Desafio Movile

Neste desafio, estudamos o problema de classificação de spams em mensagens de SMS.

Foram fornecidas pela Wavy amostras de mensagens de diferentes operadoras.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
pd.set_option('display.max_colwidth', -1)
```

```
dataset = pd.read_csv('SPAM Data _ Akari - SPAM Data.csv')
```

```
print(len(dataset))
dataset.sample(frac=1).head(10)
```

```
997
```

.dataframe tbody tr th:only-of-type { vertical-align: middle; } .dataframe tbody tr th { vertical-align: top; } .dataframe thead th { text-align: right; }

	vendor	mensagem	destino	spam	total
619	difize	Oi FRANCISCO! Abasteca e pontue com seu cartao fidelidade na REDE CAXUXA e ganhe o DOBRO de pontos em sua primeira compra! Oferta valida por 30 dias.	NaN	False	239
950	mailserr	Traga seu numero para CLARO ganhe mais internet para navegar, Whastapp a vontade, ligacoes ILIMITADAS para todo Brasil. Rocell Digital whastapp 55991282580	CLARO	False	624
118	zootude	Carlos, a CLARO ainda precisa de sua ligacao! Retorne ate o final do dia de hoje no 30038080 ou 08002088080 ou acesse Claropaguefacil.com.br	NaN	False	133
850	difize	REDE GMAXX: Oi JOSE! Preparamos uma oferta para voce! Abasteca na REDE GMAXX e ganhe o dobro de pontos em sua proxima compra! Valido 30 dias!	NaN	False	233
896	centigen	Sebastiana,mantenha seu plano da TIM em dia! Efetue o pagamento da divida.Evite a sua permanencia do debito,caso ja tenha pago, favor desconsiderar.	NaN	False	104
279	centigen	Franciele,mantenha seu plano da TIM em dia! Efetue o pagamento da divida.Evite a sua permanencia do debito,caso ja tenha pago, favor desconsiderar.	NaN	False	163
864	centigen	Renato,mantenha seu plano da TIM em dia! Efetue o pagamento da divida.Evite a sua permanencia do debito,caso ja tenha pago, favor desconsiderar.	NaN	False	608
225	centigen	Eliene,mantenha seu plano da TIM em dia! Efetue o pagamento da divida.Evite a sua permanencia do debito,caso ja tenha pago, favor desconsiderar.	NaN	False	128

	vendor	mensagem	destino	spam	total
112	centigen	Caio,mantenha seu plano da TIM em dia! Efetue o pagamento da divida.Evite a sua permanencia do debito,caso ja tenha pago, favor desconsiderar.	NaN	False	215
595	difize	Oi EMERSON! Abasteca e pontue com seu cartao fidelidade na REDE CAXUXA e ganhe o DOBRO de pontos em sua primeira compra! Oferta valida por 30 dias.	NaN	False	103

Na amostra acima, podemos ver que temos 5 features: * Vendor: Nome da empresa que enviou a mensagem * Mensagem: texto da mensagem * Destino: Operadora do destinatário da mensagem * Spam: se foi classificada como spam (true) ou não (false) * Total: Quantidade de cópias da mensagem enviadas

Abordagens

Existem inúmeras formas de explorar o problema. Primeiro, devemos olhar bem as características dos nossos dados.

```
dataset['spam'].value_counts()
```

```
False    991
True       6
Name: spam, dtype: int64
```

Aqui temos um claro problema de *skewed classes*, em que uma classe (negativa) é muito mais predominante do que a outra (positiva). Casos assim não são triviais de serem solucionados. Vamos então primeiramente explorar os dados e analisar uma possível solução.

Podemos começar vendo o que temos nesses SMSs classificados como spam

```
dataset.loc[dataset['spam'] == True]
```

.dataframe tbody tr th:only-of-type { vertical-align: middle; } .dataframe tbody tr th { vertical-align: top; } .dataframe thead th { text-align: right; }

	vendor	mensagem	destino	spam	total
772	mailserr	OI. Temos uma OFERTA especial de CELULAR ILIMITADO para todo BRASIL + 10GB de INTERNET. responda OK que retornarmos para voce ou ligue 0800 291 2253	NaN	True	2276
783	quasiyo	Ola, somos da TIM! Parabens! Seu chip esta ativado, realize uma ligacao de 30 segundos para confirmar o funcionamento. Digite se ja esta utilizando, 2 se nao.	NaN	True	495
785	quasiyo	Ola, somos da TIM! Seu chip foi ativado e liberado para fazer ligacao. Utilize com urgencia, p confirmar o sinal! Digite 1 se ja esta utilizando, 2 se nao.	NaN	True	1041

	vendor	mensagem	destino	spam	total
786	quasiyo	Ola, somos da TIM! Seu chip foi ativado e liberado para fazer ligacao. Utilize com urgencia, p confirmar o sinal! Digite 1 se ja esta utilizando, 2 se nao.	NaN	True	458
787	quasiyo	Ola, somos da TIM! Seu chip ja foi ativado e esta gerando fatura. Digite 1 se ja realizou alguma ligacao com seu chip novo, 2 se nao.	NaN	True	283
788	quasiyo	Ola, somos da TIM! Verificamos que voce ainda nao utilizou seu chip, e estamos gerando fatura! Faca uma ligacao com urgencia usando o seu chip da TIM.	NaN	True	193

A primeira coisa que nos chamou a atenção foi a grande semelhança entre a classe positiva e negativa. Notei apenas erros sutis de ortografia ou gramática. Além disso, 5 dos 6 SMS classificados como spams foram enviados pela "quasiyo". Vamos verificar se todos os SMS enviados pela quasiyo são spams.

```
dataset.loc[dataset['vendor'] == 'quasiyo'].head(5)
```

.dataframe tbody tr th:only-of-type { vertical-align: middle; } .dataframe tbody tr th { vertical-align: top; } .dataframe thead th { text-align: right; }

	vendor	mensagem	destino	spam	total
238	quasiyo	Esta difIcil lidar com as taxas de emprestimos? CLARO Q NAO! So RedeCifrao lhe apresenta a menor taxa e ainda diminui o valor da sua parcela! Resp. LIMITE (1/2)	NaN	False	781
239	quasiyo	Esta difIcil lidar com as taxas de emprestimos? CLARO Q NAO! So UBLA lhe apresenta a menor taxa e ainda diminui o valor da sua parcela! Resp. LIMITE p/+ info.	NaN	False	1128
240	quasiyo	Esta dificil lidar com as taxas de emprestimos? CLARO QUE NAO! So a REDE CIFRAO lhe apresenta a menor taxa e ainda diminui o valor da sua parcela! Resp. LIMIT	NaN	False	1923
241	quasiyo	Esta dificil lidar com as taxas de emprestimos? CLARO QUE NAO! So a REDE CIFRAO lhe apresenta a menor taxa e ainda diminui o valor da sua parcela! Resp. LIMITE	NaN	False	186
242	quasiyo	Esta dificil lidar com as taxas de emprestimos? CLARO QUE NAO! So a RedeCifrao lhe apresenta a menor taxa e ainda diminui o valor da sua parcela! Resp. LIMITE	NaN	False	1067

Confirmamos que vários outros SMSs enviados pela quasiyo não foram classificados como spam. Será que existe algum padrão nas mensagens spams?

```
dataset[dataset['mensagem'].str.contains("Ola, somos da TIM!")]
```

.dataframe tbody tr th:only-of-type { vertical-align: middle; } .dataframe tbody tr th { vertical-align: top; } .dataframe thead th { text-align: right; }

	vendor	mensagem	destino	spam	total

	vendor	mensagem	destino	spam	total
783	quasiyo	Ola, somos da TIM! Parabens! Seu chip esta ativado, realize uma ligacao de 30 segundos para confirmar o funcionamento. Digite se ja esta utilizando, 2 se nao.	NaN	True	495
785	quasiyo	Ola, somos da TIM! Seu chip foi ativado e liberado para fazer ligacao. Utilize com urgencia, p confirmar o sinal! Digite 1 se ja esta utilizando, 2 se nao.	NaN	True	1041
786	quasiyo	Ola, somos da TIM! Seu chip foi ativado e liberado para fazer ligacao. Utilize com urgencia, p confirmar o sinal! Digite 1 se ja esta utilizando, 2 se nao.	NaN	True	458
787	quasiyo	Ola, somos da TIM! Seu chip ja foi ativado e esta gerando fatura. Digite 1 se ja realizou alguma ligacao com seu chip novo, 2 se nao.	NaN	True	283
788	quasiyo	Ola, somos da TIM! Verificamos que voce ainda nao utilizou seu chip, e estamos gerando fatura! Faca uma ligacao com urgencia usando o seu chip da TIM.	NaN	True	193

Como apenas mensagens de spams tem o texto "Ola, somos da TIM!", uma solução seria classificar todas as mensagens com esse texto como spam. Mas claramente não seria um bom classificador.

Com estes testes, concluímos que as mensagens spams e não spams são muito semelhantes. É muito complicado criar um modelo de aprendizado de máquina para fazer uma tarefa que nem mesmo nós, humanos, seríamos capazes de fazer.

O único padrão que pudemos encontrar nos spams foi erros de ortografia e gramática. Provavelmente as mensagens foram classificadas como spams por este motivo. Outras features, como o total de cópias da mensagem, também não apresentaram nenhum padrão. Poderíamos propor filtrar as mensagens através de um corretor ortográfico, mas aparentemente é um padrão nas mensagens de SMS não utilizar acentuação, o que nos geraria muitos falsos positivos.

Podemos primeiramente fazer uma regressão, observar os resultados e, se necessário, melhorar a solução. Para isso, precisamos separar o conjunto em treino, validação e teste.

```
label = dataset['spam']
data = dataset.drop('spam', axis=1)

#Separar em treino e teste aleatoriamente
```

```
X_train, X_test, y_train, y_test =  train_test_split(data, label, test_size=0.15,
random_state=5)

#Separar em treino e validação
X_train, X_valid, y_train, y_valid =  train_test_split(X_train, y_train,
test_size=0.3, random_state=5)
```

```
```