BLUEEDTECH
ANA CRISTINA CHAVES
RODRIGO ANDRENELI
WAGNER BRAGA MOURA

MELHORIAS NOS MÉTODOS DE ANÁLISE DE DADOS COM *CROSS VALIDATION*

Trabalho para conclusão do Módulo 01, do curso de Cientista de dados da BlueEdTech, Turma C014. Orientado pelo Professor Rafael e Professor Flávio

2022

1. INTRODUÇÃO

O desenvolvimento humano traz consigo o aumento no volume de informações passíveis de serem acessadas e consultadas com as mais variadas finalidades. Com o advento da internet e sua respectiva evolução, cada vez mais se tornou importante esse tipo de acesso, a ponto de a humanidade precisar criar regras e legislações para que estes volumes de informações não sejam utilizados de maneira prejudicial às pessoas que as fornecem.

Neste documento não vamos abordar a questão legal do uso das informações, mas sim a importância de utilizá-las de maneira que consigamos classificar novos dados, após feito uma análise, a fim de encontrarmos padrões e seus desvios, correlações, comportamentos ou conexões ao tema a ser estudado.

A análise de dados consiste na transformação de dados em conhecimentos e *insights* relevantes. Ou seja, agregar novas informações a fim de entender o que os dados iniciais nos dizem[5].

Existem diversas formas de realizar a análise de dados, dentre elas vamos abordar a análise realizada a partir de uma estrutura de tabelas, ou conhecida como análise tabulada. Dentro deste tipo de análise existem algumas categorias, descritas abaixo[5].

Análise Descritiva: Consiste na descrição das principais características de um conjunto de dados, listando e resumindo valores mais comumente de apenas uma variável. Neste caso aparecem análises de média, mediana, moda, mínimo, máximo, porcentagem e frequência além de seus desdobramentos, como por exemplo o desvio padrão[5].

Análise Exploratória: Nesta etapa, já com os dados da análise descritiva obtidos, faz-se as correlações entre variáveis, usando técnicas de regressões e análise de variância[5].

Análise Preditiva: Baseada no acúmulo de informações prévias, constrói-se modelos a fim de realizar previsões de eventos futuros. Por exemplo, após analisar os hábitos e rotinas de uma pessoa, podemos prever com um certo grau de assertividade, onde ela estará na próxima segunda-feira às 9 horas[5].

Análise Prescritiva: Este tipo de análise baseia-se no acúmulo de análises já realizadas e tem como objetivo gerar tomadas de decisões ou sugestões de forma

automática ou semiautomática. Podemos dar o exemplo de sistemas que liberam crédito para usuários sob medida, de acordo com seu histórico de pagamentos[5].

Dentro das ferramentas de análise preditiva, existem os processos para desenvolvimento de projetos de *Machine Learning* (aprendizado de máquina), que consistem em separar os dados do *Dataset (*banco de dados) através de métodos de coleta pré-definidos que nos tragam uma amostra relevante deste banco de dados, e submeter essa amostra à modelos de algoritmos que retornarão um comportamento padrão que permita prever comportamento de dados futuros.

A partir deste momento, vamos introduzir o conceito do *Cross Validation (*validação cruzada), que é uma técnica estatística aplicada ao *Machine Learning* e conhecida pela sua eficiência em testes de desempenho, é amplamente empregada em problemas onde o objetivo da modelagem é a predição.

Consiste no particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, o uso de alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento), sendo os subconjuntos restantes (dados de teste) empregados na validação do modelo.[2]

A seguir vamos mostrar quais as melhorias que o *Cross Validation* em relação a proporção fixa, traz para o modelo de algoritmo escolhido, utilizando a base de dados Iris e o software Orange Data Mining.

1.1. BASE DE DADOS IRIS

O conjunto de dados flor Iris ou conjunto de dados Iris de Fisher é um conjunto de dados multivariados introduzido pelo estatístico e biólogo britânico Ronald Fisher em seu artigo de 1936 “O *uso* de *múltiplas* medições em *problemas taxonômicos*, como um *exemplo* de *análise discriminante linear*”.[3]

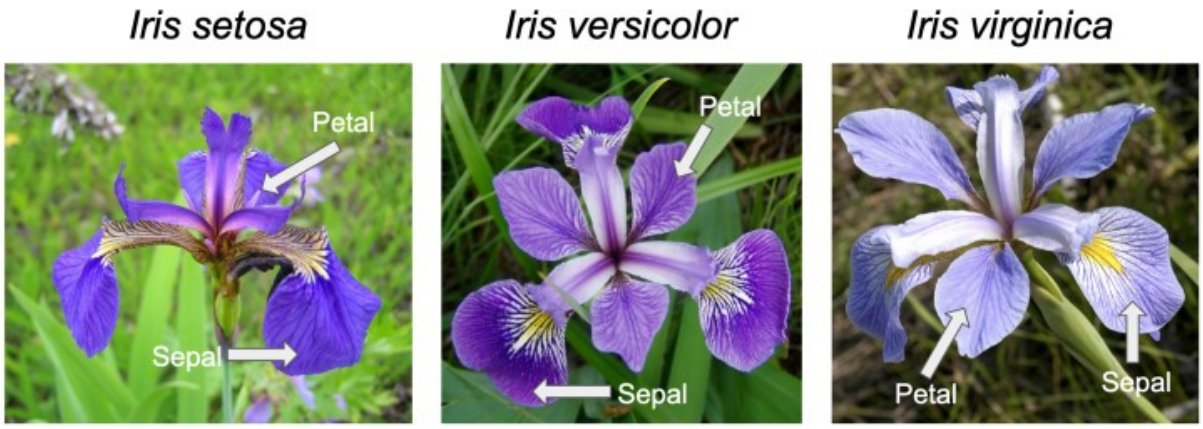
Este conjunto de dados foi criado pelo botânico Edgar Shannon Anderson (1897 - 1969) onde mediu-se o comprimento e largura de pétalas e de sépalas de três espécies de flores Iris (Setosa, Versicolor e Virgínica) com cinquenta dados de cada espécie, totalizando cento e cinquenta flores submetidas às medições.[3]

Na década de 1930 um estatístico chamado Ronald Aylmer Fisher (1890 - 1962) utilizou essas medidas para testar seus estudos em análise multivariável. De forma simplificada, Fisher investigava se a partir das medidas das 150 flores, seria possível classificar uma nova flor a partir de suas dimensões.[4]

Essas medidas são estudadas até hoje para o desenvolvimento de métodos estatísticos e de estudos computacionais.

Com base na combinação dessas quatro características (comprimento de sépala, largura de sépala, comprimento de pétala e largura de pétala), Fisher desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras.[3]

Figura 1 – Espécies da Flor Iris contidas na base de dados



Fonte: https://blueedtech.gitbook.io/dados-modulo-1-dsf1/1a-semana/aula_01

2. *CROSS VALIDATION* (VALIDAÇÃO CRUZADA)

Uma das etapas mais importantes em um projeto de *Machine Learning* é a etapa de validação dos resultados. Muitas das técnicas mais poderosas de aprendizado apresentam uma grande quantidade de parâmetros e quanto menos restrições colocarmos no nosso modelo maior a probabilidade de encontrarmos um superajustamento, ou como é mais conhecido, *Overfitting*.[1]

O *Overfitting* ocorre quando o método de aprendizado não consegue generalizar os resultados para dados que não foram utilizados no processo de treino.

Para evitar este tipo de problema, podemos aplicar processos de validação. Dentro das possibilidades existentes destes processos, abordaremos o *Cross Validation* (validação cruzada) que vem sendo cada vez mais utilizado e consiste em dividir os seus dados em base de treino e base de teste.

A validação cruzada estima o erro do método de aprendizado em observações não utilizadas no treino, ou seja, estima como o modelo construído irá se comportar aos novos dados (base de teste).

O *Cross Validation* consiste em dividir a base de dados em várias partes, sendo algumas destinadas ao treinamento do modelo e outras para realizar o teste.

Para cada interação, o método estuda as estimativas utilizando as partes destinadas ao treino (train) e verifica o erro médio com a parte destinada ao teste (test), conforme exemplificado pela figura abaixo.

Figura 2 – Interações promovidas pelo *Cross Validation*

Fonte: https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

A estimativa do erro de predição de *Cross Validation* é dada pela média dos erros médios das partes testadas.

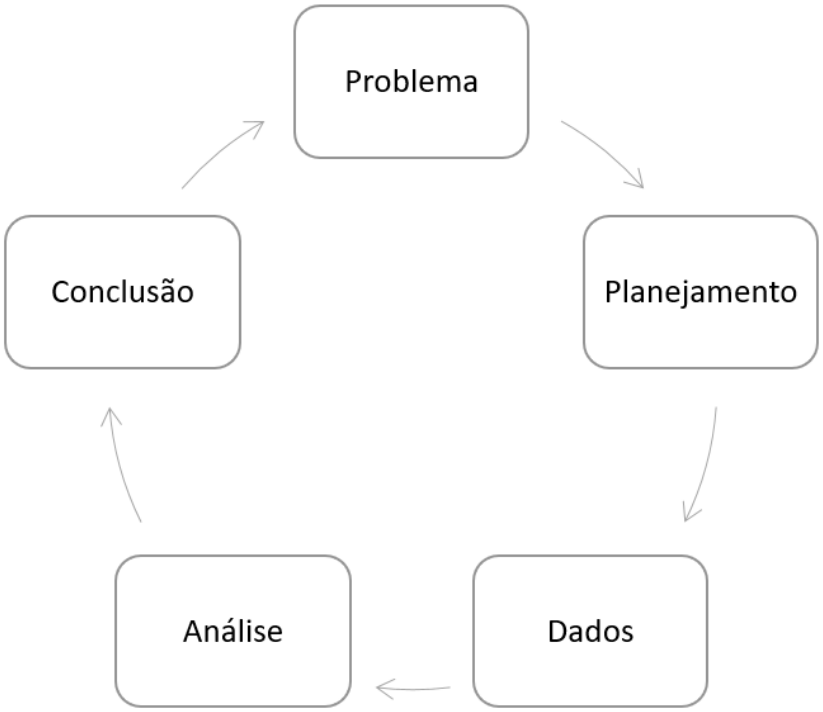
3. ANÁLISE ESTATÍSTICA EXPLORATÓRIA E CONFIRMATÓRIA

Em estatística, a análise exploratória de dados consiste em abordar os conjuntos de dados de maneira a se resumir suas características principais, frequentemente com métodos visuais.

Há várias fases durante o trabalho, a fase exploratória nos ajuda a obter percepções do conjunto de dados, enquanto a fase confirmatória realiza a confirmação dessas percepções. Sem essa fase de exploração é provável se afastar dos dados e se perder em tópicos sem sentido, sem alcançar o objetivo.

Uma proposta para organizar a análise estatística é através do ciclo investigativo PPDAC proposto por C. J. Wild e M. Pfannkuch (1999) e que diante de um contexto, investiga um problema, planeja a análise, coleta ou considera os dados, faz a análise e verifica as conclusões (Figura 3).

Figura 3 – Ciclo Investigativo



Fonte: https://blueedtech.gitbook.io/dados-modulo-1-dsf1/1a-semana/aula_01

Na fase do problema, vamos conhecer o contexto dos dados e a definição do problema. No planejamento, define-se as ações para a investigação. Na parte dos dados, descreve-se o processo de coleta dos dados. Seguimos então para a análise, realizando o tratamento e análise dos dados. Fechando então o ciclo com a conclusão tendo posicionamento crítico, reflexivo, com a comunicação dos dados podendo gerar novas ideias e novos questionamentos [6].

A partir do ciclo PPDAC, avançaremos a nossa análise.

3.1. PROBLEMA

Observamos através dos dados, e através de muitas consultas de termos, conceitos, exemplos de testes e manipulação do software Orange, nos levando às dúvidas iniciais:

- Para que serve o *Cross Validation*?
- Ele altera as métricas no *Test and Score?*
- Como analisar essas diferenças?

3.2. PLANEJAMENTO

Avaliar os dados, verificando como utilizar o *Cross Validation* no *software Orange* de modo que extraia valores de métricas confiáveis. Após o entendimento do uso da validação cruzada, compara-se os valores de resposta, podendo assim obter a informação de melhoria ao processo que se espera utilizando esta ferramenta, melhorando as métricas e evitando o *Overfitting.*

3.3. DADOS

Os dados utilizados para a análise será os da Base Iris, descrita previamente neste trabalho. Para melhor entendimento do conjunto, descreveremos abaixo o dicionário de dados da base íris.

Dicionário de Dados (Íris):

- 150 observações/dados (N = 150)

Variáveis:

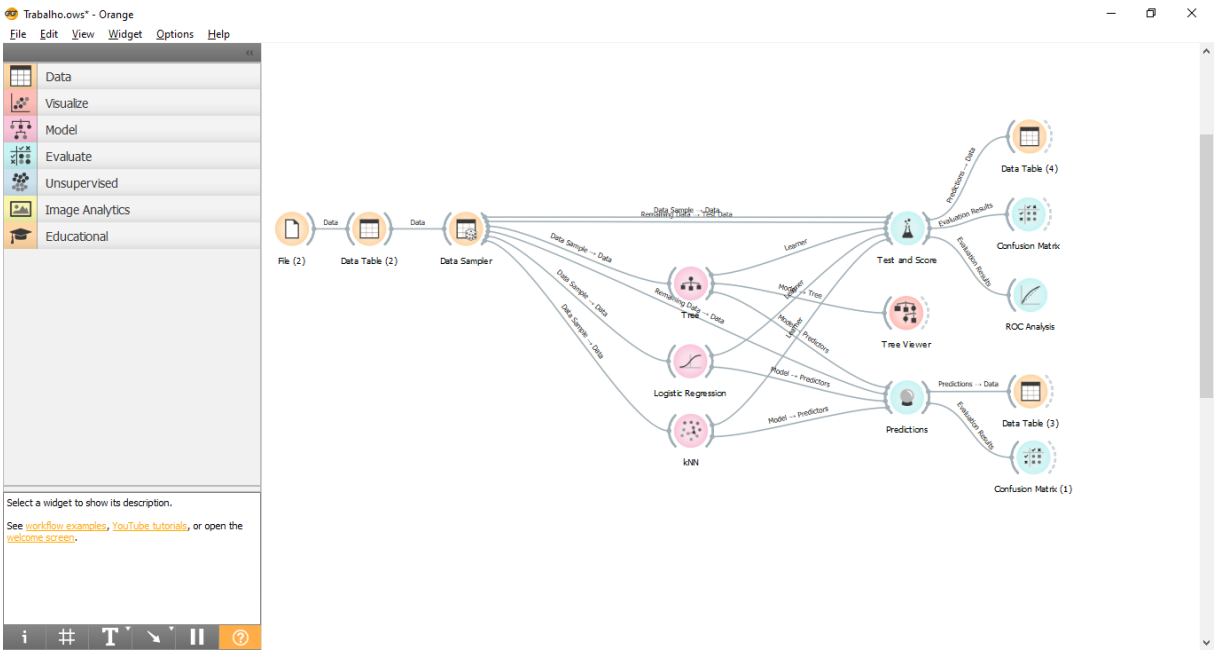
- Íris (Qualitativa Nominal): Setosa, Versicolor, Virgínica
- Comprimento da Sépala (Quantitativa Contínua) (cm): 4,3; ...; 7,9
- Largura da Sépala (Quantitativa Contínua) (cm): 3,0; …; 3,8
- Comprimento da Pétala (Quantitativa Contínua) (cm): 1,1; ...; 6,4
- Largura da Pétala (Quantitativa Contínua) (cm): 0,1; ...; 2,0

Neste dicionário percebe-se então que a variável “Espécie” é do tipo Qualitativa Nominal, pois apenas nomeia os três tipos existentes na base de dados. As demais variáveis, por se tratarem de medidas as quais foi necessário realizar suas medições uma a uma, são classificadas de Quantitativas Contínuas e têm seus valores mínimos e máximos descritos logo após a classificação expressos em centímetros.

3.4. ANÁLISE

Fizemos vários testes utilizando alguns algoritmos para melhor visualização, conforme a figura abaixo:

Figura 4 – Modelo de teste da base Íris no software Orange

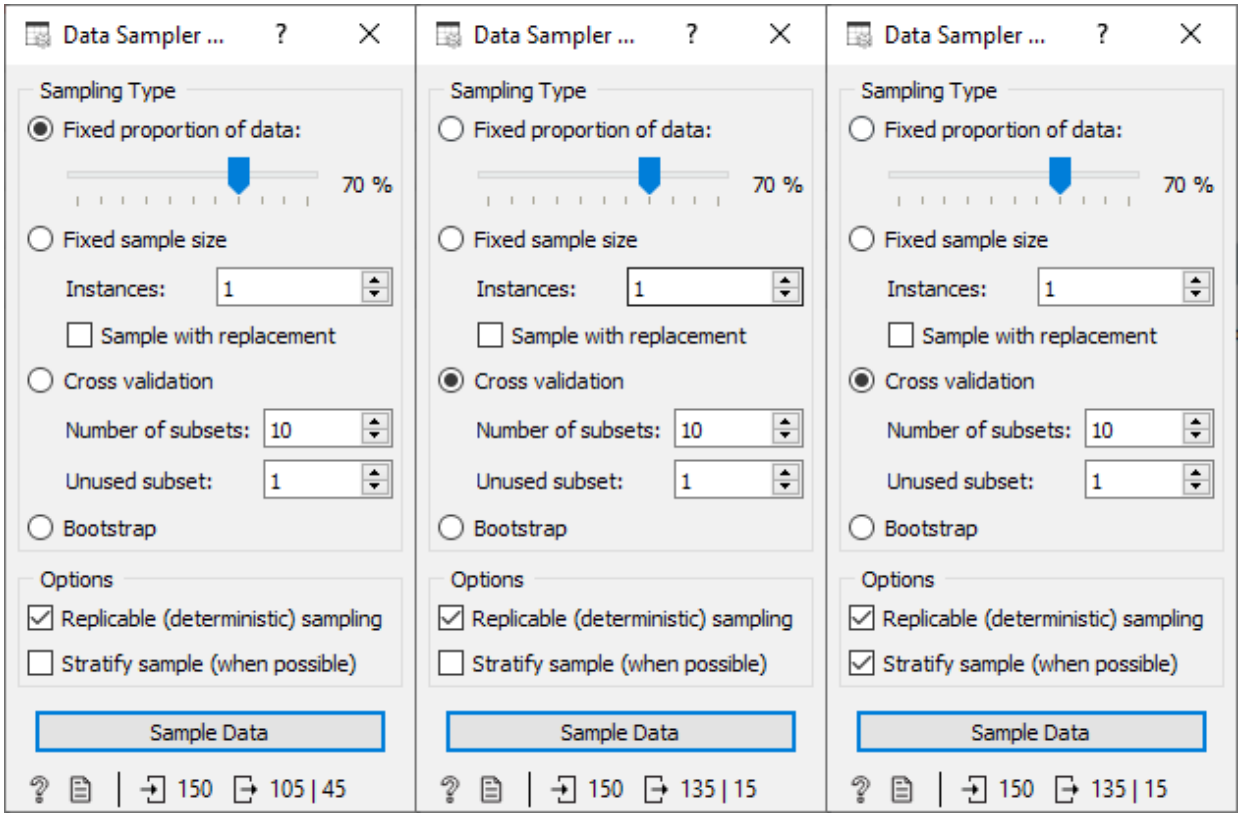


Fonte: Software Orange Data Mining

Para essa análise, utilizaremos os algoritmos, *tree, logistic regression* e *KNN*, para visualizamos o funcionamento do *cross validation*.

Neste modelo os *widgets Data Sampler* e *Test and Score* foram configurados para que pudéssemos observar a real melhoria que o *cross validation* traria para o modelo, conforme as Figuras 5 e 6 respectivamente retornando as métricas de validação.

Figura 5 – Configurações do *Data Sampler* nos três casos estudados



Fonte: Software Orange Data Mining

Utilizamos três configurações para comparamos as alterações que o *cross validation* apresenta.

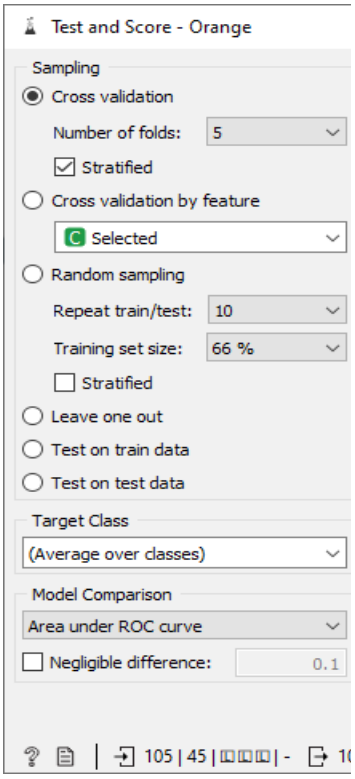
Uma base de dados fixo, cujo 70% (setenta por cento) entra para treinamento e o restante dos 30%(trinta por cento) ficariam para teste (modelo 1).

A segunda configuração, usamos o *cross validation*, onde ele separa a base de dados em 10 (dez) parte e exclui uma, sem estratificar os dados (modelo 2).

E a última, utilizamos o *cross validation*, dividindo a base em 10 partes, excluindo uma com os dados estratificados (modelo 3).

E a configuração do *test and score*, local onde serão apresentados as métricas testadas dos modelos, está no *sampling* do *cross validation, stratified* com 5 *folds (dobra)* em todas as bases.

Figura 6 – Configuração do *Test and Score*



Fonte: Software Orange Data Mining

Todos os modelos apresentaram uma boa precisão conforme a Figura 7.

Figura 7 – Comparação das métricas dos resultados

Test and Score / Data Sampler	Cross Validation (5 folds) - Stratified					
Fixed Proportion of Data	Model	AUC	CA	F1	Precision	Recall
	kNN	0,996	0,943	0,943	0,947	0,943
	Tree	0,970	0,943	0,943	0,944	0,943
	Logistic Regression	0,998	0,962	0,962	0,962	0,962
Test and Score / Data Sampler	Cross Validation (5 folds) - Stratified					
Cross Validation (10/1)	Model	AUC	CA	F1	Precision	Recall
	kNN	0,996	0,963	0,963	0,963	0,963
	Tree	0,952	0,948	0,948	0,948	0,948
	Logistic Regression	0,997	0,970	0,970	0,971	0,970
Test and Score / Data Sampler	Cross Validation (5 folds) - Stratified					
Cross Validation (10/1) - Stratify	Model	AUC	CA	F1	Precision	Recall
	kNN	0,995	0,963	0,963	0,963	0,963
	Tree	0,983	0,948	0,948	0,948	0,948
	Logistic Regression	0,997	0,956	0,956	0,956	0,956

Fonte: Software Orange Data Mining

Levando em conta apenas a precisão dos métodos estudados e conforme a figura acima, podemos notar que a primeira configuração utilizada já nos retorna bons resultados utilizando o *Cross Validation* apenas no *widget Test and Score* (modelo 1).

Quando comparamos os primeiros resultados com a segunda configuração utilizada, onde também foi ativado o *Cross Validation* no *widget Data Sampler*, pode-se perceber que todos os modelos foram reajustados, retornando precisões ainda melhores (modelo 2).

Porém, quando seguimos para a terceira configuração, onde ativamos a opção *Stratify* no *widget Data Sampler*, o modelo *Logistic Regression* retornou precisões com decréscimo, e, não houve alterações nos demais modelos (modelo 3).

A princípio a métrica não poderia chegar a 1.0 (overfitting), e em nenhum modelo mencionado isso ocorreu, utilizamos o *cross validation*, e ele melhorou sutilmente as métricas, gerando maior confiabilidade nos modelos propostos nesta análise.

3.5. CONCLUSÃO

O Cross validation é amplamente usado para melhorar as métricas dos modelos, evitando erros e overfitting, pois ele divide a base de dados em várias partes (folds) e testa mais dados de forma cruzada, generalizando mais o modelo, gerando uma métrica para cada teste, e tirando uma média das métricas o que o torna mais confiável.

Ele altera as métricas justamente para evitar o superajustamento, em alguns casos ele diminui as métricas, por generalizar mais a amostra. Neste banco de dados, pudemos observar que houve melhora das métricas, entretanto, não foi uma alteração muito significativa. Nisto não conseguimos identificar o real motivo desde fato, mas conseguimos supor que devido ao pequeno tamanho do banco de dados ou sua simetria (50 dados de cada espécie) acabamos por não necessitar de mais complexidade, não entramos no mérito de algoritmos utilizados e quais seriam melhores para o banco de dados em questão.

Foi observado a comparação entre o resultado obtido no *test and score* com a configuração *Fixed*, em contra partida com a configuração do *cross validation*, e as métricas tiveram uma alteração sútil.

Nossa maior dificuldade foi de entender a configuração e utilização do *cross validation*, dentro do *orange*, tanto no *data sampler* quanto no *test and score*. Por fim optamos por expor as várias configurações possíveis

4. REFERÊNCIA BIBLIOGRÁFICA

[1] PESSANHA, Cínthia. **Por que Cross Validation?**. Disponivel em: <https://medium.com/cinthiabpessanha/por-que-cross-validation-b4f57007834a>. (Acesso em 16/03/2022).

[2] **Validação cruzada**, Disponível em:<https://pt.wikipedia.org/wiki/Valida%C3%A7%C3%A3o_cruzada> (Acesso em 16/03/2022)

[3] BLUEEDTECH. **Orange e Análise Estatística.** Disponível em: <https://blueedtech.gitbook.io/dados-modulo-1-dsf1/1a-semana/aula_01>. Acesso em (16/03/2022)

[4] **Conjunto de dados flor Iris**. Disponível em: <https://pt.wikipedia.org/wiki/Conjunto_de_dados_flor_Iris> (acesso em 16/03/2022).

[5] **MAS O QUE SIGNIFICA ISSO?': INTRODUÇÃO À ANÁLISE DE DADOS.** Disponivel em**: <**https://escoladedados.org/tutoriais/mas-o-que-significa-isso-introducao-a-analise-de-dados/>. (Acesso em 19/03/2022)

[6] SANTANA, Eurivalda Ribeiro dos Santos; CAZORLA, Irene Maurício. **O Ciclo Investigativo no ensino de conceitos estatísticos**. Revemop, Ouro Preto, v. 2, p. 1-22, 14 out. 2020. Disponível em: <https://periodicos.ufop.br/revemop/article/view/4251>. Acesso em: 20 mar. 2022.