

# ANÁLISE DE DADOS SOBRE UM DATASET UTILIZANDO A FERRAMENTA ORANGE, PYTHON E SUAS BIBLIOTECAS

JONAS ROMBALDO DENARDI¹; LEANDRO CARLOS FERNANDES².

ICET – Instituto de Ciências Exatas e Tecnológicas

Ciência da Computação

<sup>1</sup>Graduando em Ciência da Computação - Universidade Paulista UNIP - Ribeirão Preto, jonas\_denardi@hotmail.com

<sup>2</sup>Doutor em Ciências Matemáticas e de Computação — USP, Professor Titular da Universidade Paulista nos campi de Araraquara e Ribeirão Preto, leandro.fernandes@docente.unip.br

## INTRODUÇÃO

Nunca se gerou tanta informação como nos dias atuais – e essa informação é originada a partir de dados. Dados produzidos por sistemas, celulares, sensores, câmeras, dispositivos de segurança, tudo isso em grande volume e velocidade. Nesse contexto, entra o papel da ciência de dados, trazendo ferramentas, métodos e tecnologias para analisar, visualizar e tomar decisões a partir dos dados. A ciência de dados é um processo, não um evento. É o processo de usar dados para entender o mundo, é a arte de descobrir os insights e tendências que estão escondidos atrás dessas informações.(CENTRALPRESS, 2018).

Análise de dados possui diferentes facetas e abordagens, incorporando diversas técnicas. Tem grande importância nas mais diversas áreas: ciências, estudos sociais, negócios, entretenimento, saúde, serviços financeiros, sistemas de recomendação, etc. por conta da diversidade de modelos possíveis.(WIKIPÉDIA, 2017).

Uma grande utilização atualmente foi a disponibilização do supercomputador "Watson" pela empresa IDM para auxiliar médicos no diagnóstico e na pesquisa da cura do câncer. (EXAME, 2015). Nesse mundo de análise de dados uma linguagem se destaca: Python. Mas por que os Cientistas de Dados preferem codificação em Python? Não só por causa de sua intensidade computacional, outros fatores também contribuem como sua grande comunidade, crescente número de bibliotecas de análise de dados, é fácil de aprender, escalabilidade, visualização / gráficos, etc.(CIÊNCIA E DADOS, 2018).

Um software de código aberto muito útil para análise de dados é o Orange. Contém vários algoritmos de mineração e aprendizado de máquina para pré-processamento de dados, classificação, modelagem, regressão, clustering e outras funções diversas. Seu ambiente de trabalho consiste em ferramentas para importar dados, arrastar e soltar widgets e links para conectar diferentes widgets para concluir o fluxo de trabalho.(ANALYTICS VIDHYA, 2017).

#### **OBJETIVO**

Utilizando Pyhon, Orange e conceitos de análise de dados analisar um dataset contendo resultado de exames de câncer de mama e montar classificadores (conceito de aprendizado de máquina) para predizer se um novo câncer é benigno ou maligno com base no que os classificadores aprenderam. Mostrar através de gráficos quais características influenciam mais nos resultados de benigno ou maligno.

#### **METODOLOGIA**

O trabalho foi dividido em 3 etapas: limpeza dos dados (utilizado python e pandas), execução dos classificadores e visualização dos resultados obtidos (utilizado o Orange). A etapa de limpeza evolve o entendimento dos dados e suas características a fim de saber o que cada característica representa dentro do dataset e assim podermos eliminar aquelas irrelevantes. Na figura 1 podemos ver o cabeçalho do dataset com as 5 primeiras linhas. Como a característica "Sample code number" serve apenas como um identificador único do exame não precisamos dela, podendo assim remove-la do dataset como mostrado na figura 2. Em seguida verificamos se existiam valores faltantes. A característica "Bare Nuclei" continha valores faltantes representados por "?", assim fizemos a substituição deles pela média da própria coluna como mostrado na figura 3. Ainda na etapa de limpeza poderíamos normalizar os dados dividindo o valor de cada coluna pelo seus maior valor porém o dataset já está normalizado de 0 a 10.

Figura 1 – Cabeçalho do dataset com as 5 primeiras linhas

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1	3	1	1	Benígno
1	1002945	5	4	4	5	7	10	3	2	1	Benígno
2	1015425	3	1	1	1	2	2	3	1	1	Benígno
3	1016277	6	8	8	1	3	4	3	7	1	Benígno
4	1017023	4	1	1	3	2	1	3	1	1	Benígno

Figura 2 – Remoção da coluna "Sample code number"

#Remove a coluna Sample code number pois é irrelevante na classificação data.drop('Sample code number',axis=1,inplace=True)

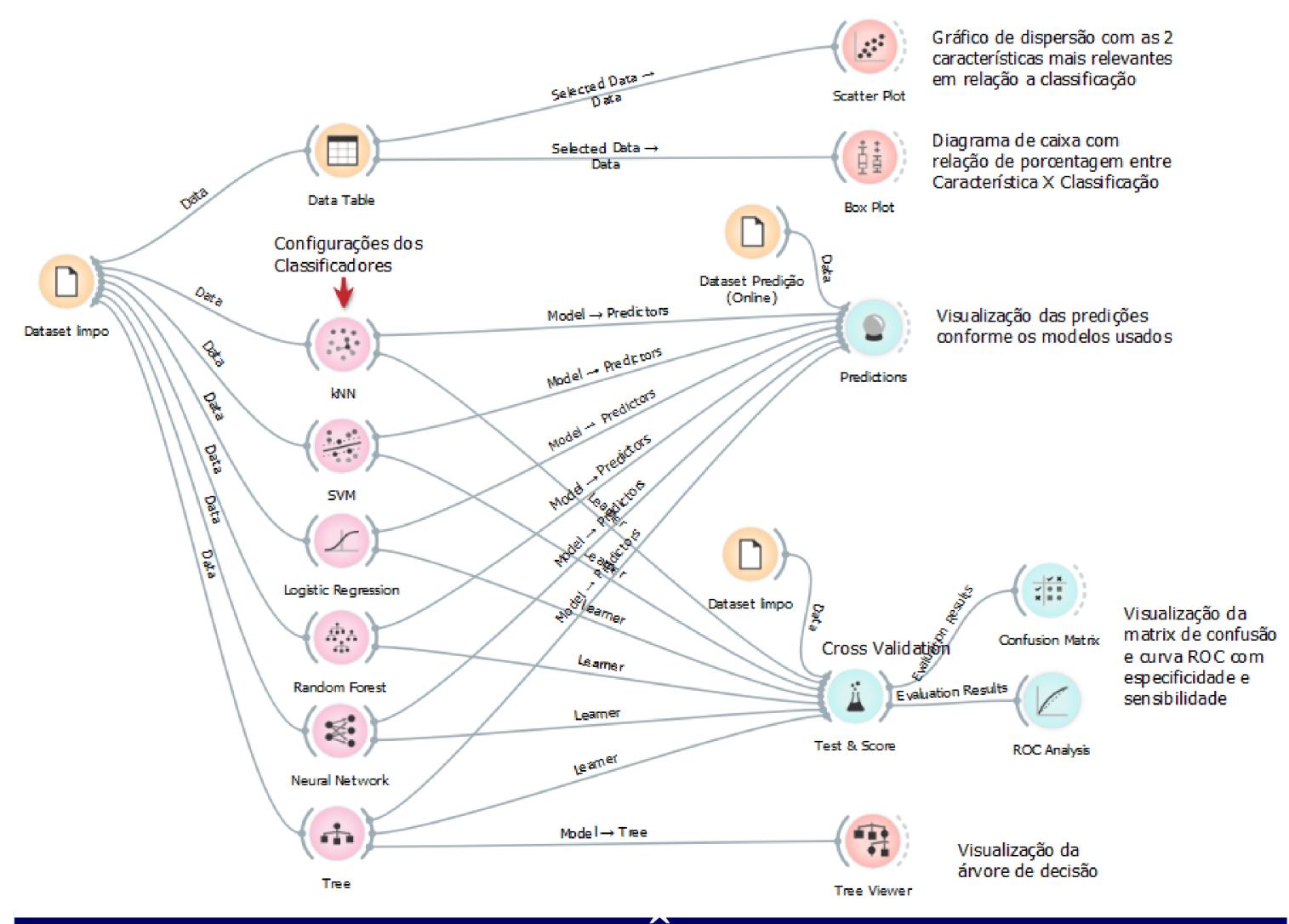
Figura 3 – Substituição de "?" pela média da coluna

#Coluna 'Bare Nuclei' possui 16 diagnósticos com valores faltantes que foram preenchidos com '?',as soluções mais simples são #remover as linhas correspondentes ou fazer a média dos valores e embutir no lugar de '?', Que é o que farei nesse caso data['Bare Nuclei'][data['Bare Nuclei'] == '?'] = int(round(data['Bare Nuclei'][data['Bare Nuclei'] != '?'].astype('int').mean(),0))

Para a etapa de execução dos classificadores utilizamos o Orange ligando os widgets necessários como mostrado na figura 4. Primeiro carregamos o dataset limpo depois ligamos ele aos classificadores que neste caso usamos 6 (KNN, SVM, Regressão Logística, Floresta Aleatória, Rede Neural e Árvore de decisão). Todos os classificadores foram ligados a um widget de Test & Score para avaliação dos classificadores e outro widget de predição, este último que serve para classificar uma nova amostra de exame.

Para a etapa de visualização dos resultados obtidos ligamos o Test & Score a outros 2 widgets: Confusion Matrix que mostra as taxas de acerto dos classificadores e ROC Analysis que mostra a curva ROC com a especificidade e sensibilidade. O dataset limpo ligamos a um widget de tabela e este ligamos a outros 2: Scatter Plot que mostra um gráfico de dispersão mostrando 2 características em relação a uma característica alvo e Box Plot que mostra um diagrama de caixa em relação a porcentagem entre determinada Característica X Classificação.

Figura 4 – Orange com os widgets conectados



#### REFERENCIAS

UCI. Breast Cancer Wisconsin (Original) Data Set. Disponível em:

<a href="https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%280riginal%29dados">https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%280riginal%29dados</a> >. Acesso em 17 de novembro de 2017.

Vasconcellos, Paulo. 28 comandos úteis de Pandas que talvez você não conheça. Disponível em:<a href="https://paulovasconcellos.com.br/28-comandos-%C3%BAteis-de-pandas-">https://paulovasconcellos.com.br/28-comandos-%C3%BAteis-de-pandasque-talvez-voc%C3%AA-n%C3%A3o-conhe%C3%A7a-6ab64beefa93>. Acesso em: 20 de maio de 2018.

Orange. Orange, Data Mining Fruitful and Fun. Disponível em: <a href="https://orange.biolab.si//>">https://orange.biolab.si//>">. Acesso em 17 de julho de 2018.

Figura 5 – Widget Test & Score

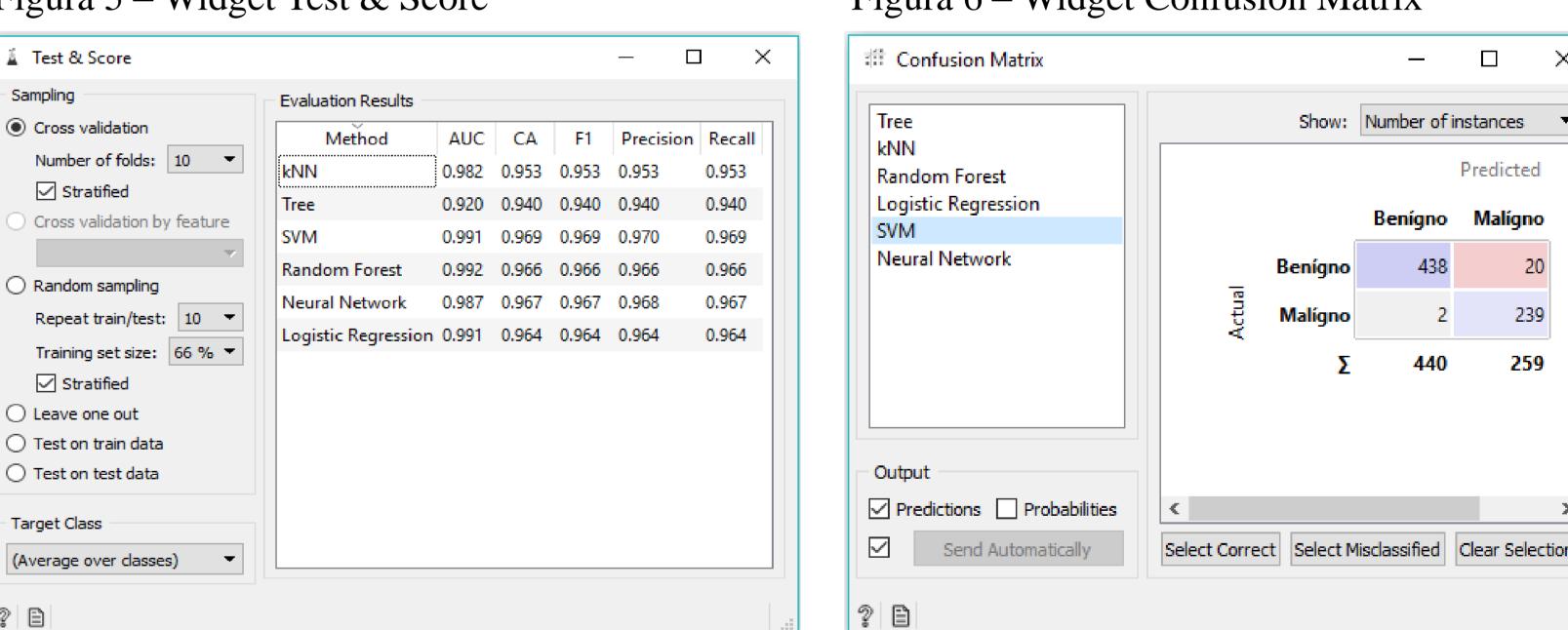


Figura 6 – Widget Confusion Matrix

Predicted

#### RESULTADO

Através do gráfico de dispersão concluímos que as 2 características que influenciam no resultado são: Bare Nuclei e Uniformity of Cell Size. Quanto menor o valor dessas duas características maior as chances do resultado ser benigno e quanto maior o valor delas maior as chances de ser maligno. No Test & Score tivemos melhores resultados em relação a acurácia (CA) nos classificadores SVM e Neural Network refletidos na matriz de confusão onde as duas foram as que obtiveram menor taxa de erros.

### CONCLUSÃO

O desenvolvimento do presente trabalho possibilitou a exploração de um dataset set aplicando conceitos e ferramentas de Análise de Dados. Ao final foi possível obter resultados satisfatórios como alta acurácia dos classificadores e entendimento de quais características do câncer influenciavam mais na classificação podendo assim explora-las mais a fundo a fim de entender quais fatores externos influenciam para o atual estado desta característica. O melhor resultado, como visto na matriz de confusão, é do classificador SVM que detecta muito bem os malignos o que teoricamente é interessante pois entre ter poucos falsos positivos ou poucos falsos negativos é melhor, no caso de câncer, ter falsos negativos. Pois diagnosticando como maligno mas o câncer sendo benigno a pessoa enfrenta o processo de tratamento porém em exames futuros irá constatar que não é maligno. Já o contrário, câncer ser maligno e diagnosticar benigno corre-se o grande risco da evolução do câncer para um estado grave até a constatação em exames futuros. O classificador pode ajudar médicos na decisão do diagnóstico mas com o ressalvo de que o classificador não substitui o médico que ainda deve dar a palavra final.