

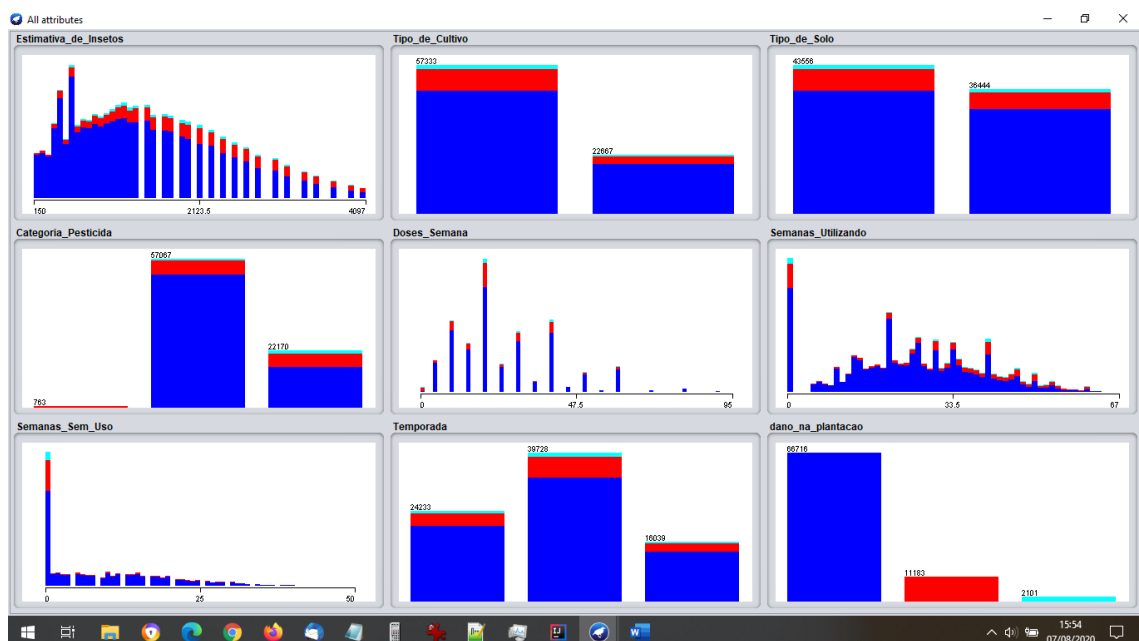
Resposta ao desafio (Processo seletivo para Estágio em Data Science)

Termino 07/8/2020

- 1) Primeiro foi retirado dos dados (Safra_2018-2019.csv) as duas primeiras colunas: **numeração da linha** e o **Identificador_Agricultor**. Ficando os demais dados.

	Identificado	Agricultor	Estimativa_de_In.
1	0, SCROP00001,	188,1,0,1,0,0,0,0,1,0	
2	1, SCROP00002,	209,1,0,1,0,0,0,0,2,1	
3	2, SCROP00003,	257,1,0,1,0,0,0,0,2,1	
4	3, SCROP00004,	257,1,1,1,0,0,0,0,2,1	
5	4, SCROP00005,	342,1,0,1,0,0,0,0,2,1	
6	5, SCROP00006,	348,0,1,1,0,,0,2,1	
7	6, SCROP00007,	348,0,1,1,0,,0,2,1	
8	7, SCROP00008,	377,1,0,1,0,0,0,0,1,2	
9	8, SCROP00009,	731,0,0,1,0,0,0,0,2,0	
10	9, SCROP00010,	1132,1,0,1,0,0,0,0,1,2	
11	10, SCROP00011,	1212,1,0,1,0,,0,3,0	
12	11, SCROP00012,	1575,0,0,1,0,0,0,0,1,1	
13	12, SCROP00013,	1575,0,1,1,0,0,0,0,2,1	
14	13, SCROP00014,	1575,1,1,1,0,0,0,0,2,1	
15	14, SCROP00015,	1575,1,1,1,0,0,0,0,2,2	
16	15, SCROP00016,	1785,1,1,1,0,0,0,0,2,1	
17	16, SCROP00017,	2138,0,1,1,0,0,0,0,1,1	
18	17, SCROP00018,	2401,0,1,1,0,,0,1,1	
19	18, SCROP00019,	2401,1,1,1,0,0,0,0,2,1	
20	19, SCROP00020,	2401,1,1,1,0,0,0,0,2,1	
21	20, SCROP00021,	2999,0,1,1,0,0,0,0,3,1	
22	21, SCROP00022,	3516,1,0,1,0,0,0,0,2,0	
23	22, SCROP00023,	3895,1,1,1,0,0,0,0,1,1	
24	23, SCROP00024,	4096,1,1,1,0,0,0,0,2,1	

A imagem abaixo são os dados carregados no Weka depois de retirar as duas colunas. Para isso foi preciso fazer um arquivo (Safra_2018_2019.arff)











A imagem a seguir mostra o arquivo (Safra_2018_2019.arff)

```

1  @relation Safra_2018_2019
2
3  @attribute Estimativa_de_Insetos real
4  @attribute Tipo_de_Cultivo {0,1}
5  @attribute Tipo_de_Solo {0,1}
6  @attribute Categoria_Pesticida {1,2,3}
7  @attribute Doses_Semana real
8  @attribute Semanas_Utilizando real
9  @attribute Semanas_Sem_Uso real
10 @attribute Temporada {1,2,3}
11 @attribute dano_na_plantacao {0,1,2}
12
13 @data
14 188,1,0,1,0,0,0,1,0
15 209,1,0,1,0,0,0,2,1
16 257,1,0,1,0,0,0,2,1
17 257,1,1,1,0,0,0,2,1
18 342,1,0,1,0,0,0,2,1
19 448,0,1,1,0,0,0,2,1
20 448,0,1,1,0,0,0,2,1
21 577,1,0,1,0,0,0,1,2
22 731,0,0,1,0,0,0,2,0
23 1132,1,0,1,0,0,0,1,2
24 1212,1,0,1,0,0,0,3,0
25 1575,0,0,1,0,0,0,1,1

```

- 2) A linguagem usada para trabalhar foi o Java, devido a algumas dificuldades em se trabalhar com a tabelas dentro do java preferiu dividir cada coluna em um arquivo txt diferente como demonstrado na figura a seguir.

Nome	Data de modificação	Tipo	Tamanho
 Categoria_Pesticida.txt	06/08/2020 14:53	Documento de Te...	26 KB
 Doses_Semana.txt	06/08/2020 14:53	Documento de Te...	34 KB
 Estimativa_de_Insetos.txt	06/08/2020 14:50	Documento de Te...	49 KB
 Semanas_Sem_Uso.txt	06/08/2020 14:56	Documento de Te...	30 KB
 Semanas_Utilizando.txt	06/08/2020 14:55	Documento de Te...	34 KB
 Temporada.txt	06/08/2020 14:56	Documento de Te...	26 KB
 Tipo_de_Cultivo.txt	06/08/2020 14:51	Documento de Te...	26 KB
 Tipo_de_Solo.txt	06/08/2020 14:52	Documento de Te...	26 KB

- 3) Para se trabalhar com o Software Weka e seus modelos, foi necessário verificar qual deles seria melhor avaliado para o problema atual, então fez uso do programa desenvolvido para esse fim chamado de (EscolheModelo.class) de modo a escolher qual teria mais relevância para o problema.

IBk	J48	NaivesBayes	RandomForest	RandomTree
76.3937 %	84.2362 %	82.7525 %	82.1338 %	75.8012 %

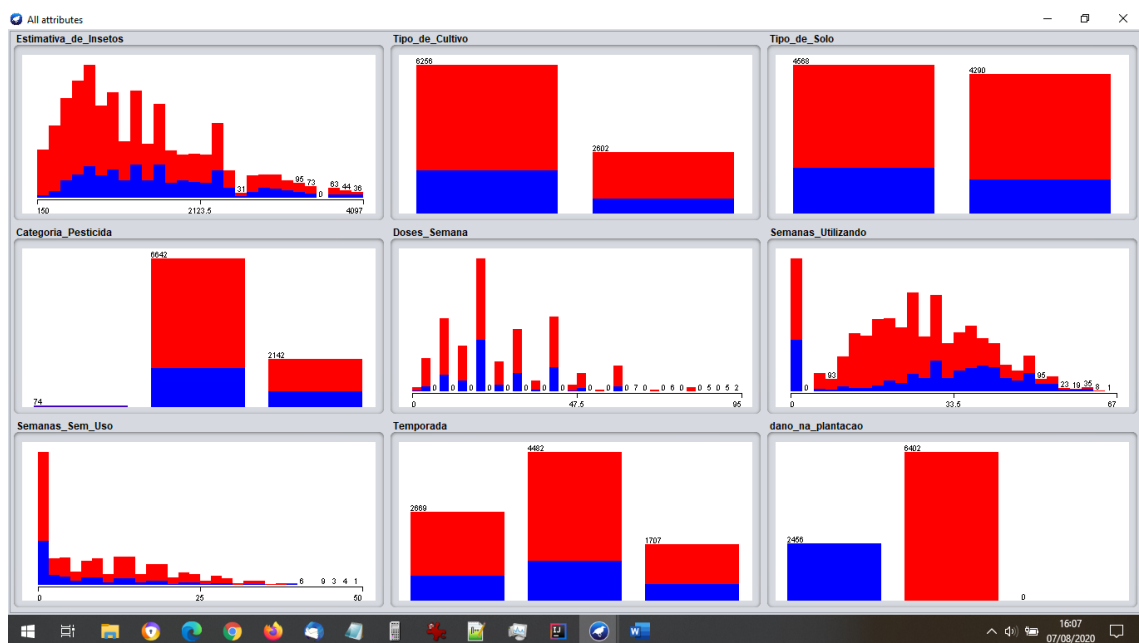
Para esse desafio usou-se o IBk com o valor de 76.3937 % das instancias corretamente classificados, embora outros tem dado valores mais altos não os usou, o NaivesBayes apresentou erro na hora de predizer os dados. Por outro lado, o RandomForest seu modelo ficou muito grande quase 200mb como pode ser visto na imagem abaixo.

Nome	Data de modificação	Tipo	Tamanho
IBk.model	05/08/2020 20:44	Arquivo MODEL	7.895 KB
J48.model	05/08/2020 20:37	Arquivo MODEL	86 KB
MLP.model	05/08/2020 19:58	Arquivo MODEL	15 KB
NaiveBayes.model	07/08/2020 15:23	Arquivo MODEL	6 KB
RdnForest.model	05/08/2020 20:20	Arquivo MODEL	196.359 KB
RdnTree.model	05/08/2020 20:38	Arquivo MODEL	3.064 KB

4) Com os modelos Feitos foi a vez de desenvolver (safra20.class)

Tal programa faz com que as colunas sejam salvas em vetores indo de 0 a 8858 então desse modo preenche os valores: Estimativa_de_Insetos, Tipo_de_Cultivo, Tipo_de_Solo, Categoria_Pesticida, Doses_Semana, Semanas_Utilizando, Semanas_Sem_Uso e Temporada com os dados que vem do arquivo (Safra_2020.csv) retirado as duas primeiras colunas. E o modelo construído no treino do (Safra_2018-2019.csv) calcula o valor do dano_na_plantacao e salva tudo em um arquivo chamado (weka_file_safra_2020.arff)

Quando se abre o arquivo (weka_file_safra_2020.arff) com os dados salvos, aparece como demonstrado na tabela abaixo.



Onde o dano_na_plantacao traz a informação que:

- tem 2556 ocorrências com valor 0=Sem Danos;
- tem 6402 ocorrências com valor 1=Danos causados por outros motivos
- tem 0 ocorrências com valor 2=Danos gerados pelos pesticidas