Universidade do Estado do Amazonas
Data: 26 de setembro de 2020
Disciplina: Fundamentos da Ciência de Dados
Professor(a): Carlos Maurício Serodio Figueiredo
Alunos: Lucas Pereira Reis

Atividade Avaliativa #4

Nesta atividade será realizado uma análise de algum *dataset* presente na ferramenta *Orange*, informando características, observações, *insights* e etc.

O *dataset* escolhido para esta atividade foi o *Kickstarter projects*. O site *Kickstarter* é uma plataforma de *crowdfunding* de projetos, onde você apresenta sua ideia de projeto (seja ela um jogo, aplicativo, filme, etc.) e pessoas podem apoiar financeiramente esta ideia, em troca de benefícios específicos (como edições *premium*, acesso antecipado, entre outros). Este conjunto de dados exibe todos os projetos desta plataforma no ano de 2016 entre os meses de fevereiro até maio, a Figura 1 apresenta informações básicas sobre o *dataset*.

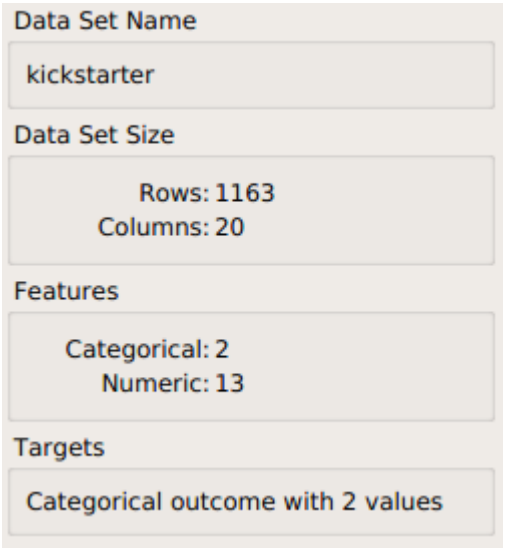

Figura 1: Informações do *dataset*.

Como podemos observar, o *dataset* apresenta um total de 1163 registros, ocupando 214.1 KB com 20 colunas. Este conjunto de dados tem como objetivo prever se um projeto será fundado, ou seja, atingiu seu objetivo financeiro em um período estabelecido. A Figura 2 exibe um exemplo dos registros encontrados e logo em seguida a Tabela 1 informa uma descrição sobre cada coluna.

	Funded	URL	Title	Year	Month	Type	Has FB	Backed Projects	Previous Projects	Creator Desc Len	Title Len	Goal	Duration	Pledge Levels	Min Pledge Tiers	Max Pledge Tiers	Proj Desc Len	Images	Videos	Has Video
1	no	https://www.kicksta...	Pixelstart: Ch...	2016	Apr	Art	1	11	2	125	57	2829.59	53	7	1.14	171.00	2001	2	1	1
2	no	https://www.kicksta...	Smart shop l...	2016	Apr	Art	1	0	0	111	27	28295.87	51	3	1.14	46.00	2508	0	0	0
3	no	https://www.kicksta...	Minimal Haus...	2016	Apr	Art	0	4	0	294	52	766.25	30	8	1.51	755.00	2325	1	1	1
4	no	https://www.kicksta...	NeoN: Alterin...	2016	Mar	Art	0	0	0	179	41	1439.10	24	5	7.00	141.00	3736	13	1	1
5	no	https://www.kicksta...	Nintendo NE...	2016	Mar	Art	0	0	0	51	41	1000.00	30	2	5.00	20.00	636	0	0	0
6	no	https://www.kicksta...	Day and Nig...	2016	Mar	Art	0	0	0	162	30	800.00	28	1	10.00	10.00	1199	0	0	0
7	no	https://www.kicksta...	Fund an Art ...	2016	Feb	Art	1	2	0	110	55	2000.00	50	9	5.00	500.00	2384	11	0	0
8	no	https://www.kicksta...	Trump that T...	2016	Mar	Art	1	1	0	12	58	1000.00	30	3	1.00	30.00	1603	1	1	1
9	yes	https://www.kicksta...	Once Upon a...	2016	Apr	Art	1	4	0	285	19	1079.32	30	16	7.00	157.00	2397	9	1	1
10	yes	https://www.kicksta...	Under the Ho...	2016	Apr	Art	1	12	8	55	49	10000.00	30	19	5.00	8000.00	1307	13	1	1
11	yes	https://www.kicksta...	KOKORO	2016	Mar	Art	0	0	0	637	6	792.28	45	5	6.00	113.00	3300	23	0	0
12	yes	https://www.kicksta...	Draw Cool Sh...	2016	Apr	Art	1	0	0	317	52	996.12	21	4	0.77	77.00	2677	6	1	1
13	yes	https://www.kicksta...	Elleforthеco...	2016	Mar	Art	0	0	0	204	45	2000.00	30	14	10.00	10000.00	864	10	0	0
14	yes	https://www.kicksta...	PT Apparel	2016	Mar	Art	1	0	0	337	10	600.00	45	6	5.00	100.00	1449	3	0	0
15	yes	https://www.kicksta...	Epocha - Han...	2016	Apr	Art	1	143	18	221	55	1000.00	25	3	20.00	500.00	3158	3	1	1
16	yes	https://www.kicksta...	The Little AB...	2016	Apr	Art	1	13	9	176	29	863.46	20	8	4.00	56.00	2141	9	1	1
17	yes	https://www.kicksta...	Burl & Fur	2016	Mar	Art	0	0	0	161	14	14000.00	30	7	10.00	250.00	10067	15	1	1
18	yes	https://www.kicksta...	Pens & Pedals	2016	Apr	Art	1	3	0	514	17	3000.00	21	5	10.00	300.00	4495	10	1	1
19	yes	https://www.kicksta...	BCU Illustrati...	2016	Feb	Art	1	0	0	26	21	1439.10	58	7	7.00	140.00	1935	2	0	0
20	yes	https://www.kicksta...	78 Tarot Car...	2016	Mar	Art	1	104	4	498	17	20000.00	29	14	1.00	250.00	17050	51	1	1

Figura 2: Tabela de dados do *dataset*

Atributo	Descrição
Funded	Atributo-alvo do *dataset*. Ele diz que se o projeto foi fundado ou não
URL	*Link* da página do projeto no *Kickstarter*
Title	Título do Projeto
Year	Ano que o projeto foi anunciado
Month	Mês que o projeto foi anunciado
Has FB	Especifica se o projeto tem página no *Facebook*
Backed Projects	Informa quantos projetos o criador do projeto já apoio
Previous Projects	Informa quantos projetos o criador do projeto já elaborou
Creator Desc Len	Tamanho de caracteres da descrição do criador
Title Len	Tamanho de caracteres do título
Goal	Quanto de dinheiro planeja-se alcançar
Duration	Duração prevista da campanha do projeto (em dias)
Pledge Levels	Níveis de benefícios do projeto
Min Pledge Tier	Valor mínimo dos benefícios presentes no projeto
Max Pledge Tier	Valor máximo dos benefícios presentes no projeto
Proj Desc Len	Tamanho de caracteres da descrição do projeto
Images	Quantas imagens a página do projeto possui
Videos	Quantos vídeos a página do projeto possui
Has Video	Especifica se a página do projeto possui vídeo

Tabela 1: Descrição de cada coluna do *dataset*.

Nesta etapa é necessário realizar uma análise exploratória dos dados para entender mais a fundo seu comportamento. O gráfico que permite ter a maior visibilidade neste *dataset* é o *Distributions*. O primeiro ponto para se observar é entender quantas projetos foram financiados com sucesso, em que mês eles foram criados e quais os tipos que mais são financiados.

O gráfico da Figura 3 mostra que 58.90% dos projetos presentes no *dataset* não foram fundados, mostrando um comportamento já esperado pois precisa-se do apoio financeiro do público para que seja possível seu sucesso, se as pessoas não aprovarem ou gostarem da ideia, é bem provável que o projeto não venha a ser finalizado.

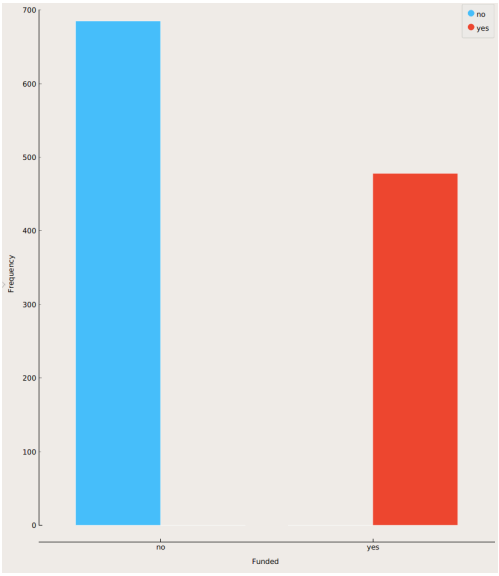

Figura 3: Gráfico indicando quantos projetos foram fundados ou não.

Após observar quantos projetos foram fundados ou não, é interessante observar como os projetos estão distribuídos ao longo dos meses. Analisando o gráfico da Figura 4 podemos entender que para esta classe (*month*) temos um desbalanceamento dos dados pois a maioria dos registros se concentram no mês de março (um total de 823), mostrando uma diferença muito grande no gráfico. Novamente observamos que em todos os meses (menos o de Maio, que possui apenas 4 registros) a quantidade de projetos não fundados é superior.

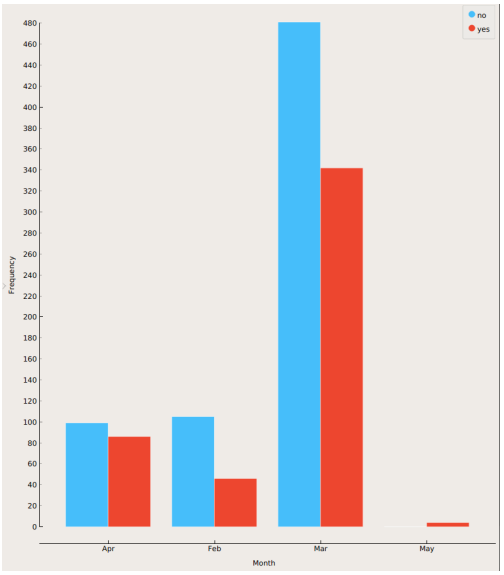

Figura 4: Gráfico indicando os projetos fundados no decorrer dos meses.

A última categoria para se analisar é o tipo do projeto, para entender quais os projetos de maior interesse do público. A Figura 5 exibe essa visão onde podemos observar uma clara preferência por determinado tipo. O público da plataforma aparenta não possuir muito interesse por projetos do tipo *App* ou *Software* que são justamente os tipos voltados para a área de tecnologia, mostrando que as pessoas que frequentavam nesta época não tinham desejo por contribuir em projetos desta área. Os projetos que mais alcançaram seus objetivos foram do tipo *Design* e *Video*, mostrando uma preferência mais artística das pessoas, os projetos do tipo *Video* costumam ser filmes ou episódios caseiros, e do tipo *Design* sendo propostas de produtos.

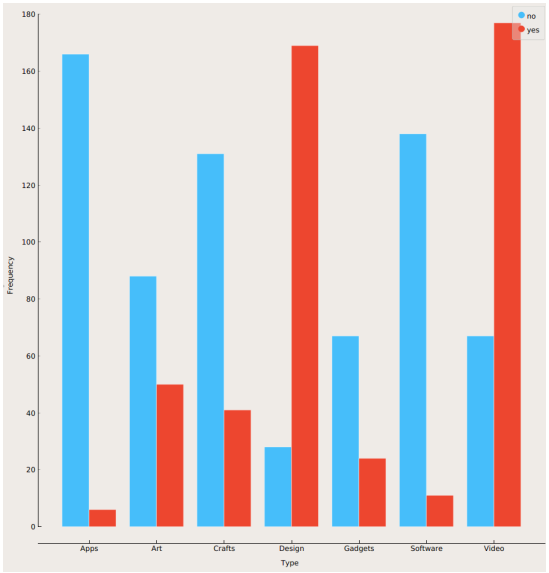

Figura 5: Gráfico indicando os projetos fundados por tipo.

Os gráficos que analisam variáveis categóricas se mostraram extremamente importantes para observamos tendências de quais projetos provavelmente serão fundados. Para este *dataset*, as variáveis numéricas das duas categorias do atributo-alvo são semelhantes, não havendo muita dispersão nos dados, mas ainda conseguimos extrair algum *insight* à seu respeito.

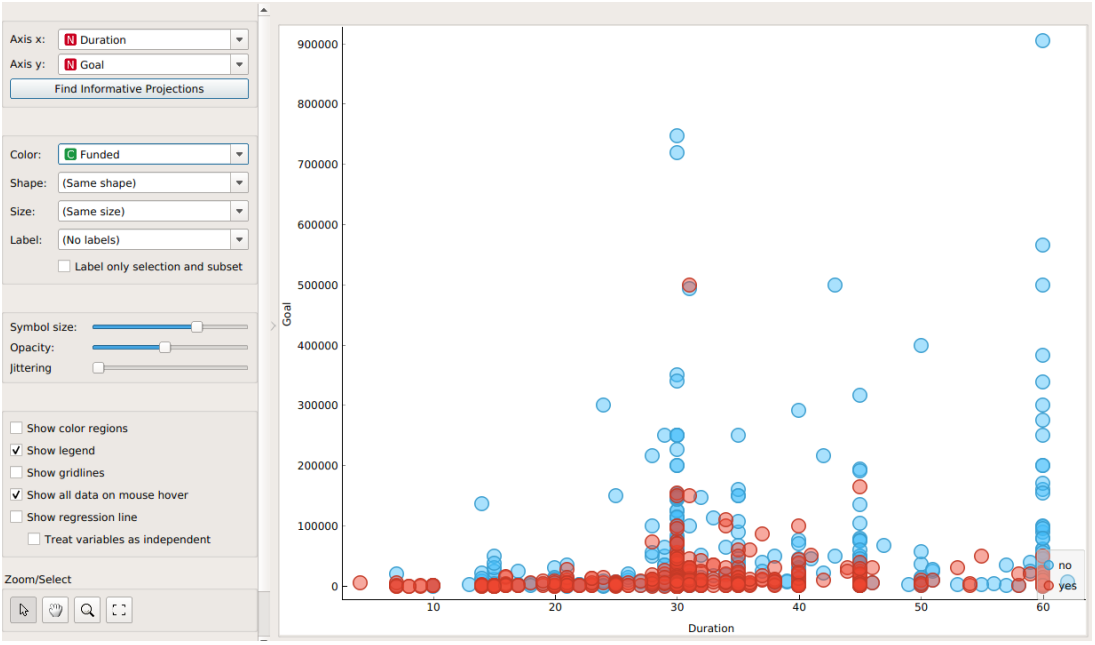

Figura 6: *Scatter Plot* dos atributos *Duration* x *Goal*.

A Figura 6 apresenta um gráfico comparando os atributos *Duration* e *Goal* onde podemos observar que projetos podem ser fundados independente da sua duração, seja ela de apenas alguns dias ou de 60 dias, mas um fator de extrema importância é que quanto maior o fundo financeiro necessário, maior as chances dele não ser fundado, podemos observar que quase todos os projetos acima de 20000 não tiveram seu objetivo comprido (temos apenas um *outlier* visível). Esse gráfico permite entender que o público da plataforma evita financiar projetos com *Goal* muito alto, e assim buscando projetos que tenham um valor menor, independente de sua duração.

Após a análise dos dados, inicia-se a etapa da criação de modelo. Primeiramente, optou-se por remover colunas indesejadas, pois algumas remetem apenas ao tamanho de caracteres de tal campo. Também foi removido a coluna ano pois o *dataset* só possui dados do ano de 2016 e a coluna URL, pois informa apenas o *link* de acesso. A Figura 7 exibe as colunas selecionadas e as colunas removidas.

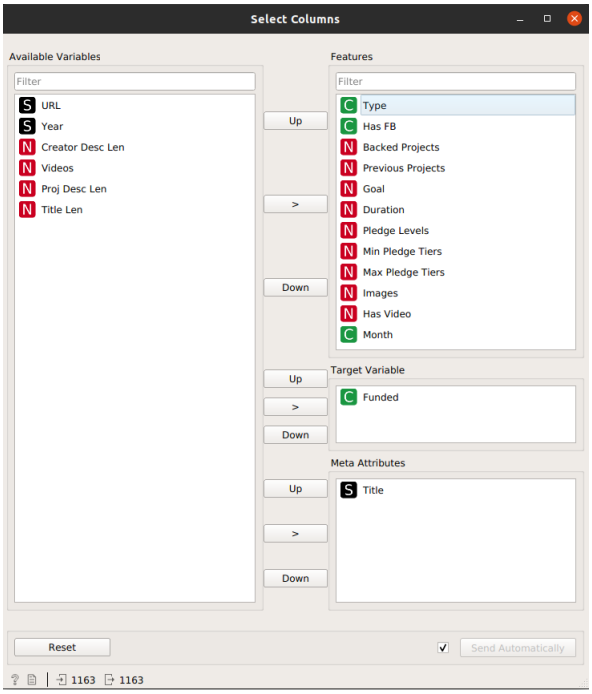

Figura 7: Visualização das colunas selecionadas e removidas.

Após selecionar as colunas desejadas, é necessário particionar o *dataset* em dois, um de treinamento e outro de teste através do *widget Data Sampler*, a proporção utilizada foi 70%. Para esta aplicação, foi selecionado diferentes modelos de aprendizagem supervisionada para treiná-los com o *dataset* de treinamento utilizando seus parâmetros padrões. Com os modelos treinados, é necessário avaliar qual obteu os melhores resultados, para isto foi utilizado o *widget Test and Score*, que recebe todos os modelos e os *datasets* de treinamento e teste, este *widget* exibe diferentes métricas conforme pode ser visto na Figura 8.

Model	AUC	CA	F1	Precision	Recall
SVM	0.859	0.796	0.797	0.798	0.796
RandomForest	0.902	0.833	0.832	0.832	0.833
NeuralNetwork	0.916	0.848	0.848	0.850	0.848
DecisionTree	0.739	0.790	0.787	0.788	0.790
AdaBoost	0.789	0.805	0.803	0.803	0.805

Figura 8: Métricas de desempenho dos modelos obtidos

A métrica utilizada para esta atividade será o *F1-Score*. Podemos observar que dentre os cinco modelos selecionados, a *Neural Network* obteve o melhor resultado, com um total de 84,8%, a *Random Forest* alcançou um valor próximo, sendo assim o segundo melhor modelo, É importante ressaltar que em todos os modelos foram utilizados seus parâmetros padrões.

Para a *Decision Tree*, que foi o terceiro melhor modelo, é possível visualizar como foi realizado suas decisões internamente através do *Tree Viewer*, conforme pode ser visto na Figura 9.

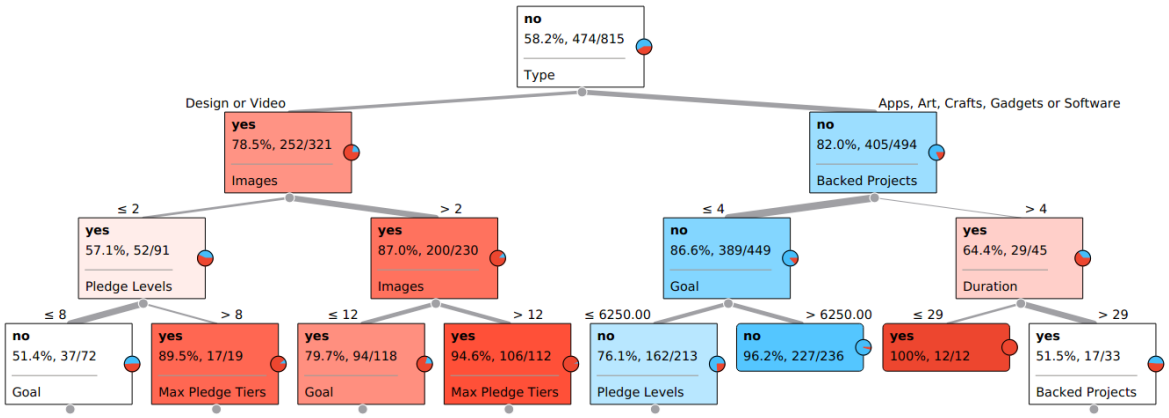

Figura 9: Visualização interna da *Decision Tree*.

Podemos observar logo de início que a coluna mais decisiva é a *Type*, pois ela divide bem que projetos foram fundados conforme visto na Figura 5. No segundo nível da árvore podemos identificar os atributos *Images* e *Backed Projects* que também se tornaram importantes para a decisão, podemos deduzir então que se projetos do tipo *Design* ou *Video* tiveram pelo menos duas imagens, é bem provável que será fundado. Do outro lado da árvore, se o autor já tiver ajudado outros projetos e se a campanha durar bastante, ainda existe a chance do projeto ser fundado.

Por fim, além de observar as métricas obtidas para cada modelo, é interessante também observar sua matriz de confusão. O *widget* não permite visualizar as matrizes de todos os modelos, então escolheu-se apenas a matriz do melhor modelo, a *Neural Network*, que pode ser visto na Figura 10.

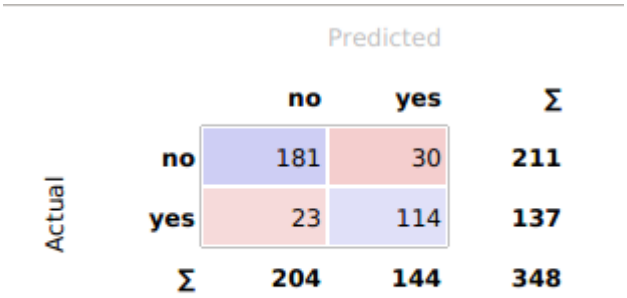

Figura 10: Matriz de confusão da *Neural Network*.

A utilização do *Orange* trouxe uma grande facilidade para analisar dados, treinar e avaliar modelos. Vale ressaltar que como estamos limitados aos *widgets* da ferramenta, não temos tanta flexibilidade quanto teríamos utilização programação em si, mas para projeto simples ou apenas para estudo, é possível desenvolver rapidamente um estudo completo. Por fim, a Figura 11 mostra uma visão geral dos *widgets* utilizados para esta atividade.

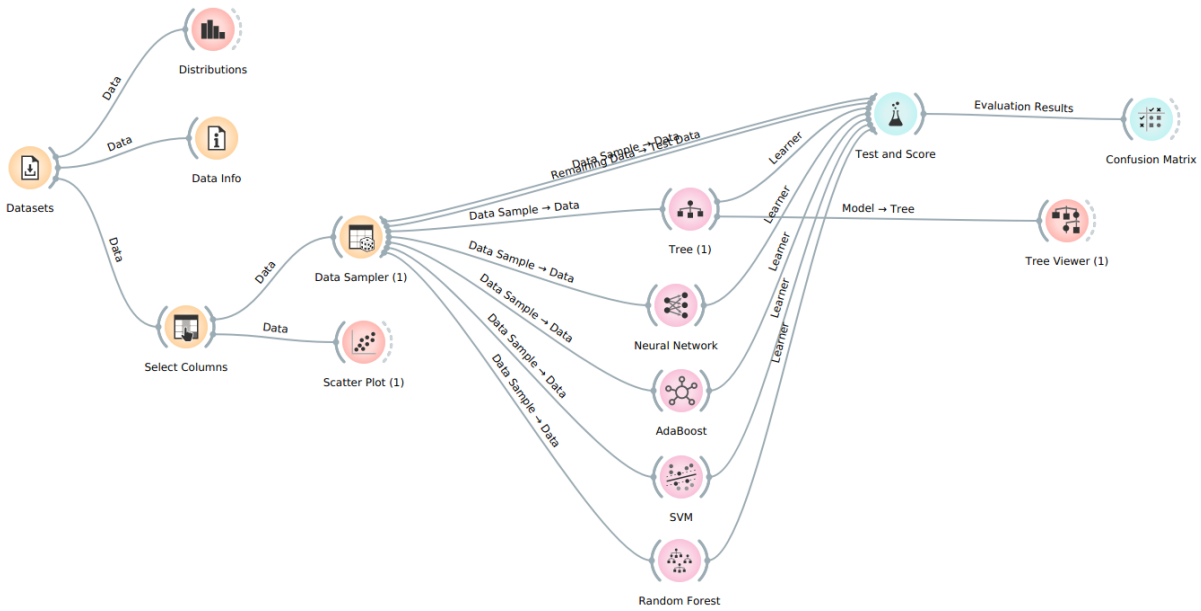

Figura 11: Visão geral da aplicação desenvolvida.