Uma imagem com texto

Descrição gerada automaticamente

**Machine Learning aplicado à previsão do financiamento de um projeto**

**No âmbito de Inteligência Artificial**

**Aluno:**

**18825 – João Fernandes**

**18836 – Carlos Martins**

**Professor:**

**Joaquim Gonçalves**

**Licenciatura em Engenharia de Sistemas Informáticos**

Braga, janeiro, 2022

Índice

[Índice de Figuras 2](#_Toc93664406)

[1. Introdução 3](#_Toc93664407)

[1.1. Contextualização 3](#_Toc93664408)

[2. Implementação 4](#_Toc93664409)

[2.1. Dataset 4](#_Toc93664410)

[2.2. Orange 4](#_Toc93664411)

[3. Conclusão 9](#_Toc93664412)

[4. Bibliografia 10](#_Toc93664413)

# Índice de Figuras

[Figura 1 - Excerto do dataset utilizado 4](#_Toc93664328)

[Figura 2 - Tratamento de variáveis 4](#_Toc93664329)

[Figura 3 - Resultados dos modelos escolhidos 5](#_Toc93664330)

[Figura 4 - Modelo inicial 6](#_Toc93664331)

[Figura 5 - Selecionar linhas corretamente previstas como funded 6](#_Toc93664332)

[Figura 6 - Selecionar linhas corretamente previstas como not funded 6](#_Toc93664333)

[Figura 7 - Modelo final 7](#_Toc93664334)

# Introdução

O objetivo deste trabalho prático passou por interpretar um dataset com informações acerca de projetos. Este dataset era composto por um conjunto de variáveis, das quais uma chamada “Funded” que nos dizia se um projeto tinha sido aceite ou não para posteriormente ser financiado.

Posto isto, foram utilizados algoritmos e outras ferramentas disponíveis na plataforma Orange para obter as probabilidades de financiamento dados os diversos campos.

## Contextualização

O presente relatório foi realizado devido à necessidade de documentar todo o trabalho realizado, de forma a explicar todo o processo feito para conseguir chegar ao objetivo pretendido. O Orange permitiu isso, pois é machine learning open source e permite a visualização de dados. Podemos construir fluxos de trabalho de análise de dados de forma visual, com várias ferramentas e diversificadas.

# Implementação

## Dataset

O dataset em questão foi fornecido pelo professor e tivemos que fazer apenas ligeiras alterações para conseguirmos executá-lo no Orange.

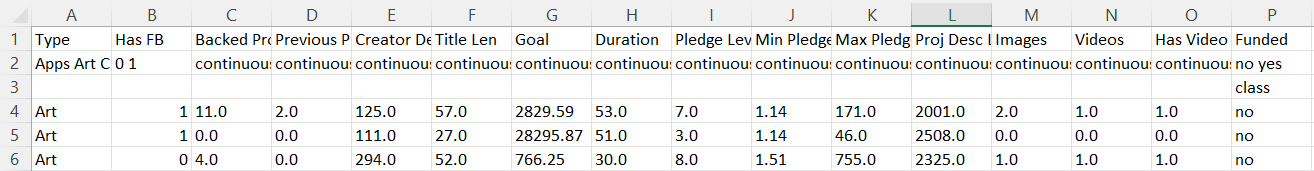


Figura 1 - Excerto do dataset utilizado

Se fosse utilizado o dataset no Orange como vemos na figura anterior, o mesmo não o ia conseguir executar corretamente por causa das linhas 2 e 3. Por isso, foram eliminadas essas linhas e seguimos com o projeto.

## Orange

Começamos por ler o ficheiro e definir o tipo de variável e o seu role. O resultado foi o seguinte:

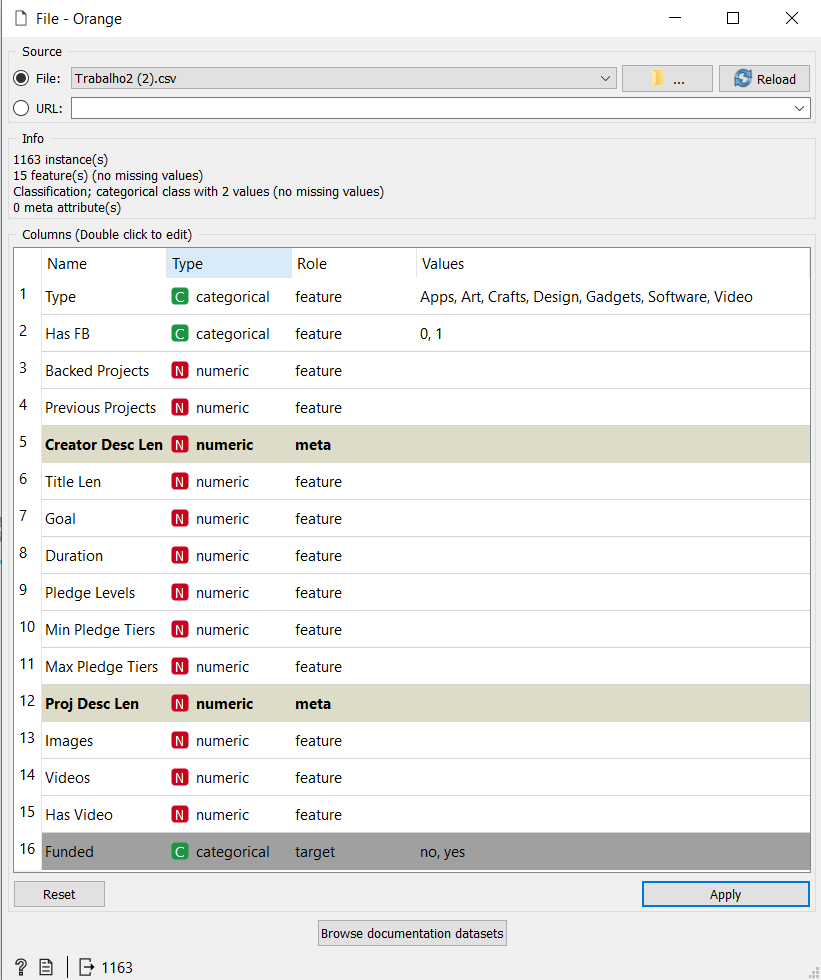


Figura 2 - Tratamento de variáveis

Como é possível, a variável “Funded” foi definida como target, porque é a variável dependente e que estamos a tentar prever. As variáveis “Creator Desc Len” e “Proj Desc Len” foram definidas como meta, pois como estavam definidas como feature, elas não acrescentavam valor nenhum. As restantes foram definidas como feature, porque vão ser usadas para prever o valor da variável target.

Posto isto, prosseguimos à aplicação de todos os modelos e decidimos escolher os três que nos davam melhores resultados. Com a sampling definida para “Random sampling” em que a repetição de treino/teste era feita 10 vezes e o tamanho do set de treino era 30%, os modelos com melhores resultados foram a regressão logística, redes neuronais e gradient boosting.

Verificamos que quando aumentávamos o tamanho do set de treino o modelo tinha taxas de acerto mais altas, mas decidimos reduzi-lo ao ponto de ter um resultado não muito mau, nem muito bom, porque se ficássemos num tamanho de treino grande a probabilidade de o modelo ficar “viciado” nos resultados de treino era elevada.

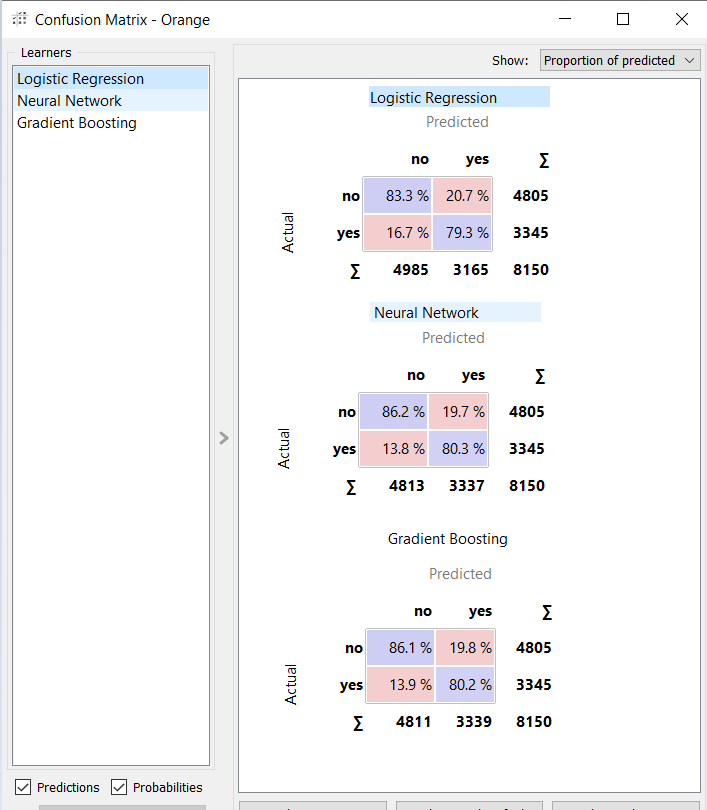


Figura 3 - Resultados dos modelos escolhidos

Pela matriz de confusão conseguimos verificar que em média 85.2% dos projetos que não foram aceites para financiamento são bem previstos. Já a média de projetos que forem aceites para financiamento e que foram corretamente previstos já não é tão alta, ficando nos 79.9%.

Até a este ponto, o modelo é o seguinte:

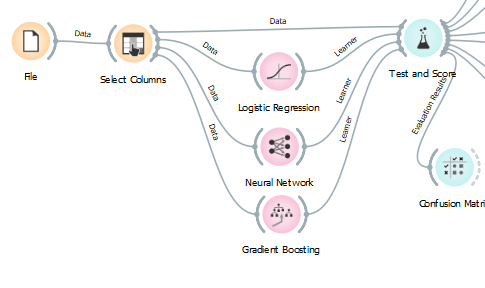


Figura 4 - Modelo inicial

Como os resultados ainda possuem uma grande margem de erro (previsões erradas) decidimos selecionar apenas as linhas que foram previstas corretamente como sim e não.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 5 - Selecionar linhas corretamente previstas como funded

Uma imagem com texto

Descrição gerada automaticamente

Figura 6 - Selecionar linhas corretamente previstas como not funded

Repetimos isto para cada um dos modelos utilizados.

Em seguida concatenamos os dados das diferentes tabelas geradas pela aplicação destas seleções e assim obtivemos todos os dados previstos corretamente (como funded e not funded).

Por fim, aplicamos os modelos outra vez e dessa forma obtivemos uma taxa de acerto substancialmente maior.

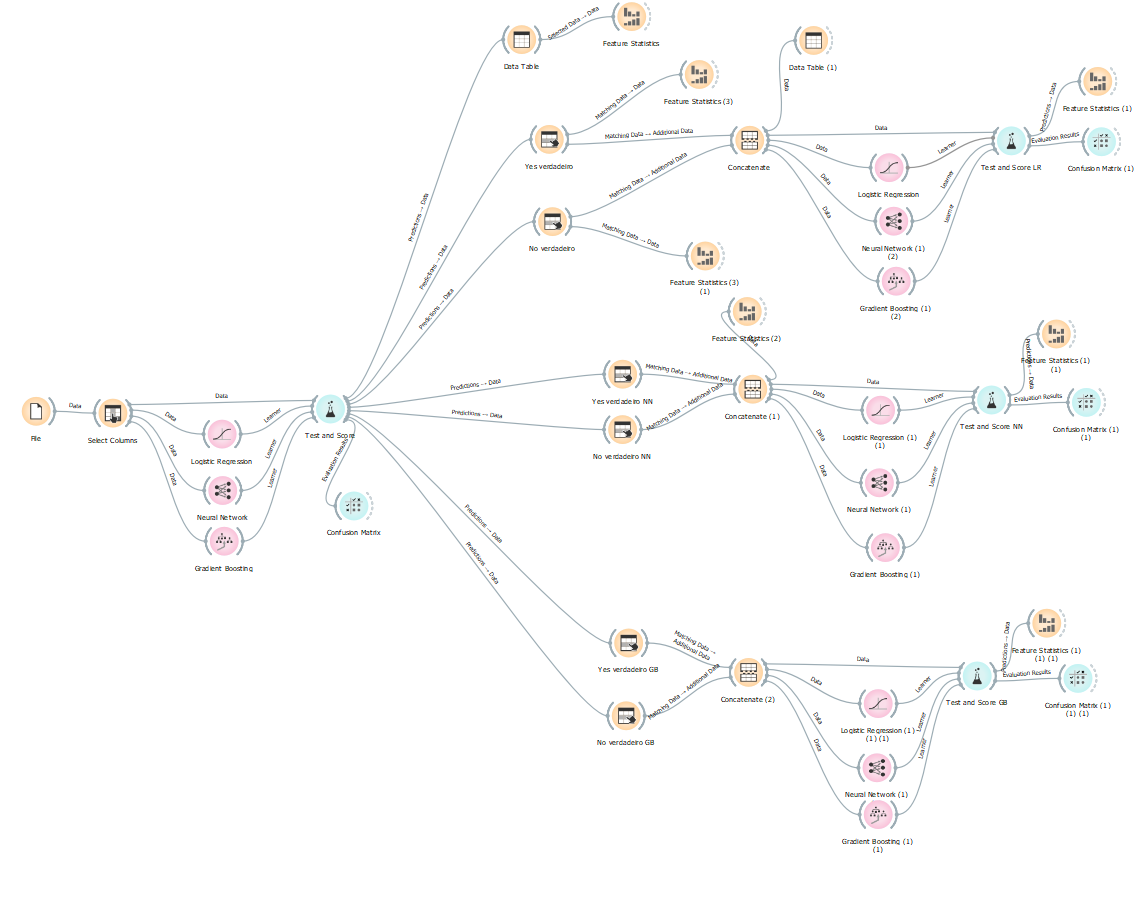


Figura 7 - Modelo final

A matriz de confusão com a média de previsões finais para cada um dos modelos é:

**Regressão logística**

Os valores utilizados para aqui foram os valores que no modelo inicial a regressão logística caracterizou corretamente como funded e not funded.

|  |  |  |
| --- | --- | --- |
|  | **no** | **yes** |
| **no** | 97,23 | 4,53 |
| **yes** | 2,77 | 95,47 |

**Redes neuronal**

Os valores utilizados para aqui foram os valores que no modelo inicial a rede neuronal caracterizou corretamente como funded e not funded.

|  |  |  |
| --- | --- | --- |
|  | **no** | **yes** |
| **no** | 98,40 | 2,37 |
| **yes** | 1,60 | 97,63 |

**Gradient boosting**

Os valores utilizados para aqui foram os valores que no modelo inicial o gradient boosting caracterizou corretamente como funded e not funded.

|  |  |  |
| --- | --- | --- |
|  | **no** | **yes** |
| **no** | 95,07 | 5,60 |
| **yes** | 4,93 | 94,40 |

Utilizámos também o módulo filter statistics e reparamos nos seguintes dados.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Projetos** |  |  |
|  | **Financiados** | **Não financiados** | |
| **Duração** | 30.30 | 36.10 |  |
| **Valor pedido** | 10915.60 | 685199.40 |  |
| **Qt. Imagens** | 15.50 | 3.01 |  |

As conclusões retiradas com o filter statistics foram que, em média, os projetos que são financiados possuem menos duração, pedem menos financiamento e possuem menos fotografias.

# Conclusão

Chegando ao fim do semestre e fazendo uma retrospetiva da cadeira, cremos que foi uma disciplina trabalhosa e complicada, mas muito interessante, porque a inteligência artificial é uma área importante e, certamente, vai ser ainda mais importante para o futuro.

Neste trabalho conseguimos abordar diversos temas lecionados nas aulas e também aprender sobre a existência de outros como, por exemplo, aprendemos que existem outros modelos para além dos apresentados nos powerpoints pelo professor Joaquim Gonçalves.

Para concluir, podemos dizer que, com este trabalho, conseguimos ganhar um maior conhecimento sobre o conteúdo lecionado em aula, porque tivemos que aplicar a maioria dele.

# Bibliografia

* PowerPoint e exemplos de código disponibilizados pelo professor Joaquim Gonçalves.
* <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/evaluate/testandscore.html>
* <https://datascience.stackexchange.com/questions/38188/classifying-data-in-orange-whats-the-difference-between-features-target-va>
* <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/naivebayes.html>
* <https://www.youtube.com/watch?v=hfMLZyWXBjI>
* <http://neohsu.com/2018/07/01/data-mining-practice-orange-dataset/>
* <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/loading-your-data/index.html>
* <https://www.youtube.com/watch?v=nNNl7Jws5Mk>