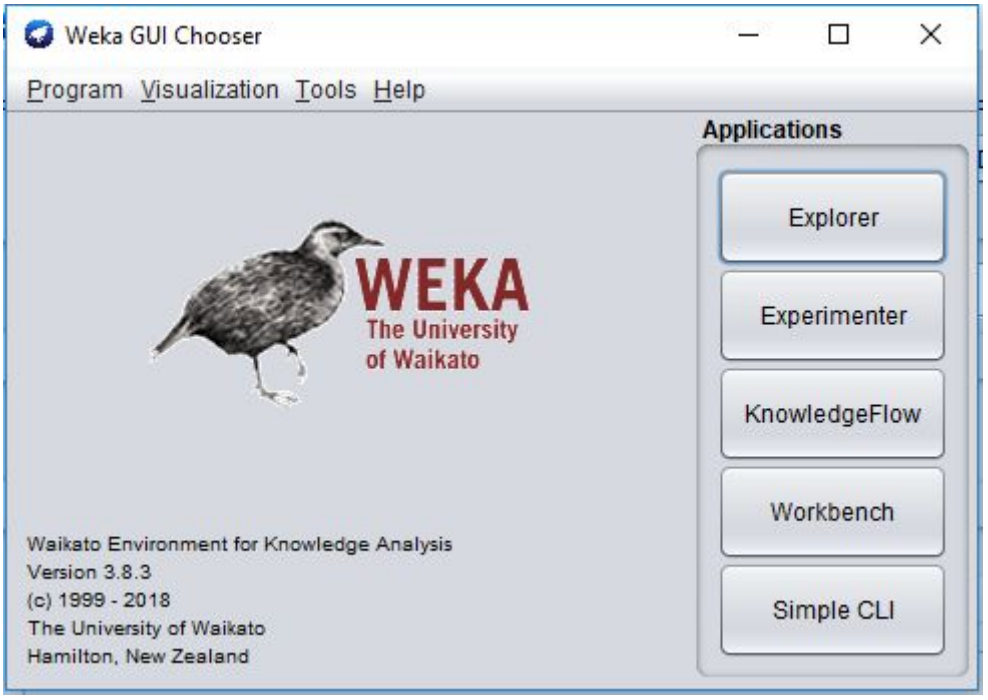
ACH2016 Inteligência Artificial
Exercício Programa 2 (EP2)
Classificação

Alunos:
Denise Keiko Ferreira Adati - 10430962
Fernando Karchiloff Gouveia de Amorim - 10387644
Gabriela Brindo Domingues - 9788030

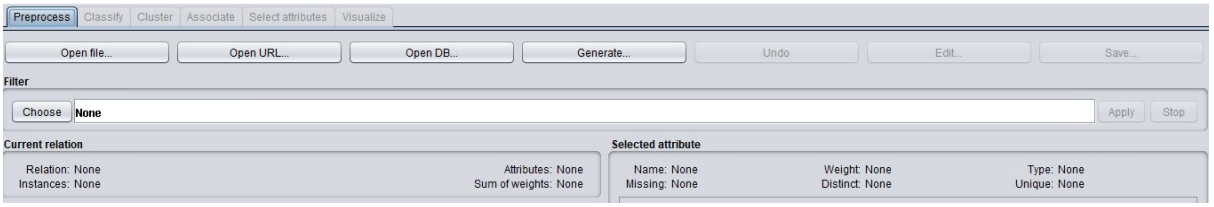
São Paulo
2019

Acessando o *Weka*



Para acessar o programa que faz as análises no *Weka*, acesse o menu de "*Explorer*".

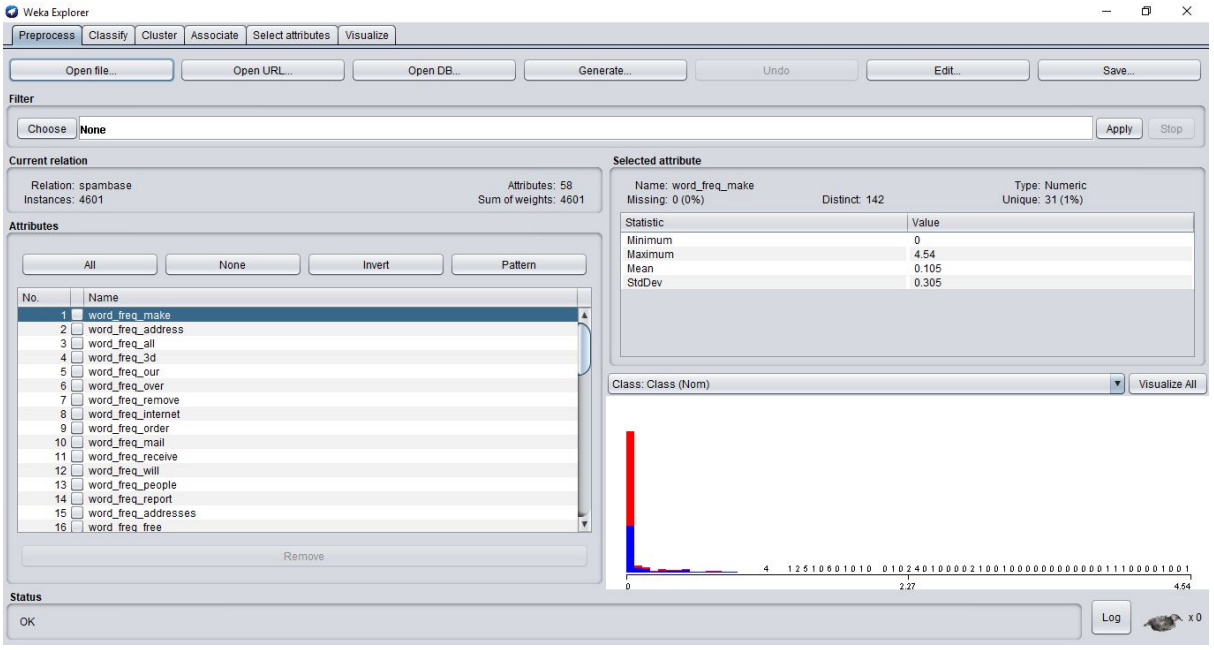
Selecionando o *dataset*



Para selecionar os *datasets* que devem ser utilizados para o tratamento de dados e execução dos algoritmos, deve ser acessado pela aba de "*Preprocess*", o *dataset* desejado, no caso ele se encontra como arquivo *.data* no computador, este arquivo deve vir acompanhado de outro com mesmo nome e extensão *.names*. Use a opção "*Open File...*" para encontrar esse *dataset*.

Durante a execução do trabalho, preferimos manter os *datasets* em *.arff*, formato que o programa disponibiliza para serem salvos os *datasets*. Transformamos então o *dataset spambase.data* para *.arff* para facilitar o trabalho.

Uma vez carregado, a tela será parecida com algo assim:

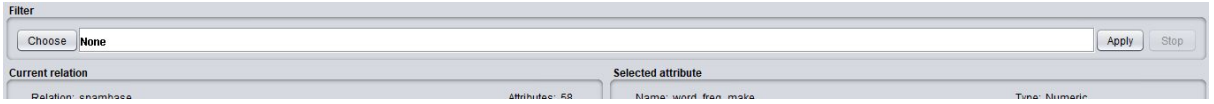


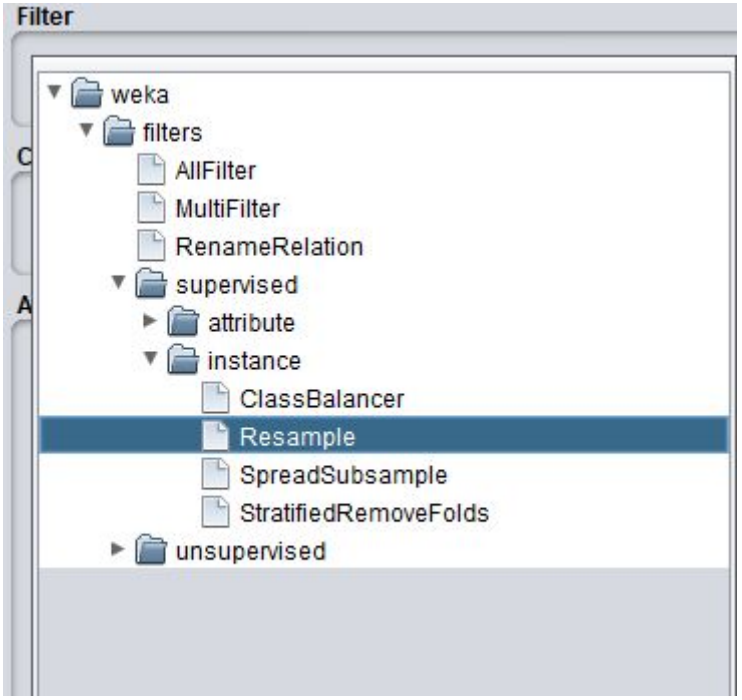
Uma vez carregado o *dataset spambase*, podemos começar a trabalhar nos dados e nos algoritmos.

Conjunto de Teste e Treinamento

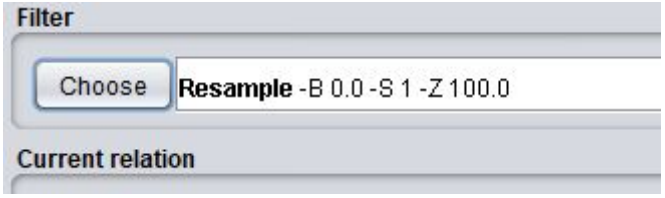
Nesta seção será demonstrado como foram feitas as separações dos conjuntos de treinamento e teste.

Para que a separação seja possível, devemos ter nosso *dataset* aberto no "*Explorer*" e na aba "*Preprocess*". Feito isso, vamos selecionar o filtro que irá fazer a separação dos dados do nosso *dataset*. No campo "*Filter*", selecione "*Choose*" e escolha pela pastas, o filtro que iremos usar, no caso será o "*Resample*", que se encontra em "*Supervised*" e na subpasta "*Instance*".

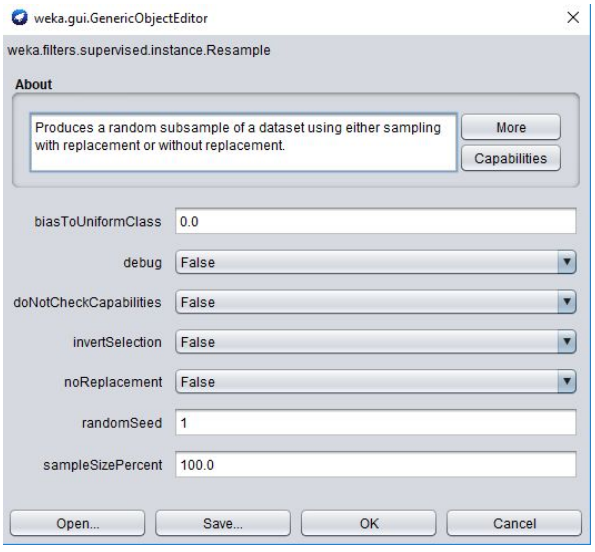




Selecionando com um clique, o filtro será colocado na caixa de "*Filter*".

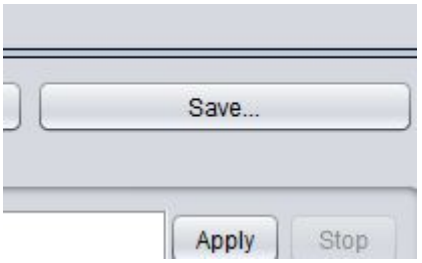


Para customizarmos as opções do filtro, vamos clicar em cima do campo com seu nome, uma caixa de diálogo aparecerá para personalizarmos o filtro.



Colocaremos em "*sampleSizePercent"*, a porcentagem do *dataset* que gostaríamos que ele selecionasse aleatoriamente e colocasse em um novo *dataset*. O "*Resample"* escolhe aleatoriamente as instâncias e tenta manter a mesma porcentagem de classes nestes *datasets*. Para o conjunto de treinamento, usaremos 70% do *dataset* original. E para o de teste, usaremos 30% para completar os 100% do *dataset*.

Após colocarmos as opções e fecharmos a caixa de propriedades, podemos então aplicar o filtro com "*Apply"*. Ele fará as alterações no *dataset*, e então podemos salvar esse novo *dataset* para usarmos nos algoritmos e aplicar outros filtros.



Para o trabalho, enunciamos os *datasets* de treinamento como *'spambase-70'*, e os de teste como *'spambase-30'*, suas variações pelos hífens subsequentes. Como por exemplo: *'spambase-70-discretizado'*, *'spambase-70-padronizado'* etc.

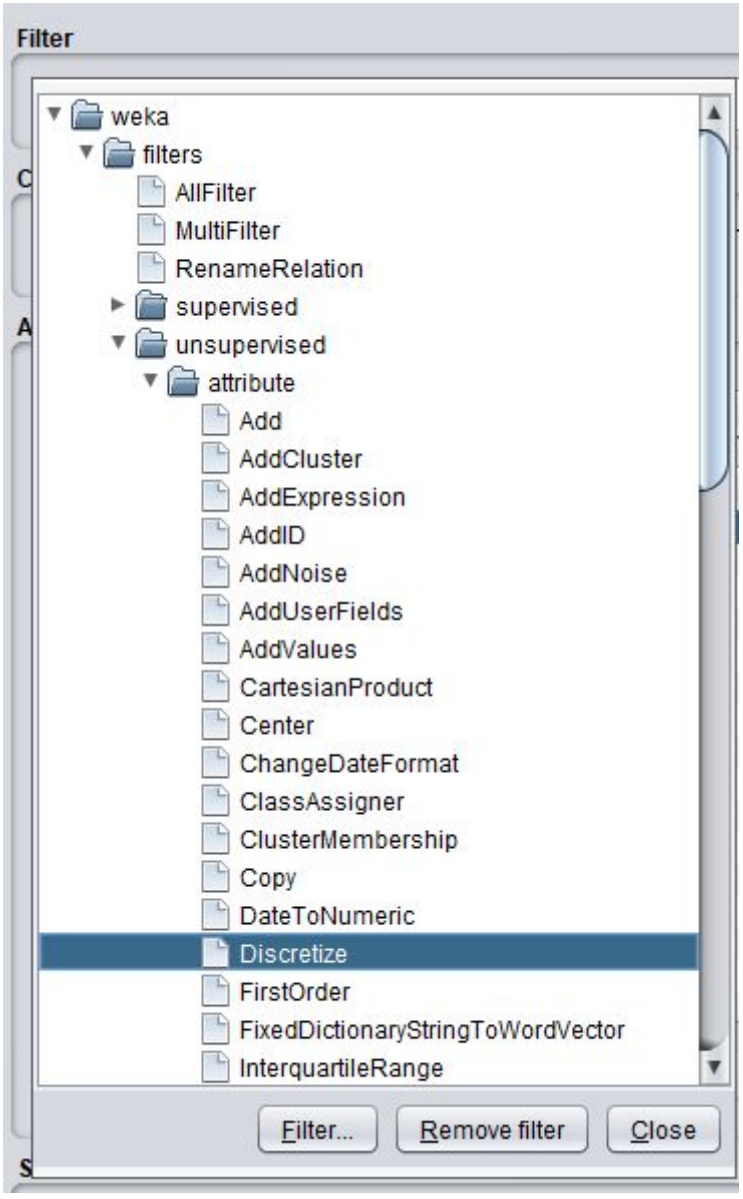
Obs: as modificações no dataset podem ser desfeitas, então é crucial salvar os datasets corretos, caso contrário, problemas podem ocorrer caso use mais de um filtro sem trocar de dataset ou voltar ao original.

Discretização

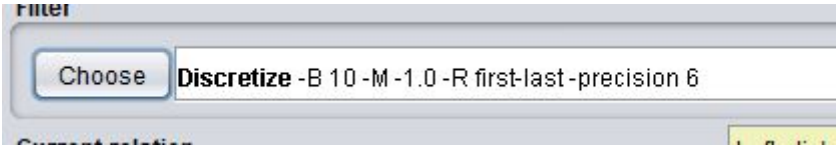
Para que seja possível trabalhar com o algoritmo de Naive Bayes, devemos discretizar os nossos dados a fim de evitar problemas nas contas que ele deve fazer. Sendo assim, vamos aplicar a discretização em ambos os conjuntos, o de treinamento e o de teste, aplicar um com discretização e outro sem pode gerar inconsistências na execução do algoritmo.

Usando os *datasets* que separamos, vamos acessar abrir-los com o *"Open File…"* para aplicarmos a função de discretização. Uma vez aberto o *dataset*, temos de ir novamente na parte de *"Filter"* para selecionar o novo filtro que fará a discretização dos nossos dados.

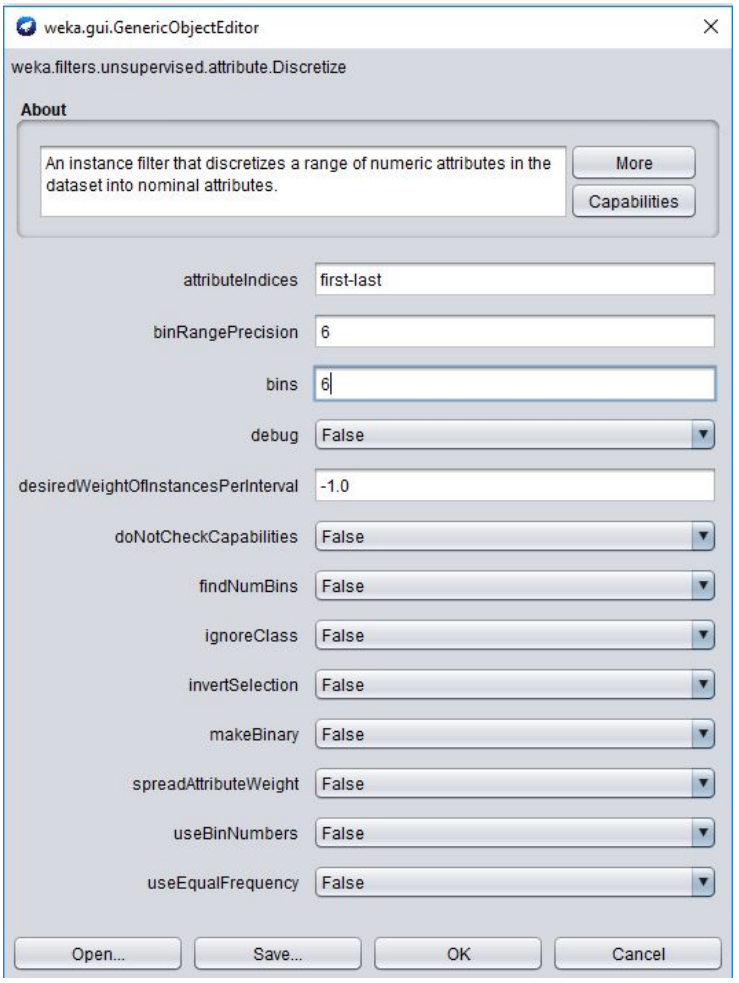
O novo filtro que usaremos é o *"Discretize",* que se encontra na pasta "*Unsupervised"*, e na subpasta *"Attribute"*.



Uma vez selecionado o filtro, vamos nas opções deste filtro, fazendo o mesmo procedimento feito durante o *"Resample"*.



Vamos então modificar as propriedades da função de discretização.



A parte que nos interessa são as *"bins"*, ou pacotes, latões, vamos definir como exigido no enunciado, 6 divisões como parâmetro para a função fazer a discretização corretamente.

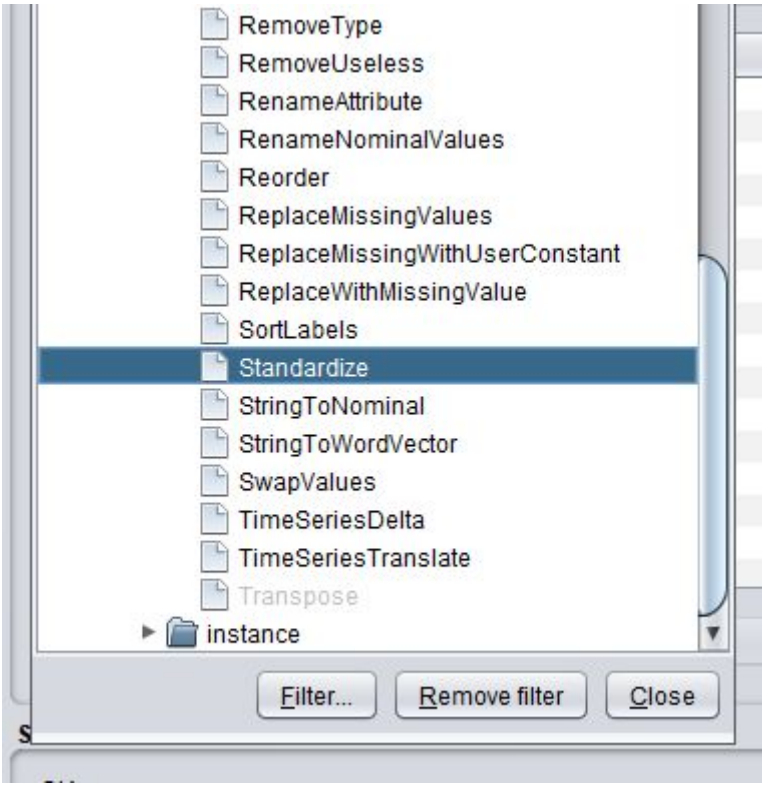
Após isso, aplicamos com *"Apply"* e salvamos esse *dataset* com um nome de *'spambase-70-discretizado'* para usarmos posteriormente. Isso também deve ser feito com o conjunto de teste, afinal eles devem ser compatíveis.

Padronização

Como exigido no enunciado do exercício programa, devemos padronizar o nosso *dataset* para fazermos execuções dos algoritmos como comparação sem a padronização. Lembrando que tanto o de treinamento quanto o de teste devem seguir o mesmo padrão para evitar inconsistências.

Fazermos o mesmo procedimento de carregamento do *dataset*, é interessante afirmar que se pode colocar um *dataset* filtrado para ser filtrado novamente, nesse caso, poderíamos ter um *dataset* que foi discretizado.

A função que vamos aplicar neste *dataset* é a função de *"Standardize"*, que se encontra na parte *"Unsupervised"* e na subpasta *"Attribute"*.

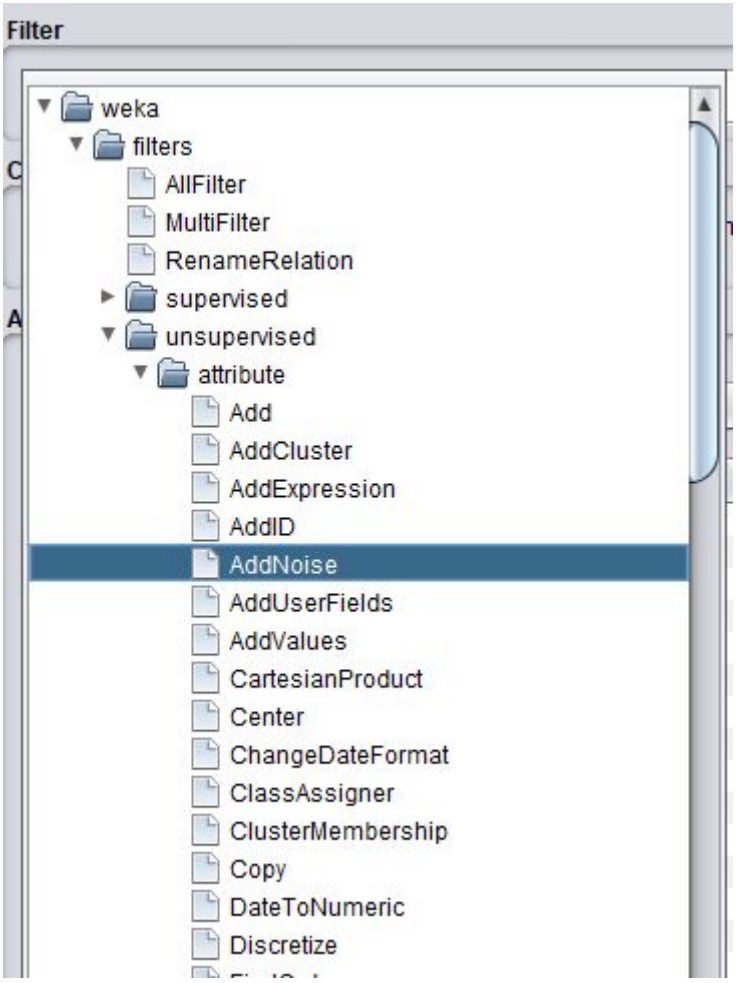


Após a seleção da mesma, só é necessário aplicar as alterações com o "*Apply"* e salvar o *dataset*.

Ruído

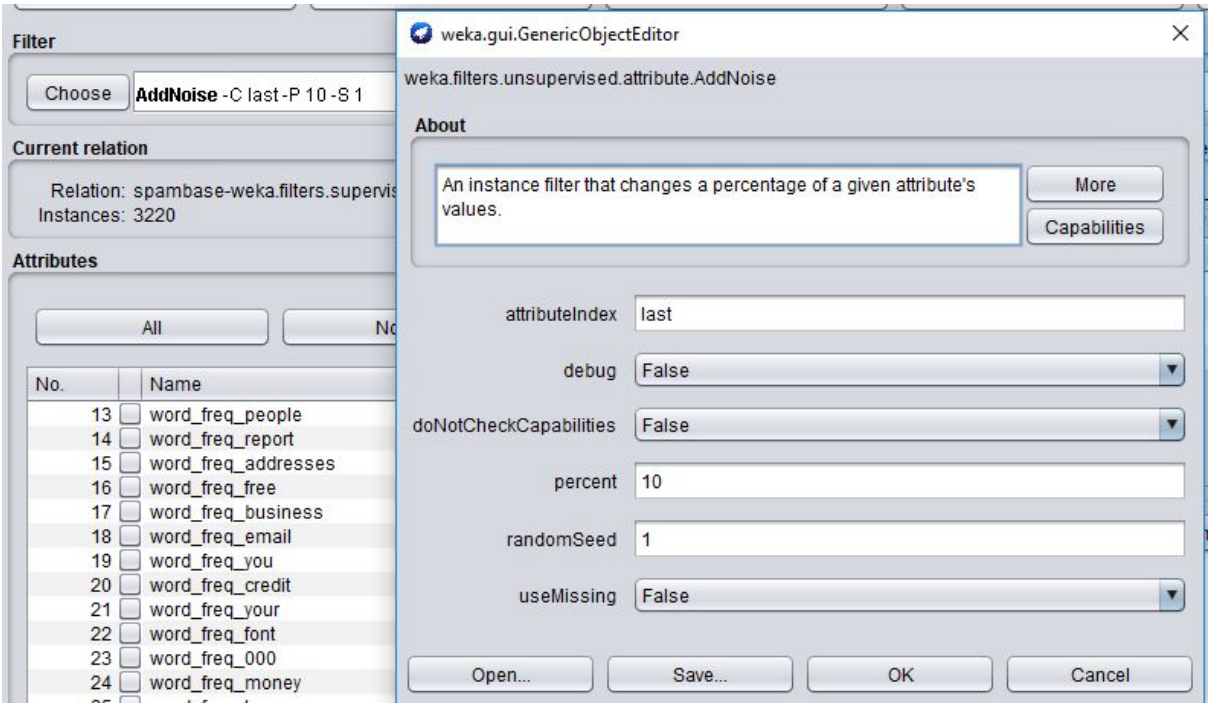
Outra parte exigida no trabalho, era a aplicação de ruídos nos dados do *dataset*, os ruídos podem ser aplicados em todos os *datasets* com filtros, também será uma exigência para que a parte de comparação com ruídos possa ser feita.

Para aplicarmos ruídos em nossos *datasets*, utilizaremos a função *"AddNoise"*, que se encontra na parte *"Unsupervised"* e na subpasta *"Attribute"*.



Adicionado o filtro ao *dataset*, podemos então modificar os parâmetros do filtro para colocarmos quanto de ruído nós queremos em nossos *datasets*, somente o conjunto de treinamento deve possuir o ruído, os de teste devem permanecer sem ruído.

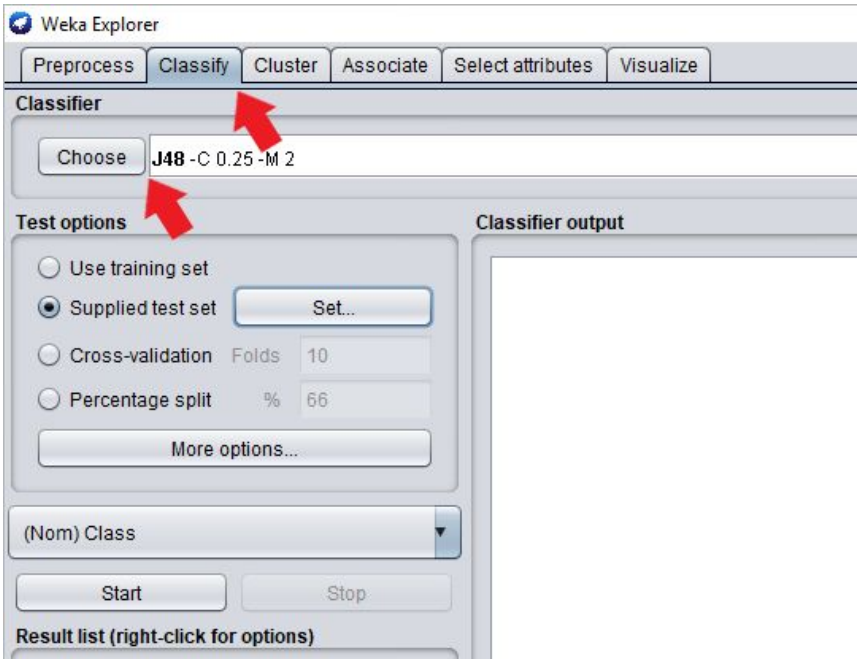
Para modificarmos a porcentagem de ruído nos *datasets*, abrimos a caixa de propriedades do filtro.



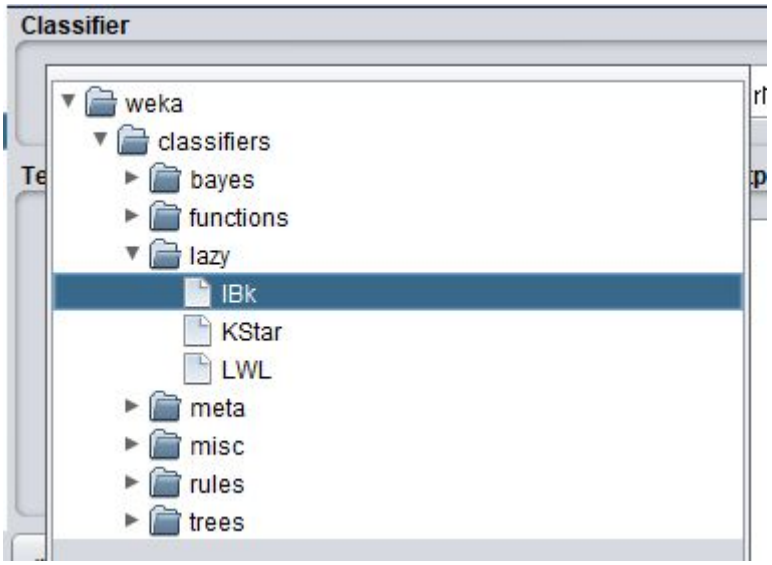
Na opção *"percent"*, inserimos quanto de ruído gostaríamos em nosso *dataset*, no trabalho, varia de 0% a 25%, Lembrando que não é necessário utilizar essa função para 0% de ruído, só utilizar o mesmo *dataset* sem o ruído.

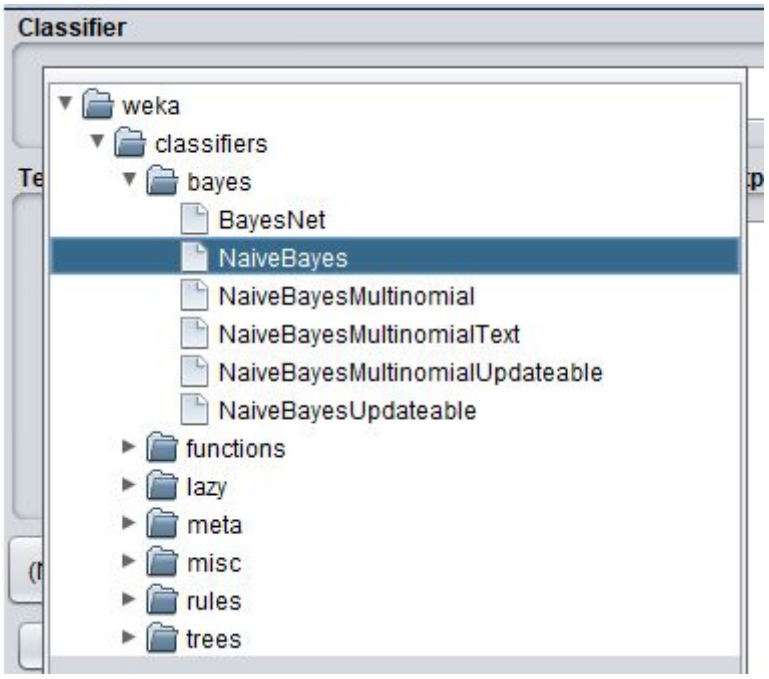
Execução de algoritmos

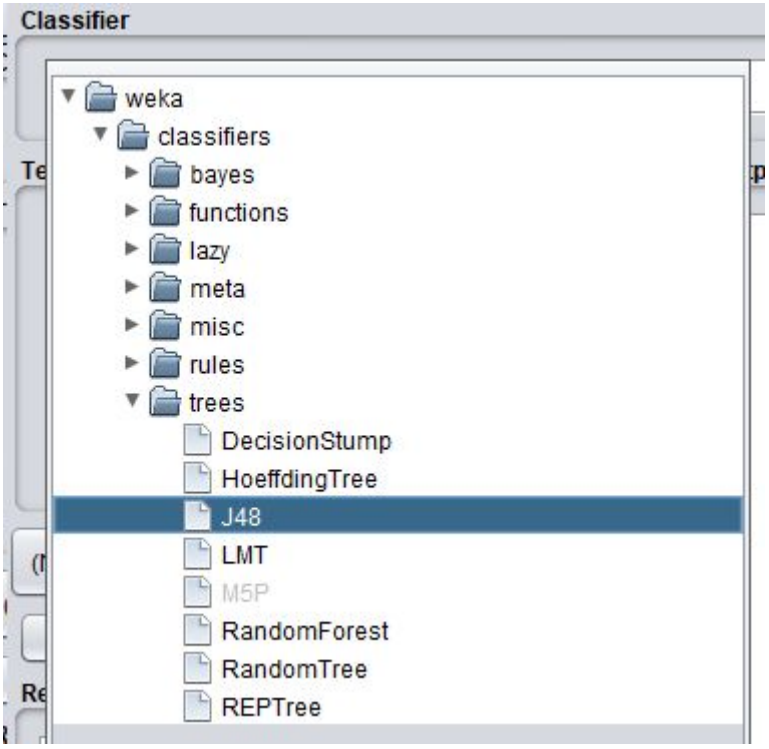
A execução dos algoritmos é feita na aba *"Classify"* e para selecionar o algoritmo desejado aperte o botão *"Choose"*, como mostrado abaixo:



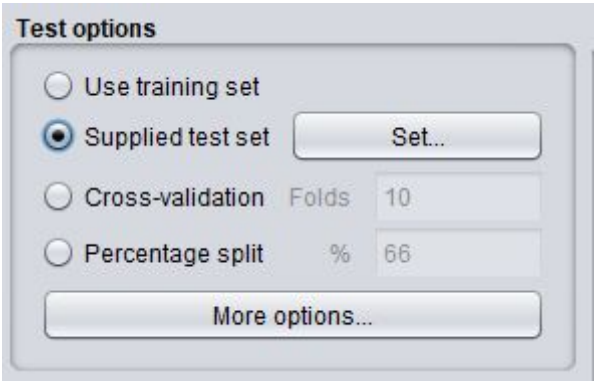
A seguir a seleção dos algoritmos KNN, Naive Bayes, e Árvore de decisão respectivamente:







Para selecionar o conjunto de Teste que os algoritmos irão executar sobre o conjunto de treinamento, selecione a opção *"Supplied test set"* como mostrado abaixo:

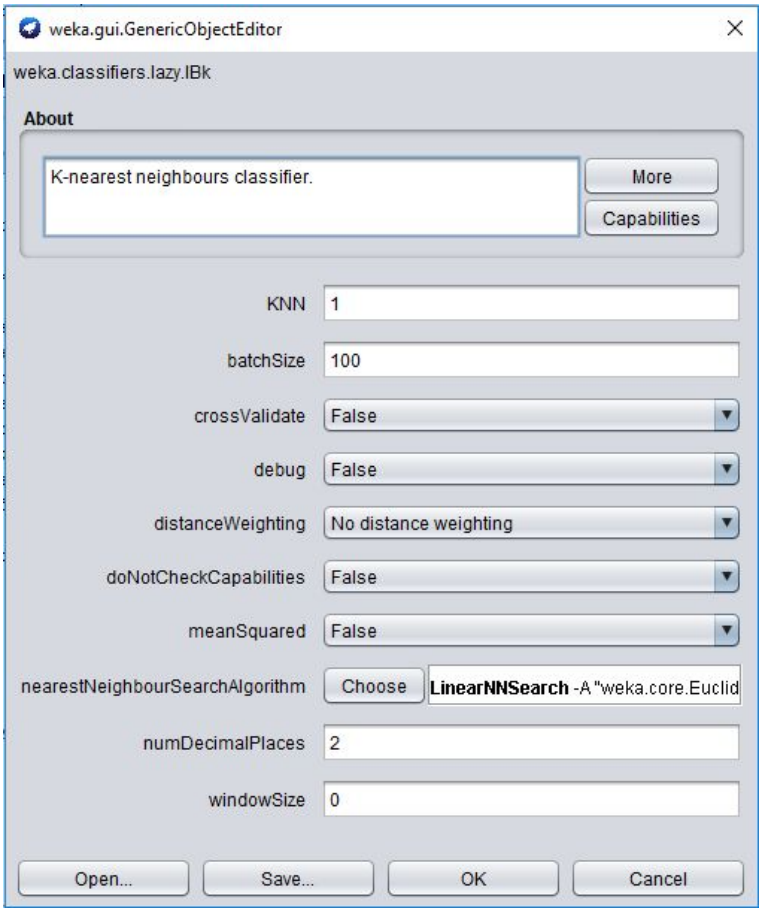


Logo após aperte o botão *"Set..."* e em seguida o botão *"Open file..."* e selecione o arquivo de dados de teste

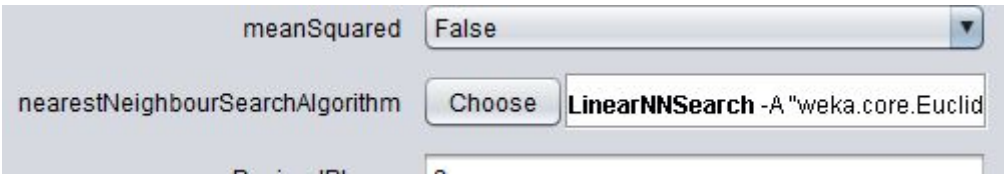
Esse procedimento de seleção do conjunto de teste é realizado para todos os algoritmos testados neste trabalho.

KNN

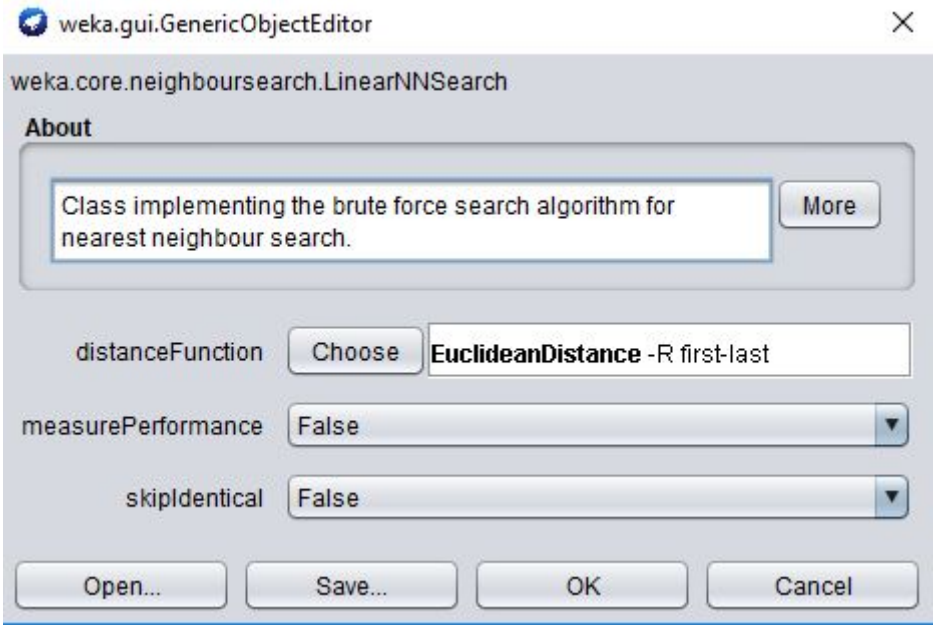
Com o algoritmo KNN selecionado, acessamos as opções do algoritmo pela barra com seu nome, abrindo a caixa de opções.



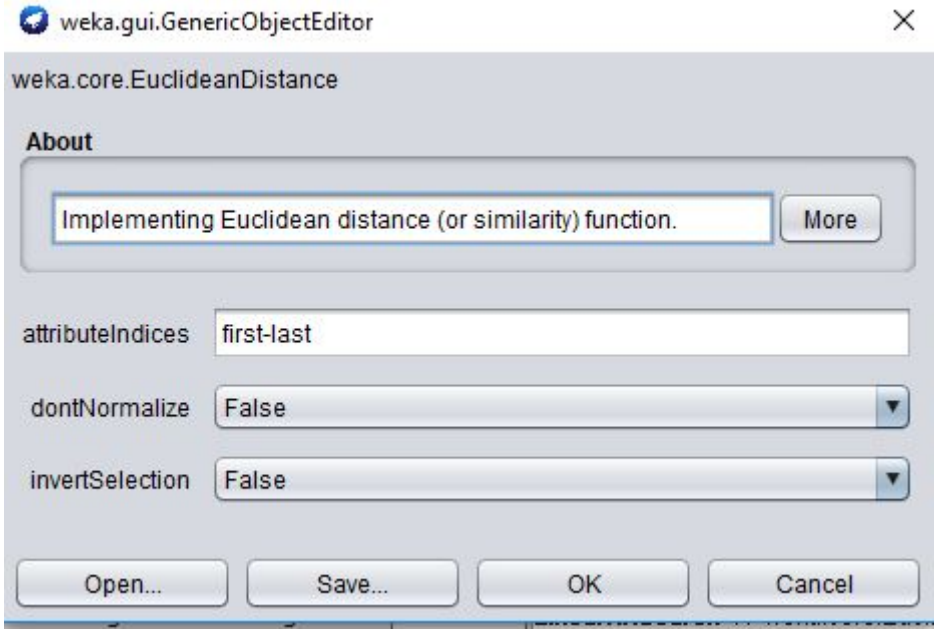
Existem duas opções que nos interessam nesta janela de propriedades, o algoritmo de busca pelo vizinho mais próximo e a quantidade de centróides a serem utilizados. A variação na quantidade de centróides que podem ser utilizados no algoritmo pode ser definida em *"KNN"*. Para alterarmos o algoritmo de distância utilizado, selecionamos a opção *"Choose"* ao lado de *"nearestNeighbourSearchAlgorithm"*, as opções que nos interessam são as distâncias da busca linear, *"LinearNNSearch"*.



Para selecionarmos qual distância queremos utilizar, clicamos no texto da caixa, e outra janela aparecerá.



Em *"distanceFunction"*, vamos selecionar entre as distâncias Euclidiana e de Manhattan, a de Manhattan será utilizada somente uma vez durante a execução do trabalho. Para não termos problemas com as distâncias, clique também no nome da mesma para abrirmos outra caixa de edição.



Nesta caixa de edição, devemos colocar *“dontNormalize”* para *“True”*, para evitar que ele normalize a distância e gere resultados inconsistentes. Feito isso, o KNN estará pronto para rodar e analisar seus dados.

Naive Bayes

Para a execução do Naive Bayes deve ser seguido a função apresentada anteriormente, não há configurações a serem feitas para o algoritmo, a única garantia que deve ser feita é que os dados que forem inseridos, sejam discretizados, procedimento explicado anteriormente.

Árvore de Decisão

Com o algoritmo de árvore de decisão selecionado, abra as opções do algoritmo clicando duas vezes em seu nome. A opção destacada abaixo, *‘unpruned’* faz com que o algoritmo execute com ou sem poda, sendo que o valor ‘*true’* corresponde a árvore sem poda e ‘*false’* a árvore com poda