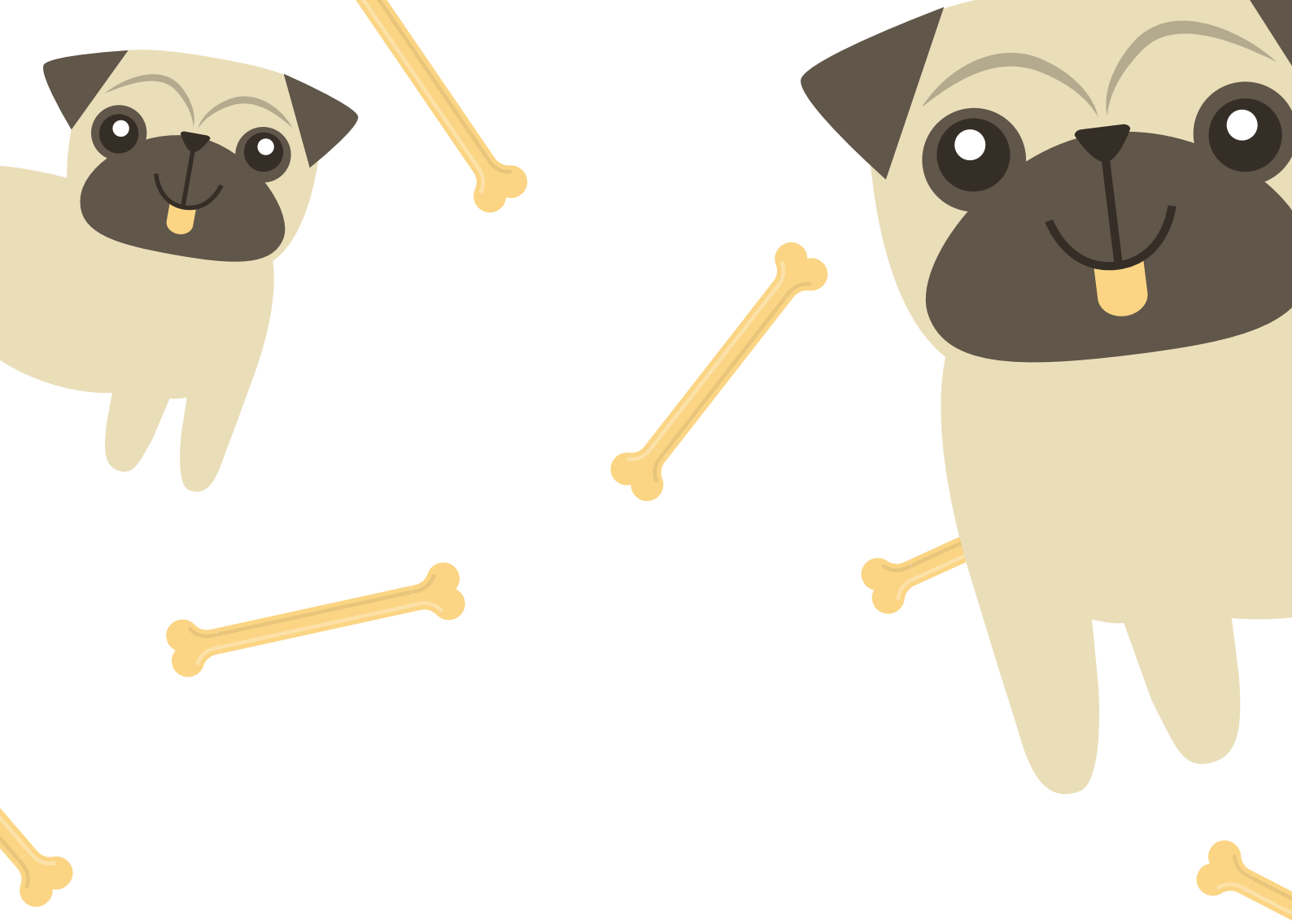


CAPSTONE - COURSERA

ADOPTION SPEED IN MALAYSIA

Rodrigo Alcarva

INDICE

1 - INTRODUÇÃO

2 - REVISÃO DA LITERATURA

3 - TRABALHO EMPIRICO

3.1 - CONTEXTO DE DADOS
3.2 - RECOLHA DE DADOS
3.3 - PREPARAÇÃO DE DADOS
3.4 - EXPLORAÇÃO DE DADOS
3.5 - MODELIZAÇÃO DE DADOS
3.6 - AVALIAÇÃO
3.7 - APRESENTAÇÃO DE RESULTADOS

4 - RESULTADOS E DISCUSSÃO

5 - CONCLUSÃO

6 - REFERÊNCIAS

INTRODUÇÃO

Com o intuito de compreender que features influenciam a rapidez de adoção de um animal de estimação decidimos elaborar o trabalho que se segue. Embora todo este processo seja feito recorrendo a um dataset um pouco restrito visto que, apenas nos dá informação sobre a Malásia, esperamos que os resultados que podem provir do nosso trabalho possam ser usados para analisar a rapidez de adoção bem como os features que a influenciam noutros países do mundo.
Assim sendo, neste trabalho esperamos que os nossos conhecimentos de programação bem como as ferramentas de Machine Leartning aprendidas no âmbito da cadeira nos possibilitem entender padrões de adoção e características que influenciam os mesmos.
Sem nenhuma informação e usando só a nossa cultura geral, temos a ideia generalizada que os animais como câes e gatos são usados na gastronomia asiática, coisa um facto que cada vez mais está a diminuir,
Neste momento cada vez a preocupação pelos animais está em crescimento,e tendo isso em conta, decidimos com o dataset que temos, analisar de algum modo esse crescimento,

REVISÃO DA LITERATURA

Em relação a artigos que possam ajudar a identificar trabalhos semelhantes desenvolvidos por outros autores, foi difícil encontrá-los, a nível de artigos académicos não encontrámos, mas no local onde fomos buscar o dataset que usámos (www.kaggle.com), encontrámos trabalhos sobre esse dataset, e serão esses que iremos colocar neste relatório.
Dos trabalhos que encontrámos muitos usavam técnicas que nunca aprendemos. Nós como pessoas menos experientes nesta área, fomos lendo vários trabalhos, também para ir recolhendo ideias para o nosso trabalho, mas também ir aprendendo o que devemos fazer ou não.

Trabalhos:

https://www.kaggle.com/jaseziv83/an-extensive-eda-of-petfinder-my-data

https://www.kaggle.com/erikbruin/petfinder-my-detailed-eda-and-xgboost-baseline

https://www.kaggle.com/wrosinski/baselinemodeling

TRABALHO EMPIRICO

3.1 CONTEXTO DE DADOS

É uma etapa importante de todo o ciclo, onde usamos a técnica dos 5 W´s (Why, Who, What, Where, When).
Assim sendo, como já foi dito fomos buscar os nossos datasets ao Kaggle, onde obtámos pelo tema de adoção de cães e gatos na Malásia, onde vimos que tinham excelentes datasets que poderíamos explorar.
Usando a técnica dos 5 W´s, o porquê da análise deste dataset, explorámos aqui a parte da cultura deste país. Em relação de quem iremos analisar, iremos analisar os cães e gatos adotados na Malásia. Relativo a o que iremos analisar, será a velocidade de adoção de cada um dos animais e quais os fatores mais importantes. Em relação a onde, a análise como já foi dito será na Malásia. E relativo a quando, os dados são do ano passado.

3.2 RECOLHA DE DADOS

Para enquadrar-mos o professor no nosso trabalho, começámos por fazer pesquisa por datasets no site da camara de Lisboa. Começamos por explorar um dataset sobre turismo em Lisboa, um tema muito interessante para nós, e depois de termos explorado e analisado com grande detalhe, quando chegámos à parte das tecnicas de modelização reparámos que afinal o nosso dataset não era bom.
Acabámos por recomeçar do 0, e aí fomos ao site Kaggle onde havia muitos datasets por explorar, sendo que o que escolhemos, tinha uma otima classificação e esta no top. Assim decidimos por analisar este dataset com este tema.
Aqui recolhemos vários CSV´s :

- train.csv
- breed_labels.csv
- labels.csv
- state_labels.csv

3.3 PREPARAÇÃO DE DADOS

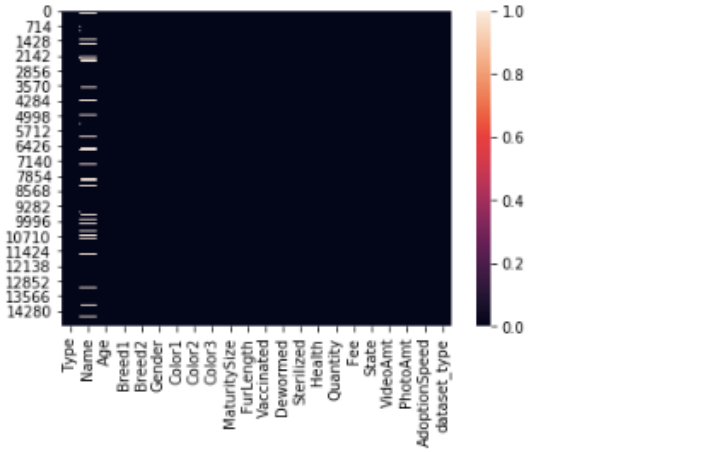
Em relação à preparação de dados, os dados textuais e distintos não são interessantes para a nossa análise pelo que eliminámos estas colunas. Porém, decidimos manter as colunas Name e Breed pois embora sejam textuais e distintas queremos ver se têm influência no target: AdoptionSpeed.

```
In [5]: df = df.drop(['Description','RescuerID','PetID'], axis=1)
```

No ponto dos Missing Values, averiguámos a sua existência. Caso se verificasse que existissem iriamos tratá-los para, no final, obtermos melhores resultados.

```
In [7]: df.isnull().sum()
Out[7]: Type              0
        Name           1257
        Age               0
        Breed1            0
        Breed2            0
        Gender            0
        Color1            0
        Color2            0
        Color3            0
        MaturitySize      0
        FurLength         0
        Vaccinated        0
        Dewormed          0
        Sterilized        0
        Health            0
        Quantity          0
        Fee               0
        State             0
        VideoAmt          0
        PhotoAmt          0
        AdoptionSpeed     0
        dataset_type      0
        dtype: int64
```

```
In [8]: sb.heatmap(df.isnull())
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x2cd3df64ef0>
```



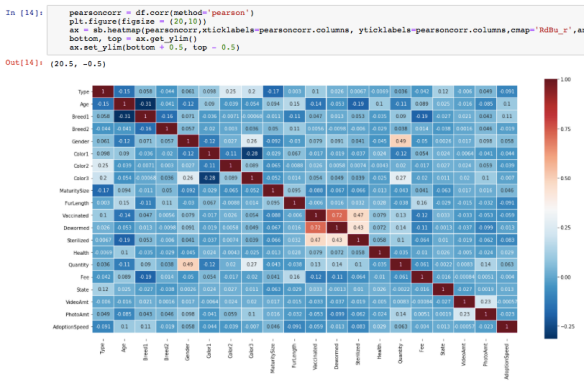
Como podemos ver, neste dataset os únicos Missing Values que encontrámos foram na coluna Name, assim tratámos estes valores da seguinte maneira:

```
In [9]: df['Name'] = df['Name'].fillna('Unknown')
```

A seguir aos Missing Values, realizámos Data Balancing para cada coluna.

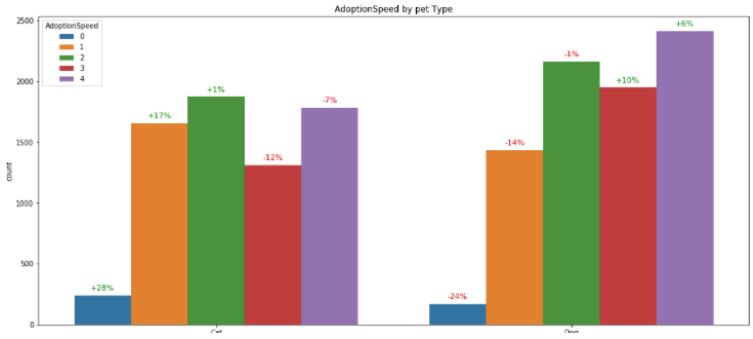
3.4 EXPLORAÇÃO DE DADOS

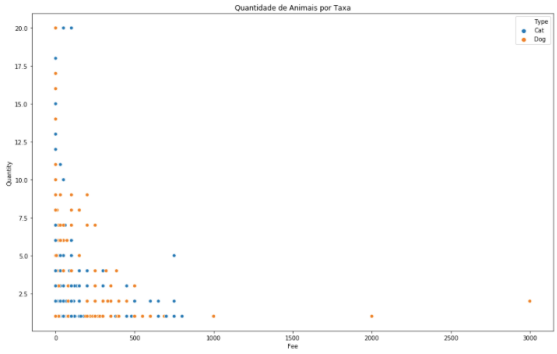
Em relação à exploração de dados fizemos 3 pontos, a Granularidade, a Variância e a Correlação. Mas também usámos a construção de gráficos para perceber melhor o target e as features que faziam sentido usar.

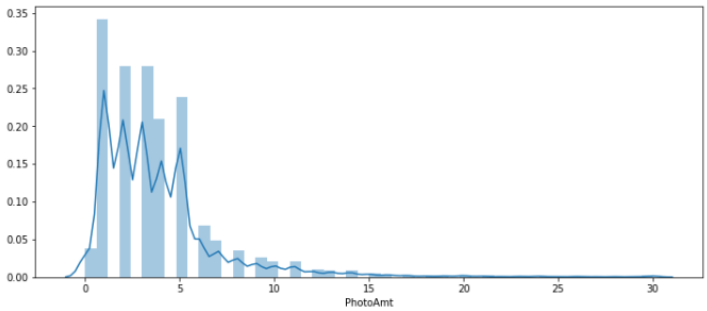


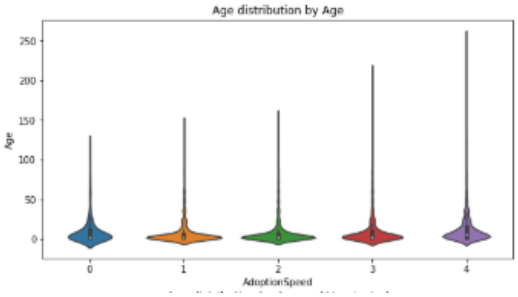
Na figura do lado, vemos a correlação entre cada coluna, e podemos reparar que a zona central, em relação a Vaccinated e Dewormed existe uma grande correlação. Onde serão colunas que no final iremos concluir que não são assim tão importantes para o AdoptionSpeed

A seguir à correlação, explorámos a parte da construção de gráficos, onde ligámos várias colunas ao nosso target.
Em relação aos gráficos para além de termos utilizado o nosso target, tabmém utilizámos o Type, Name, Age, Breed, Gender, Color, MaturitySize, FurLength, Health, Fee e State.
Deixamos aqui alguns exemplos:









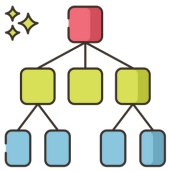
3.5 MODELIZAÇÃO DE DADOS

Em relação à modelização de dados, recorremos a 5 técnicas: Regressão Linear com OLS, Regressão Linear com RDSE. Regressão Logística e Árvore de Decisão com Regeressão.

3.6 AVALIAÇÃO

Regressão Linear com OLS

Uma vez que, neste trabalho são desejados grandes níveis de detalhe decidimos recorrer ao uso da Regressão Linear com statsmodel. Porém, embora os resultados previstos se encontrem bastante próximos dos resultados obtidos o coeficiente de determinação tem um valor muito baixo pelo que, decidimos não continuar a avaliação do nosso dataset através deste método.



Árvore de Decisão com Regressão

De modo a expandirmos os nossos modelos de previsão decidimos sair da área das Regressões Lineares e mergulhar na área das Decision Tree. Porém, através deste método só conseguimos observar uma pequena parte da amostra pelo que, ao avaliarmos as percentagens de erro verificámos que estas são muito altas, acima dos 100% em alguns casos, logo o modelo preditivo tem bastantes chances de estar errado. Assim, descartámos este método na análise preditiva do nosso dataset.

Regressão Logistica

A Regressão Logística é um caso especial de regressão linear que tem como objetivo prever um modelo de valores a partir de um conjunto de observações. O valor de precisão neste modelo é de 0.8 numa escala de 0 a 1, o que nos indica que estamos perante um bom modelo. Porém, é necessário continuar a investigar. Para isso recorremos à Regressão Linear com RDSE.

Regressão Linear com RDSE

Decidimos analisar apenas as feautures Fee e Age pois, o conjuntos de observação das outras features são demasiado pequenos para poderem ser avaliados. Não só o valor Final de RMSE é muito baixo como os valores previstos se assemelham aos valores esperados pelo que, podemos concluir que, as previsões feitas anteriormente estão de acordo com as previsões do nosso modelo.

LgR

Análise de Clusters

Com o intuito de corroborar as observações registadas acima, decidimos recorrer à anãlise de clusters para analisar como a Fee e a Age influenciam a Adoption Speed . Após a análise ficámos aptos para corroborar as observações acima bem como garantir mais uma vez que as nossas previsões estavam corretas.

3.7 APRESENTAÇÃO DE RESULTADOS

LnR w/ stastemodels

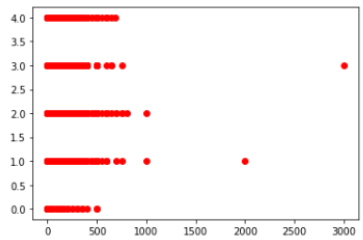
OLS Regression Results

Dep. Variable:	AdoptionSpeed	R-squared:	0.062
Model:	OLS	Adj. R-squared:	0.062
Method:	Least Squares	F-statistic:	71.23
Date:	Tue, 10 Dec 2019	Prob (F-statistic):	3.35e-197
Time:	22:52:27	Log-Likelihood:	-23237.
No. Observations:	14993	AIC:	4.650e+04
Df Residuals:	14978	BIC:	4.662e+04
Df Model:	14		
Covariance Type:	nonrobust		

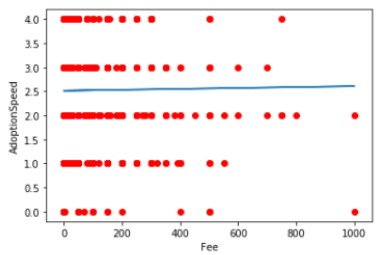
	coef	std err	t	P>|t|	[0.025	0.975]
const	-35.3731	12.014	-2.944	0.003	-58.922	-11.824
Type	-0.1764	0.019	-9.059	0.000	-0.215	-0.138
Age	0.0090	0.001	15.700	0.000	0.008	0.010
Breed1	0.0030	0.000	17.527	0.000	0.003	0.003
Breed2	0.0001	7.8e-05	1.723	0.085	-1.85e-05	0.000
Gender	0.0824	0.016	5.196	0.000	0.051	0.114
MaturitySize	0.0692	0.018	3.955	0.000	0.035	0.104
FurLength	-0.1975	0.016	-12.236	0.000	-0.229	-0.166
Vaccinated	-0.1033	0.021	-4.869	0.000	-0.145	-0.062
Dewormed	0.0953	0.020	4.820	0.000	0.057	0.134
Sterilized	-0.1312	0.019	-6.835	0.000	-0.169	-0.094
Health	0.1839	0.047	3.891	0.000	0.091	0.277
Quantity	0.0396	0.007	5.379	0.000	0.025	0.054
Fee	0.0004	0.000	2.995	0.003	0.000	0.001
State	0.0009	0.000	3.104	0.002	0.000	0.001

Omnibus:	3159.362	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	651.568
Skew:	-0.141	Prob(JB):	3.26e-142
Kurtosis:	2.018	Cond. No.	5.33e+07

LnR w/ RMSE

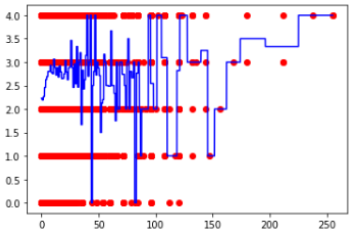


O Valor Final de RMSE é 1.1832541457255896

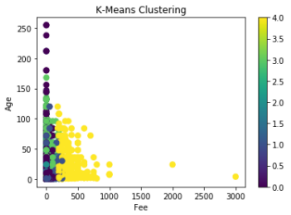


	precision	recall	f1-score	support
0	1.00	0.31	0.47	116
1	0.91	1.00	0.95	956
2	0.63	0.98	0.76	1202
3	0.95	0.30	0.46	990
4	1.00	1.00	1.00	1234
accuracy			0.82	4498
macro avg	0.90	0.72	0.73	4498
weighted avg	0.87	0.82	0.79	4498

Decision Tree Regression



Cluster Analysis



RESULTADOS E DISCUSSÃO

Após a análise dos resultados obtidos através dos métodos de modelação de dados supracitados conseguimos compreender que os features que realmente importam para a velocidadede adoção são: a age e a fee .
Os modelos preditivos corroboraram o que já tinhamos verificado na análise inicial: os animais mais novos são adotados muito mais rapidamente enquanto que, os animais mais velho têm muito mais dificuldade em serem adotados.
Embora os animais com uma fee nula sejam mais adotados verificamos também que, os animais com grandes fees são adotados mais rapidamente. Atribuímos este acontecimento ao facto de as pessoas preferirem pagar por animais de estimação "melhores", ou seja, animais saudáveis e treinados.
Tal como nós também outros autores vêem estas como uma das principais features que influencia a velocidade de adoção como podemos verificar no seguinte link: https://www.kaggle.com/artgor/exploration-of-data-step-by-step. Não só estes autores concordam com o facto de estas serem características cruciais na velocidade de adoção como também concordam com a evolução que as mesmas têm.

CONCLUSÕES

No final deste trabalho e após encontrados padrões de adoção e características que influenciam os mesmos foi-nos possível perceber quais são os métodos de modelação de dados mais indicados para o estudo da velocidade de adoção de animais de estimação bem como as features que mais influenciam a mesma. Assim sendo e, como dito introdutoriamente, esperamos que este trabalho possa servir como base à análise das taxas de velocidade de adoção em mais países que a Malásia.
Esperamos assim que, este trabalho possa ajudar próximos autores no seu estudo sobre adoções.

REFERÊNCIAS

Martins, J. P. , Programaçao em Python: Introdução a Programaçao Utilizando Múltiplos Paradigmas , IST Press., 2015

PetFinder.myAdoption Prediction. Disponível em: <https://www.kaggle.com/c/petfinder-adoption-prediction/overview>. Acesso em: 23 nov. 2019

How to Interpret Regression Analysis Results: P-values and Coefficients . Disponivel em :<https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients>. Acesso em: 10 dez. 2019

Logistic Regression In Python . Disponivel em: <https://towardsdatascience.com/logistic-regression-python-7c451928efee> . Acesso em: 10 dez 2019

What does RMSE really mean? . Disponivel em: <https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e> . Acesso em: 9 dez 2019

Python | Decision Tree Regression using sklearn. Disponível em: <https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/>. Acesso em : 10 dez 2019

Python Pandas Data Frame Basics.. Disponível em: <https://towardsdatascience.com/python-pandas-data-frame-basics-b5cfbcd8c039> . Acesso em: 10 dez 2019

Pandas – Render DataFrame as HTML Table . Disponivel em: <https://pythonexamples.org/pandas-render-dataframe-as-html-table/> . Acesso em 01 dez 2019