0.01

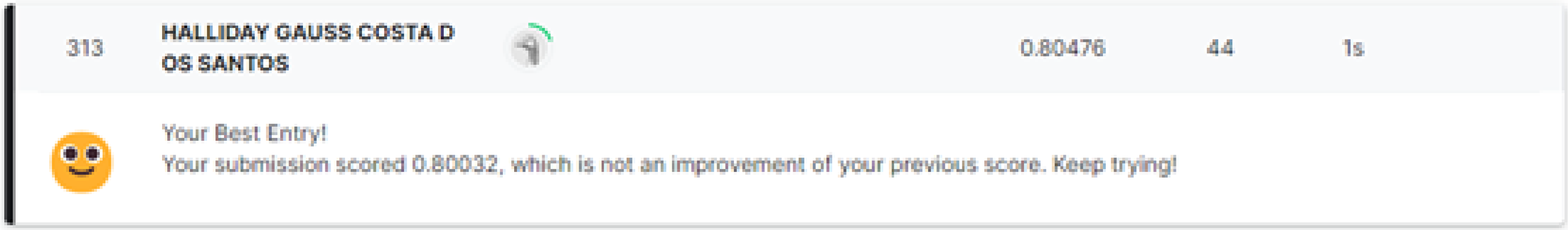
Metodologia - Classificação - Validação

Selecionado por 20 vezes uma porcentagem aleatória da base de treino entre 10% e 90%.

	precision (s)	recall (s)	f1_score (s)
0	0.81252(0.00835)	0.877596(0.003617)	0.843785(0.005267)
1	0.870136(0.004656)	0.802067(0.00734)	0.834698(0.004919)
accuracy	0.83939(0.004783)	0.83939(0.004783)	0.83939(0.004783)

Resultados

- **Pontuação de 80.41% no Kaggle.**
- **Top 1: 87.30% ; Top 5: 81.72%.**



Conclusão

- **Esse trabalho foi de suma importância para verificar os resultados de uma DNN voltada para dados tabulares, além da diversão obtida ao participar da competição.**
- **Verificar as intâncias difícieis de classificar, com e sem one-hot-encoding.**
- **XGBoost e TabNet.**

Referências

KAGGLE. Kaggle: Your Home for Data Science. Disponível em: <https://www.kaggle.com/>

Solheim Bojer, Casper, and Jens Peder Meldgaard. "Kaggle forecasting competitions: An overlooked learning opportunity." arXiv e-prints (2020): arXiv-2009.

Ekinci, Ekin, S. İlhan Omurca, and Neytullah Acun. "A comparative study on machine learning techniques using Titanic dataset." 7th international conference on advanced technologies. 2018.

Batista, Gustavo Enrique de Almeida Prado. Pré-processamento de dados em aprendizado de máquina supervisionado. Diss. Universidade de São Paulo, 2003.

Silva-Ramírez, Esther-Lydia, et al. "Missing value imputation on missing completely at random data using multilayer perceptrons." Neural Networks 24.1 (2011): 121-129.

Armina, Roslan, et al. "A review on missing value estimation using imputation algorithm." Journal of Physics: Conference Series. Vol. 892. No. 1. IOP Publishing, 2017.

Arik, Sercan Ö., and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 8. 2021.

Referências

Kothari Yash. "Predicting the Survivors of the Titanic -Kaggle, Machine Learning From Disaster". Academia.edu, 2018.

Alessandro, Lombardi, Zacchei Filippo Supervisors Polvani Niccolo, and Krapp Lucien Fabrice. "Mechanism of Action (MoA) Prediction-Kaggle Competition."

Obrigado!