Universidade Federal de Rio Grande do Norte
Departamento de Informática e Matemática Aplicada
Disciplina: Aprendizado de Máquina

Disciplina: Aprendizado de Máquina
Discentes:
Gabriel Cristian Melo da Silva
Gabriel Estácio de Souza Passos
Rodrigo Faustino de Sousa

Docente: SAMUEL DA SILVA OLIVEIRA

CheckPoint – Modelos Não Supervisionados

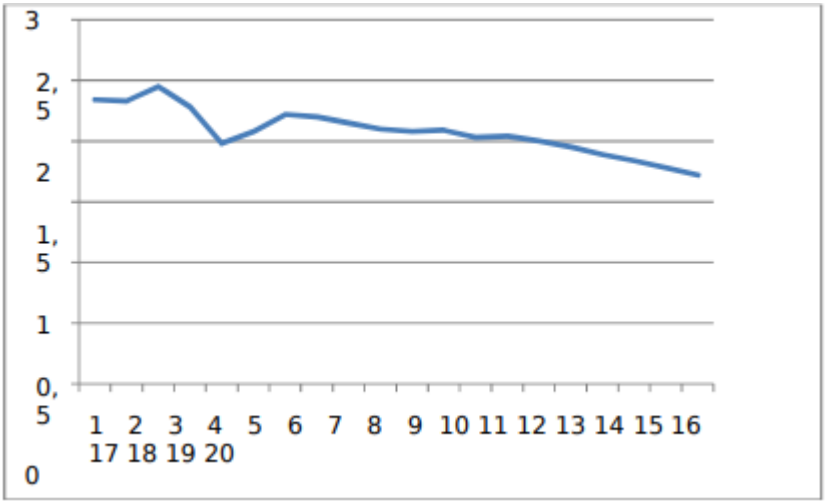
1. Objetivos:

O principal objetivo desta parte do trabalho prático é analisar como os métodos e técnicas não supervisionados vistos em sala de aula se comportam em uma aplicação prática. Para tal, o aluno terá o seguinte objetivo:

I. **Aprendizado não supervisionado:** Analisar os três principais métodos não supervisionados vistos em sala de aula (k-means, Hierárquico e Expectation–Maximization (EM)) e tentar definir o melhor número de grupos para o problema em questão, baseado nos resultados desses algoritmos.

2. Metodologia do trabalho

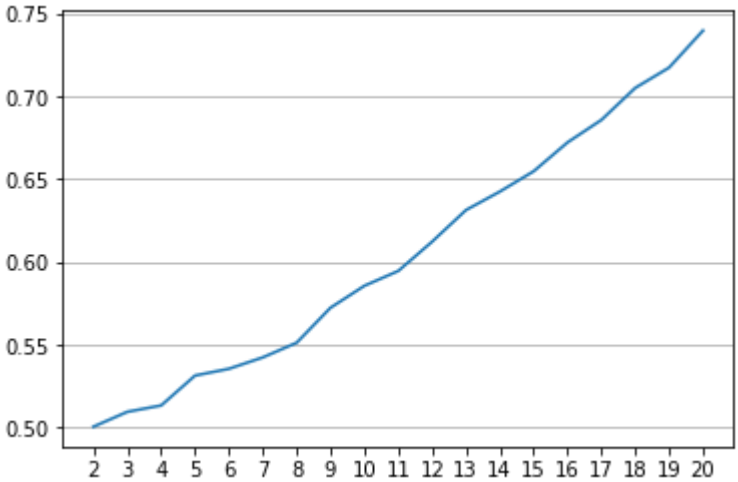
Como já mencionado, três métodos de aprendizado não supervisionados de máquina devem ser utilizados neste trabalho, que são: k-means, Hierárquico aglomerativo e EM. Para a utilização destes algoritmos na base de dados que foi escolhida, é preciso retirar o atributo classe e, então, o mesmo está apto para utilização.



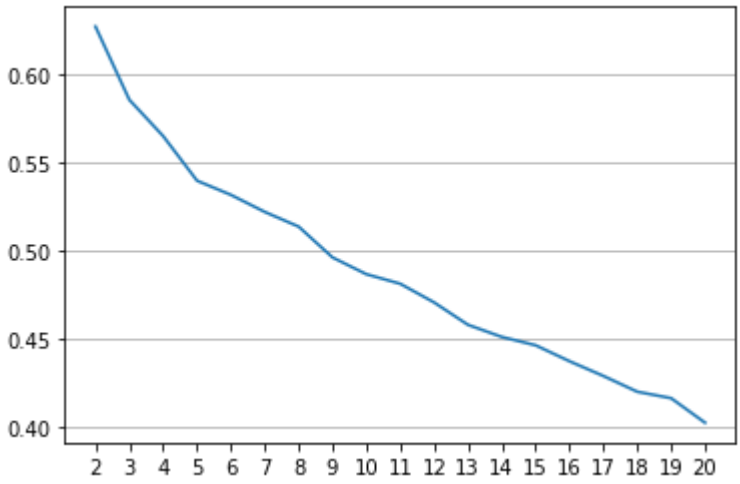
2.1.1. **k-means:** serão feitos experimentos com *k* variando de 2 até 20. Por ser um método com inicialização, para cada valor de *k*, serão feitas 5 execuções (variando o valor da seed). Após os experimentos, serão calculados os índices DB e Silhouette de todos as partições geradas (agrupamentos construídos). Uma vez calculados os índices DB e Silhouette, a média e o desvio padrão por valor de *k* são calculados para cada índice. Coloque os resultados em um gráfico (modelo acima) para cada índice, onde o eixo x representa o valor de k e o eixo y representa o respectivo índice.

Para o DB, defina como o número de grupos mais adequado o que tiver o MENOR índice DB. O comportamento inverso é esperado para o Silhouette, quanto maior, melhor para a partição. Também defina o melhor número de grupos para o Silhouette.

DB K-Means

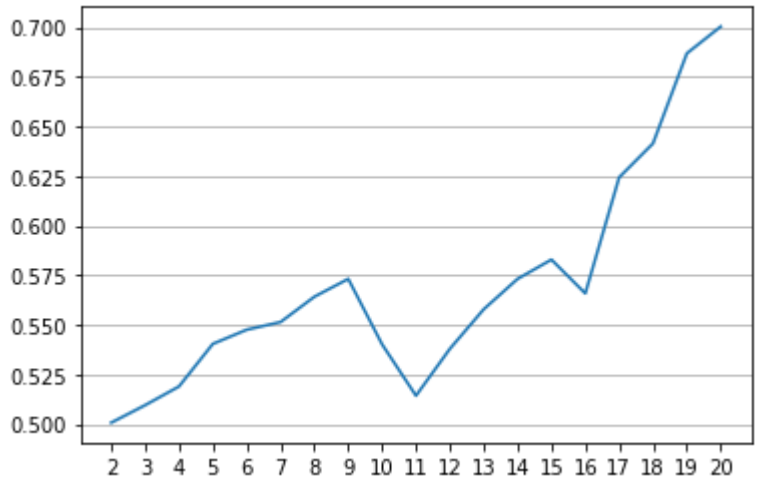


Silhouette K-Means

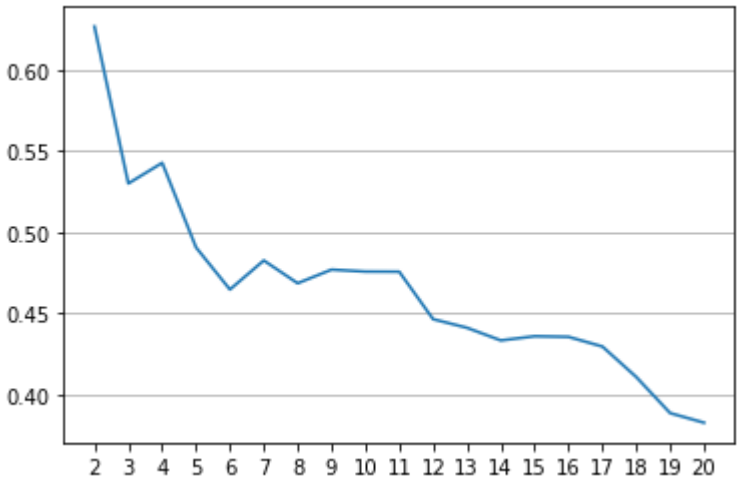


2.1.2. **Hierárquico aglomerativo:** Assim como no k-médias, serão feitos experimentos com o número de grupos variando de 2 até 20. Como este é um algoritmo determinístico, é necessário fazer apenas uma execução do algoritmo por valor de *k*, calculando na sequência os mesmos índices discutidos anteriormente. Além disso, crie os mesmos gráficos mostrados na seção 2.1.1, e por fim, defina qual o melhor número de grupos para os três índices.

DB Hierárquico Aglomerativo

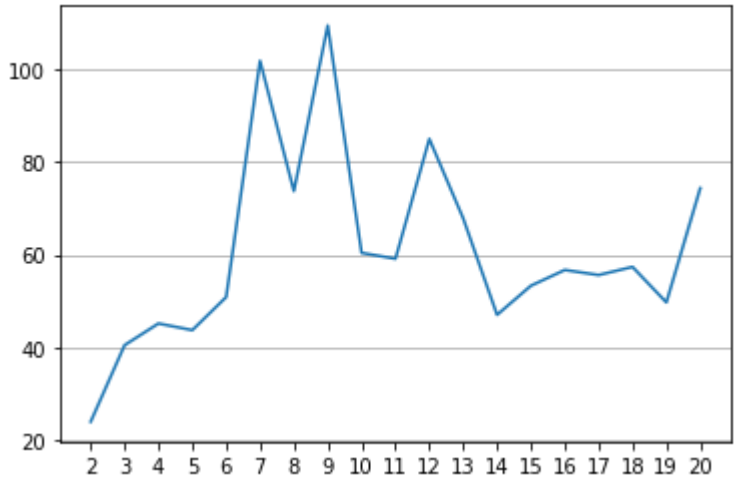


Silhouette Hierárquico Aglomerativo

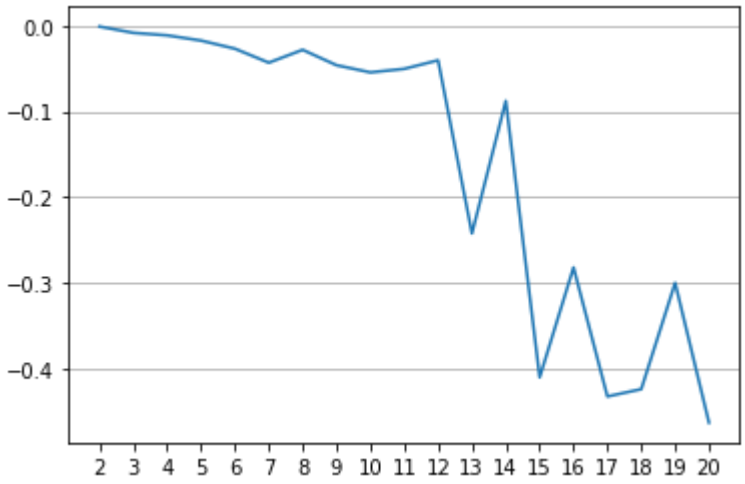


2.1.3. **Expectation Maximization (EM):** serão feitos experimentos com *k* variando de 2 até 20. Por ser um método com inicialização, para cada valor de *k*, serão feitas 5 execuções (variando o valor da seed). Após os experimentos, serão calculados os dois índices para todas as partições geradas (agrupamentos construídos). Uma vez calculado os índices, a média e o desvio padrão por valor de *k* são calculados para cada índice. Coloque os resultados em um gráfico para cada índice, onde o eixo x representa o valor de k e o eixo y representa o respectivo índice.

DB EM



Silhouette EM



Nesta fase, devemos responder às seguintes perguntas:

- **Qual foi o número de grupos definido pelo k-means baseado nos índices de validação (DB e Silhouette)?**

R. Com base nos índices, concluímos que o número de grupos que obteve melhores resultados foi 2 (Valor mais baixo do índice DB e mais alto do Silhouette).

- **Qual foi o número de grupos definido pelo Hierárquico baseado nos índices de validação?**

R. Com base nos índices, concluímos que o número de grupos que obteve melhores resultados foi 2 (Valor mais baixo do índice DB e mais alto do Silhouette).

- **Qual foi o número de grupos definido pelo EM baseado nos índices de validação?**

R. Com base nos índices, concluímos que o número de grupos que obteve melhores resultados foi 2 (Valor mais baixo do índice DB e mais alto do Silhouette).

- **Foi o mesmo resultado para os três algoritmos, se não, por que será que está havendo divergência?**

R. O resultado foi o mesmo nos três casos.

- **Para a sua base de dados, qual o melhor resultado (partição mais semelhante a categorização original), k-means, hierárquico ou EM?**

R. Com base na magnitude dos índices, o K-Means e o Hierárquico Aglomerativo mantiveram o valor do índice DB abaixo de 1 e o Silhouette acima de 0,6. Já no EM, isso não ocorreu, até no caso de menor valor (k=2), o valor do índice DB já ultrapassa 20 e o Silhouette se encontra em valores negativos. Esses valores negativos indicam que as amostras podem ter sido agrupadas em clusters inadequados. Dado isso, o K-Means e o hierárquico obtiveram resultados semelhantes, mas como os gráficos do K-Means indicaram mais estabilidade nos índices, cremos que ele tenha obtido o melhor resultado. O fato do k=2 ter nos dados os melhores resultados faz bastante sentido com a natureza do nosso problema, de decidir se o jogador foi ou não MVP da temporada (que é justamente uma classe binária).