Segunda entrega do projeto final de ‘Soluções em Mineração de Dados’ - Relatório

Aluno: Andrey Moutelik

Descrição das features:

xm e ym são a posição (X,Y) de cada boid;

xVeln e yVeln são o vetor de velocidade;

xAm e yAm são o vetor de alinhamento;

xSm e ySm são o vetor de separação;

xCm e yCm são o vetor de coesão;

nACm é o número de boids em um raio de Alinhamento/Coesão;

nSm é o número de boids no raio de Separação.

Esses atributos são repetidos para todos os m boids, onde m = 1,...,200. (12*200 = 2 400)

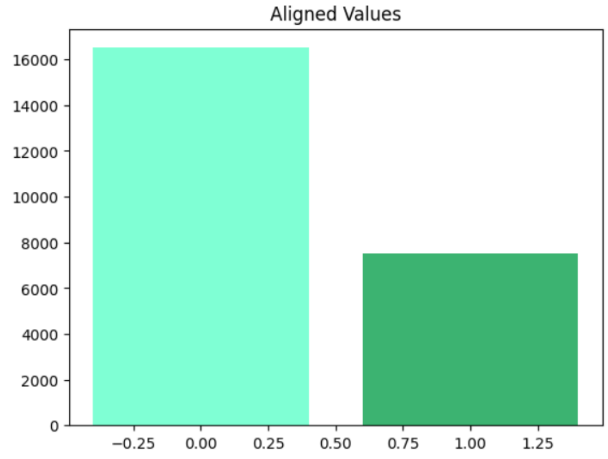
Resultando em **2400 FEATURES**

O dataset ‘Swarm Behavior Data’ contém três CSVs distintos, Aligned, Flocking e Grouped, cada um classificando os dados respectivamente em Aligned ou Not Aligned, Flocking ou Not Flocking, e Grouped ou Not Grouped.

Seguindo o artigo de EDA passo pelas etapas iniciais: checagem de valores faltantes (não há valores faltantes em nenhum dos datasets), checagem de linhas duplicadas (também não há)

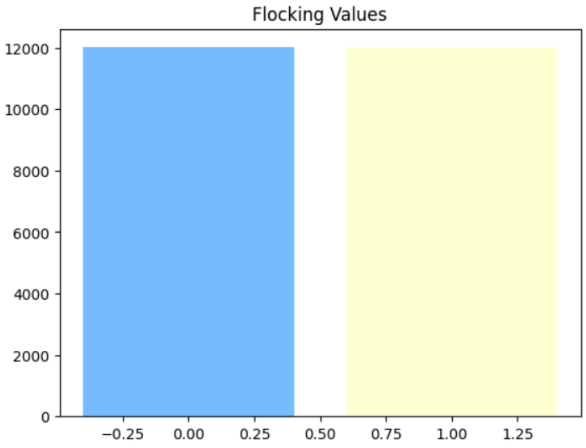
Um passo adicional que fiz foi verificar se os datasets, com exceção do atributo final de classe, são idênticos. Aligned e Grouped são, Flocking é diferente dos dois.

Checo o balanceamento dos datasets:



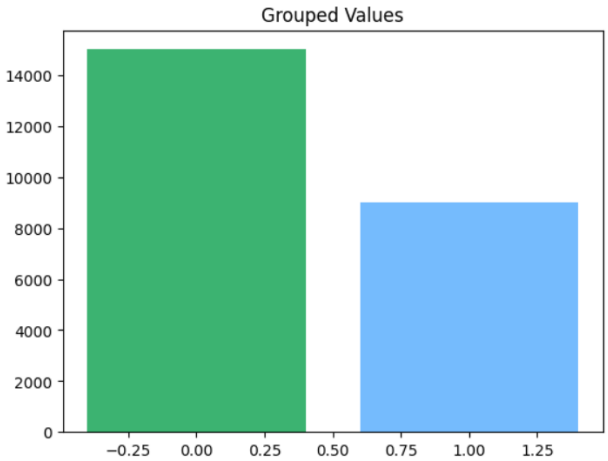

```
0    16511
1     7505
Name: Class, dtype: int64
```

Com relação a Aligned, vemos um pequeno desbalanceamento, com cerca de 69% correspondendo a primeira classe




```
0    12008
1    12008
Name: Class, dtype: int64
```

Já com relação a Flocking, temos o dataset perfeitamente balanceado, com 12008 elementos de cada classe




```
0    15010
1     9006
Name: Class, dtype: int64
```

Por fim, com relação a Grouped, voltamos a ter um pequeno desbalanceamento, com uma classe representando 62,5% do total.

Agora vamos falar sobre o elefante na sala:

Não é realista fazer uma análise exploratória feature por feature, como é descrito no artigo de EDA, neste dataset. Temos simplesmente features demais, e eu as vejo praticamente como uma “caixa preta”, pois não posso inferir conclusões a partir delas. Nosso interesse é identificar o comportamento do todo, e, portanto o atributo do comportamento de um ‘boid’ fora do contexto geral não tem potencial de previsão nenhum.

Com relação a identificação de outliers, também não acredito que seja algo aconselhável devido ao contexto do problema. Os dados provêm das simulações usadas para identificação de padrões de comportamento de grupos, e eu não sei como determinar, para cada feature do dataset, o que seria um outlier ou não.

Por fim com relação aos passos seguintes do projeto, vejo como essencial a redução no número de features no pré-processamento. Pretendo aplicar Principal Component Analysis (PCA), para identificar os atributos mais importantes e reduzir drasticamente o número dessas features. Note que cada dataset deverá terminar com features diferentes, se eu insistir na ideia de continuar com as três classificações distintas.

Fora isso, também utilizarei técnicas para lidar com o desbalanceamento das classes de Aligned e Grouped.