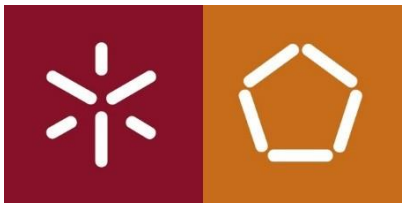


Universidade do Minho
Escola de Engenharia
Mestrado em Engenharia Informática

Sistemas Baseados em Similaridade

Ano Letivo de 2020/2021

Trabalho Prático 2

Conceção e implementação de um Sistema de Recomendação para livros

Artur Bernardo da Silva Ribeiro – a82516
Hugo Miguel Ramos Alves de Faria – pg44415
João Pedro Matos Ribeiro Soares – pg42836
Simão Pedro Martins Gonçalves, pg42850

Novembro, 2020

Índice

- Quais os domínios a tratar, quais os objetivos e como se propõe a atingi-los;
- Qual a metodologia seguida e como foi aplicada;
- Descrição e exploração detalhada do dataset e de todo e qualquer tratamento efetuado ao mesmo;
- Descrição dos workflows criados e com que objetivo (não se pretende uma descrição nodo-a-nodo). Quais os principais nodos e como foram configurados, entre outros detalhes que seja oportuno fornecer;
- Descrição detalhada do Sistema de Recomendação desenvolvido e dos paradigmas e técnicas implementadas; descrição dos problemas/desafios encontrados e como foram resolvidos;
- Sumário dos resultados obtidos e respetiva análise crítica; 2 de 2 Entrega e Avaliação
- Apresentação de sugestões e recomendações após análise dos resultados obtidos e dos modelos desenvolvidos para melhoria do Sistema de Recomendação.

indice de ilustaçoes

1. Introdução

Neste trabalho, temos como principal objetivo aprimorar os conhecimentos obtidos na unidade curricular relativamente à conceção e desenvolvimento de Sistemas de Recomendação utilizando a plataforma *Knime*.

Posto isto, numa fase inicial, começamos por escolher um *dataset* para a realização deste trabalho e após alguma pesquisa selecionamos um *dataset* correspondente à classificação de livros por parte dos utilizadores. Isto porque, consideramos importante usar um dataset que contenha as avaliações dos leitores, de forma a que seja possível aplicarmos diversas técnicas de recomendação.

Posto isto, o principal objetivo deste trabalho será recomendar livros a um utilizador com base nas suas leituras, avaliações, idade, e, também, com base na leitura dos restantes utilizadores.

2. Quais os domínios a tratar, quais os objetivos e como se propõe a atingi-los

O principal propósito deste trabalho é, através de avaliações de livros introduzidas pelo utilizador, e, também, tendo em conta a idade dos utilizadores, sugerir novas leituras com base nas avaliações dos outros leitores ou com base na idade do utilizador. Para isso, iremos utilizar um dataset inicial com 1.4 milhões de avaliações de forma a conseguirmos dar recomendações fidedignas.

Para tal, foram implementadas três técnicas de recomendação:

- ✓ Non-Personalized Top-N (Collaborative Filtering)
- ✓ Memory based (Collaborative Filtering)
- ✓ Association Rules

3. Qual a metodologia seguida e como foi aplicada

Depois de definido o propósito deste trabalho, é agora importante explicarmos a metodologia que iremos utilizar para conseguir atingir os objetivos do trabalho.

Inicialmente, iremos realizar uma exploração do *dataset*, de forma a percebê-lo melhor através dos gráficos que iremos gerar. De seguida, realizaremos um tratamento dos dados para que possamos, primeiramente, obter um bom *dataset*, e, posteriormente, construir um bom sistema de recomendação de livros. Para terminar, iremos realizar uma análise crítica dos resultados que iremos obter. Veremos mais à frente isto de forma detalhada.

Assim, de uma forma mais sintetizada, podemos dividir a metodologia que seguimos e aplicamos em:

- ✓ Exploração, análise e tratamento dos dados;
- ✓ Extração de conhecimento dos dados;
- ✓ Conceção e otimização do sistema de recomendação;
- ✓ Obtenção e análise crítica de resultados;

4. Descrição e exploração detalhada do dataset e de todo e qualquer tratamento efetuado ao mesmo

Este ponto do trabalho, é essencial para cumprirmos de forma eficiente os objetivos do trabalho. Assim sendo, iremos realizar uma pequena descrição e exploração do *dataset*.

Descrição das *features* do dataset

Inicialmente, consideramos que era essencial entender o conteúdo do *dataset* antes de passarmos a uma exploração do mesmo de forma mais detalhada e pormenorizada.

O nosso *dataset* está dividido em três ficheiros .csv, um com os dados relativos aos livros, outro com os dados relativos aos utilizadores e ainda outro com os dados relativos às avaliações de cada livro que cada utilizador leu.

Veremos na tabela seguinte os *features* que o nosso *dataset* possui, sendo que, iremos apresentar uma pequena explicação do que se trata cada um.

Users dataset

Features	Descrição da feature
Id	Identificador único utilizador
Location	Localização do utilizador
Age	Idade do utilizador

Books dataset

Features	Descrição da feature
ISBN	Identificador Internacional do livro
Title	Nome do titulo do livro
Author	Nome do autor do livro
Year of Publication	Ano de publicação do livro
Publisher	Editora que publicou o livro
Image-URL-S	URL que nos permite aceder a uma imagem pequena do livro
Image-URL-M	URL que nos permite aceder a uma imagem média do livro
Image-URL-L	URL que nos permite aceder a uma imagem grande do livro

Ratings dataset

Features	Descrição da feature
ISBN	Identificador único utilizador
UserId	Identificador Internacional do livro
Rating	Classificação de um livro por um utilizador

Preparação do *dataset* final

De modo a obtermos um *dataset* final, com todas as *features* necessárias para construir o nosso sistema de recomendação, foi necessário preparar todos os *datasets*. Tanto o *dataset Books*, como o *dataset Ratings* tiveram uma preparação mais fácil e direta através da delmitação das colunas por uma vírgula. Relativamente ao *dataset Users*, a preparação não foi tão direta, isto porque, a *feature Location* continha aspas

e vírgulas nos seus elementos. Para resolver esta situação, foi feita uma separação das colunas pelas aspas, de seguida, foram retiradas as vírgulas das colunas *Id* e *Age*, e, por fim, retiradas as colunas a mais e renomeadas as colunas necessárias.

Posteriormente, foram articulados os *datasets Books* e *Ratings* através da *feature ISBN*, e o *dataset* resultante foi articulado com o *dataset Users*, através da *feature Id.*

Exploração do *dataset*

- **Exploração empírica**

Através de uma primeira análise empírica, reparamos que tinhamos três *features* que não iriam ser necessárias para a finalidade do nosso trabalho, sendo elas as imagens dos livros. Posto isto, optamos por descartar, no tratamento de dados, as três *features* relativas às imagens pequenas, médias e grandes dos livros.

- ***Missing Values***

Através do nodo *GroupBy*, foi feita uma contagem dos *missing values*. Foram encontrados valores em falta na idade dos utilizadores, nos autores, nas editoras e, também, nas imagens grandes. A *feature* com maior registo foi a idade dos utilizadores com 277899 valores em falta.

- ***Data Explorer, Statistics, GroupBy, Box Plots***

Através dos nodos *Data Explorer* e *Statistics*, conseguimos compreender melhor como algumas *features* do nosso *dataset* variavam, tal como, a idade dos utilizadores e as suas avaliações aos livros que leram. Com a complementação da informação dos *boxplots* e das agrupações a estas informações já obtidas, conseguimos recolher várias informações importantes dos mínimos, máximos, valores médios, valores mais comuns das nossas *features*, e isto foi bastante importante para perceber melhor o nosso *dataset*, que tipo de valores tinhamos e se o *dataet* possuía valores inválidos.

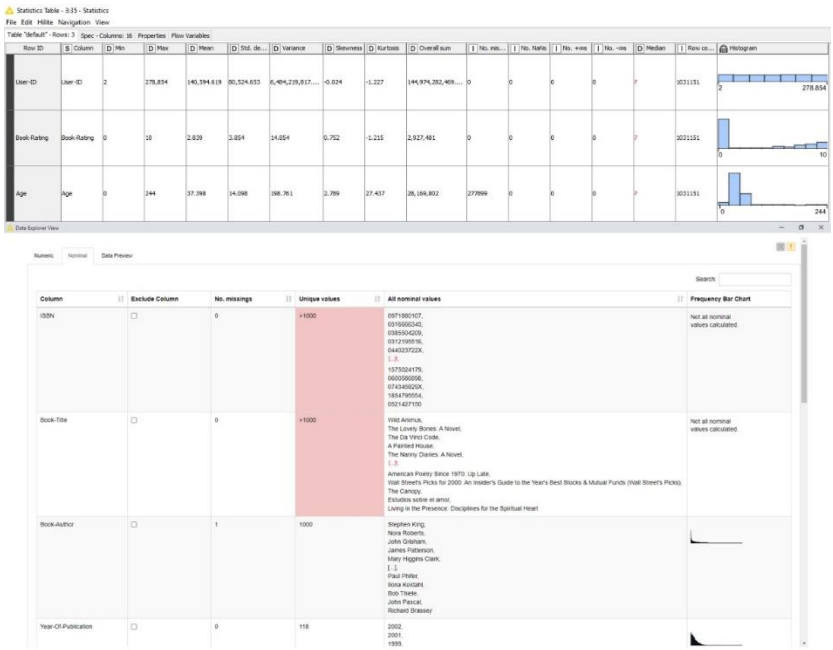


Figura 2 : Statistics e Data Explorer – Resultados obtidos

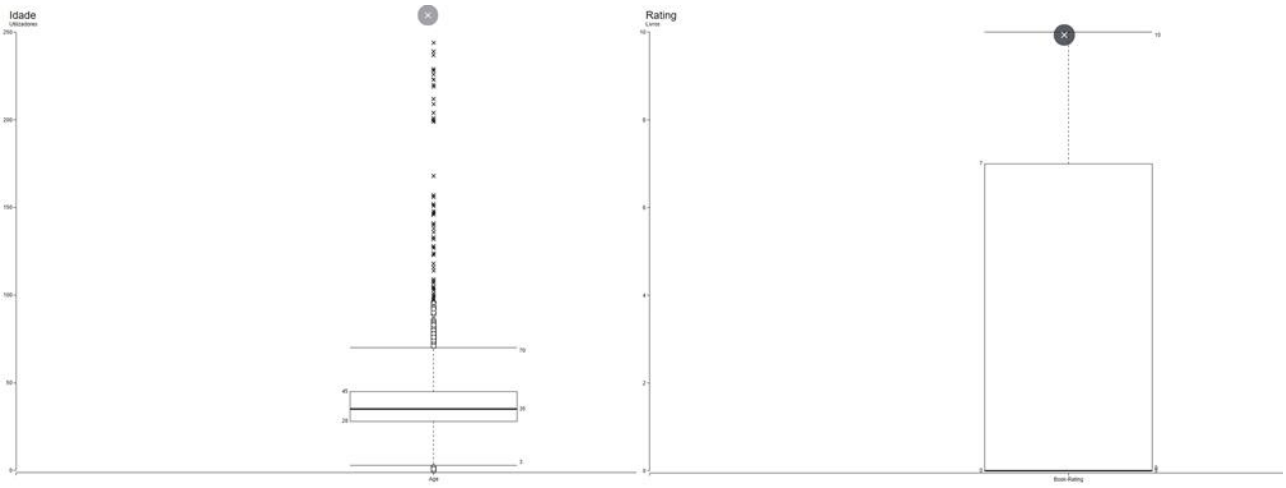


Figura 3 : Box Plot - Idade dos utilizadores e ratings

- **Correlação**

Através do nodo *Linear Correlation*, denotamos que não existe nenhuma correlação a realçar entre as *features* do nosso *dataset*.

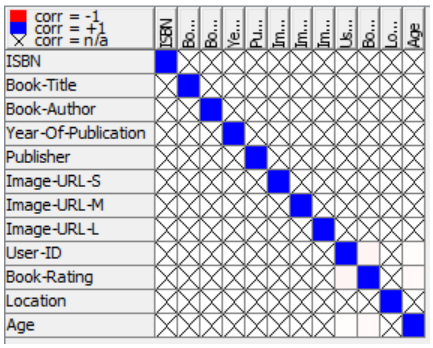


Figura 4 : Linear Correlation – Correlação entre as variáveis do dataset

Conclusões da exploração do *dataset*

Tendo em conta a análise exploratória realizada e uma primeira visualização dos dados, podemos retirar algumas conclusões e correções a fazer, *a posteriori*, no tratamento dos dados. Relativamente à *feature* idade, reparamos que continha um número considerável de valores em falta e, também, que dentro dos valores da *feature* havia valores inválidos, tendo em conta a natureza da variável. Isto porque, havia idades a variar entre os 0 e os 244 anos. Ainda dentro desta *feature*, achamos que irá ser pretinente, numa fase mais avançada, fazer um *bin* das idades. Em relação aos tipos de algumas *features*, reparamos que era necessário para fazer *cast* para inteiro de algumas, tais como, o *Id* do utilizador, a idade e o ano de publicação dos livros. No que diz respeito ao ano de publicação dos livros, verificamos que existem valores inválidos, tal como nomes, e, também, existiam vários livros com o ano de publicação zero, o que, numa fase inicial, pensamos que seriam manuscritos, por exemplo, de filosofia grega, escritos antes do ano zero mas que estariam assinalados como o ano zero. Posteriormente, achamos por bem descartar estes livros, ficando assim com os livros desde o ano de 1897 até 2020.

Tratamentos dos dados do *dataset*

No tratamento de dados começamos por eliminar todos os elementos do nosso *dataset* que possuíam *missing values*, assim sendo, foram eliminados elementos devido à falta de valores nas *features* idade, autor, editora e imagens grandes. De seguida, foram excluídas as imagens pequenas, médias e grandes do *dataset* principal. Posteriormente, foi feito o cast das *features* ano de publicação, *Id* do utilizador e idade para inteiro. Seguidamente, através do nodo *Row Filter*, foi definido um intervalo do ano de publicação dos livros que ficou definido entre 1897 e 2020, e conseguimos assim, além de eliminar anos de publicação que não interessam, eliminar anos de publicação que eram inválidos. Por fim, tendo em conta as idades que achamos minimamente aceitáveis para haver uma leitura e avaliação por parte do utilizador, definimos o intervalo de idades entre os 8 e os 95 anos.

Visualização dos dados do *dataset*

Após uma exploração e tratamento dos dados, é necessário e interessante obter uma visualização dos dados de modo a que seja possível, de forma mais fácil, haver uma extração de conhecimento dos dados. Todos os nodos utilizados para a visualização de dados foram agrupados num *Component*. Para que estas visualizações de dados fossem possíveis foi realizado algum tratamento de dados através dos nodos *GroupBy*, *Sorter* e *Row Filter* para que fossem agrupados, ordenados e limitados o tipo e número de dados que queriamos visualizar.

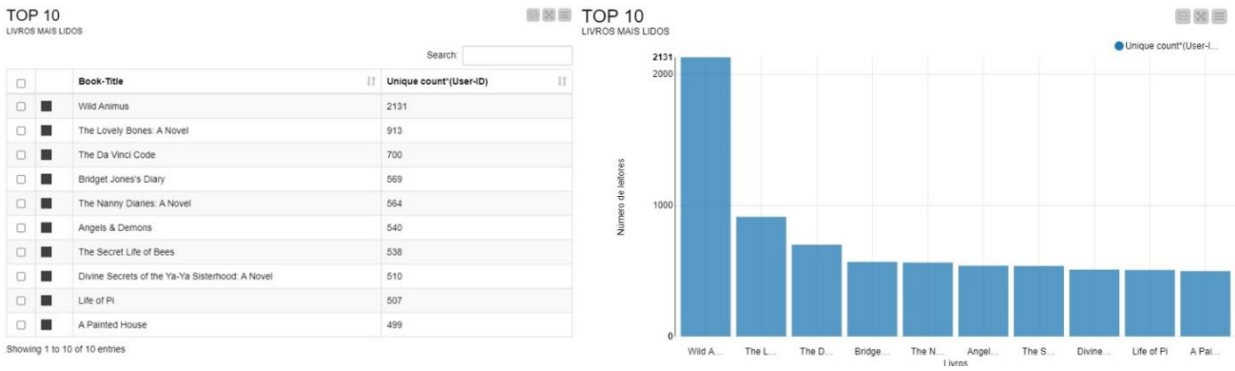
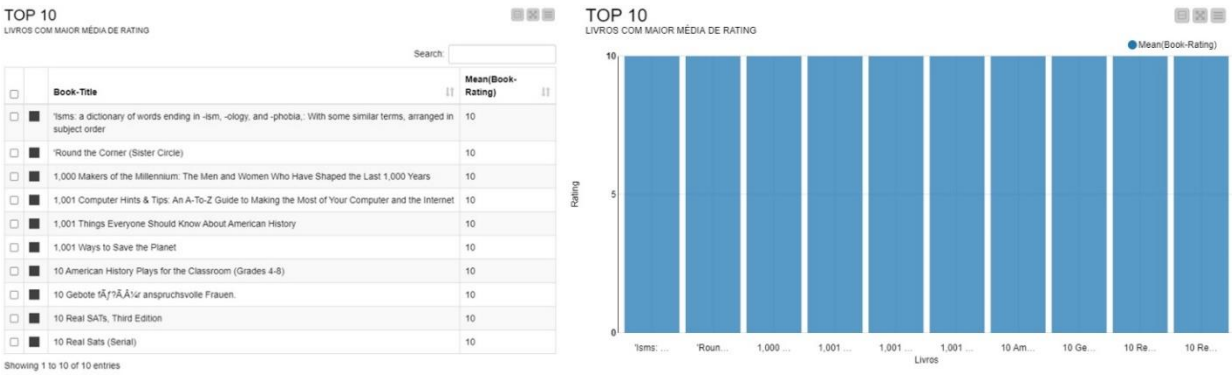


Figura 5 : Top 10 livros mais lidos



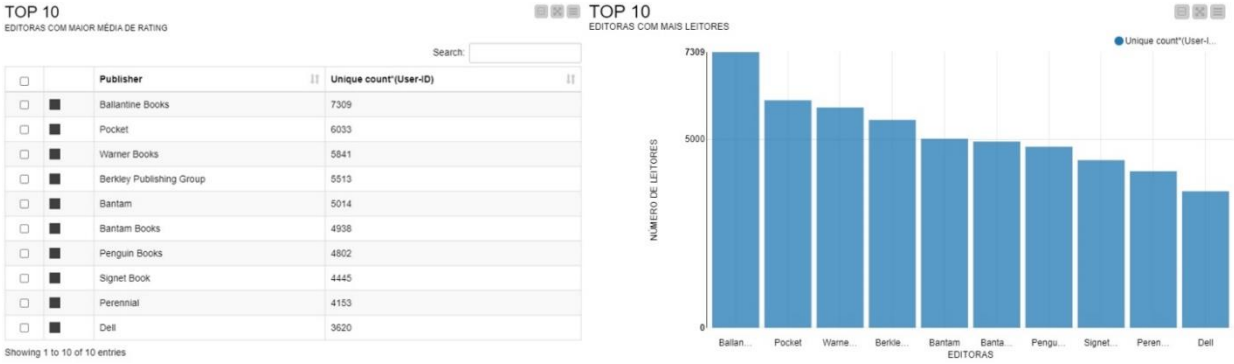


Figura 6 : Top 10 livros com maior média de rating

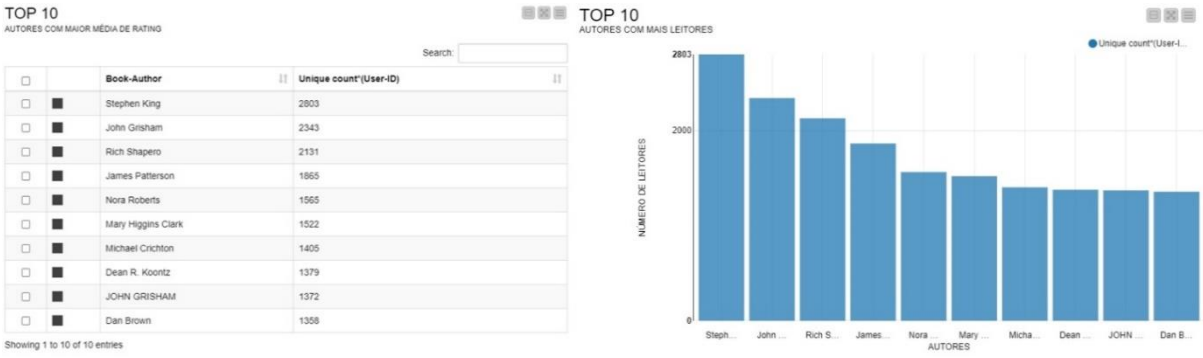


Figura 9 : Top 10 editoras com mais leitores

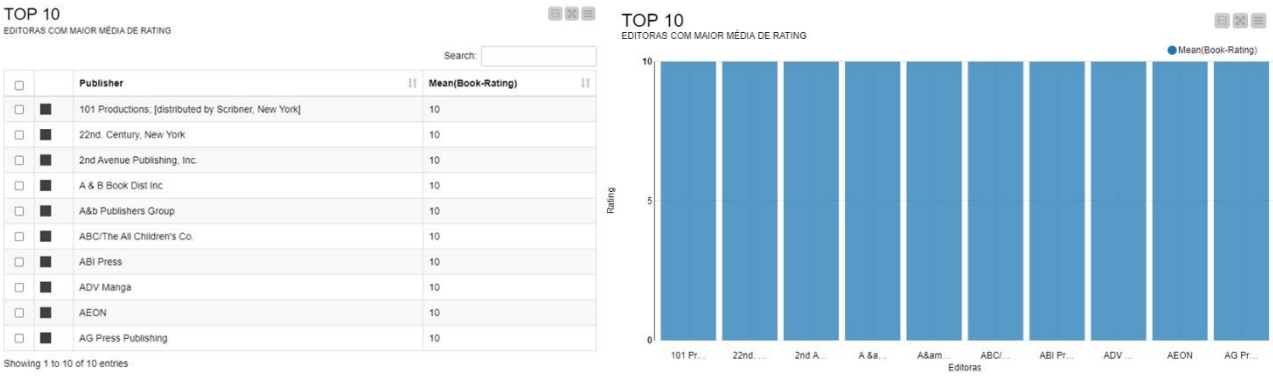


Figura 10 : Top 10 editoras com maior média de rating

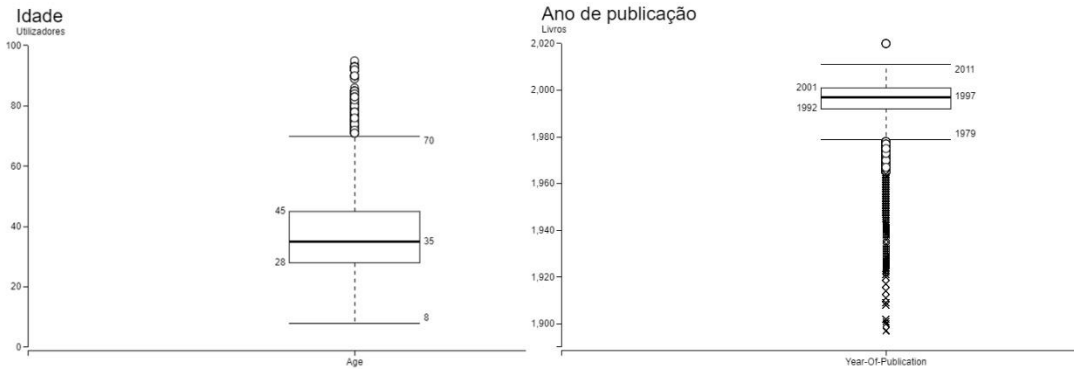


Figura 11 : Box Plot - Idade dos utilizadores e ano de publicação dos livros (Pós-tratamento dos dados)

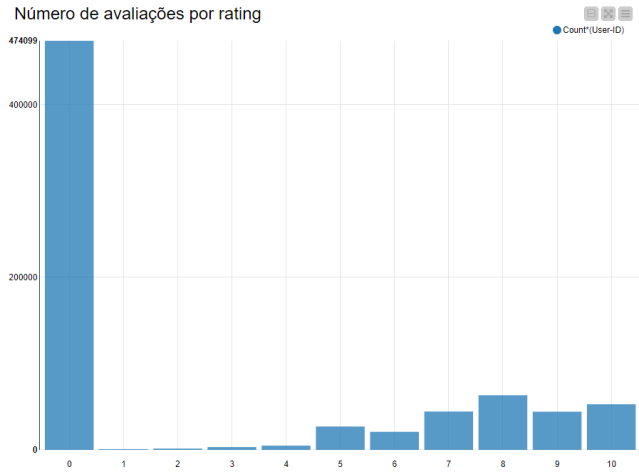


Figura 12 : Número de avaliações por rating

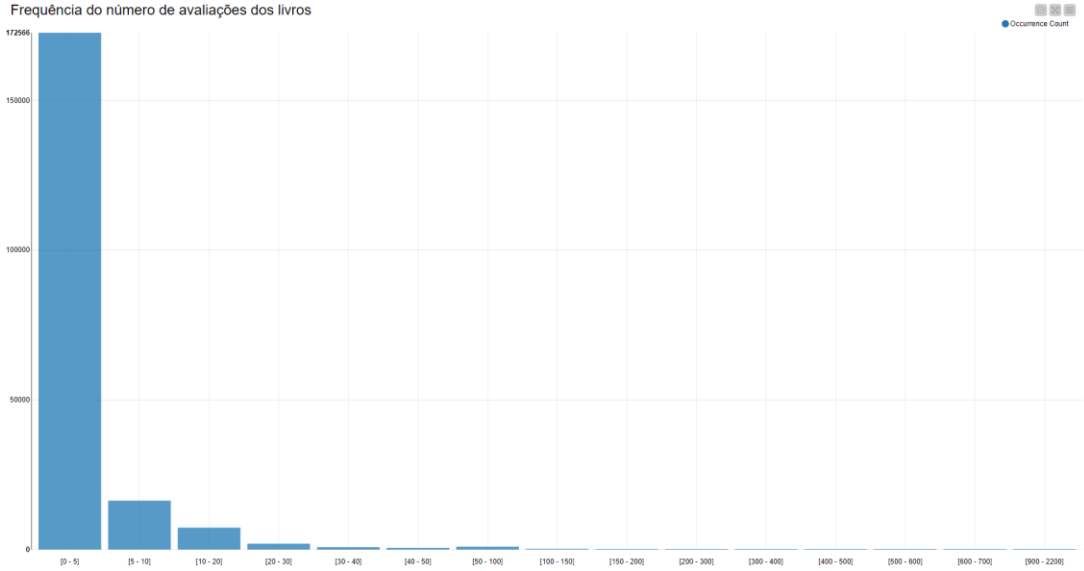


Figura 13 : Frequência do número de avaliações dos livros

Extração de conhecimento dos dados do *dataset*

Através de todos os processos de análise, tratamento e visualização de dados podemos agora extrair conhecimento importante dos dados do *dataset*. Assim sendo, podemos reparar em alguns dados interessantes, tais como:

- Os livros mais lidos não são os livros com notas de avalição mais altas (Figura 5 e Figura 6).
- Um dos problemas do *Non-Personalized Top-N* é o viés, isto porque, tem bastantes utilizadores que têm tendência a dar, sistematicamente, notas de avaliação altas. O que no caso do nosso *dataset* não acontece, sendo que das 737693 avaliações realizadas, 474099, ou seja, mais de metade foram de nota zero (Figura 12).
- A maioria dos livros têm poucas avaliações (Figura 13).

5. Descrição dos workflows criados

Após a análise, exploração, tratamento e visualização dos dados, e antes de passarmos aos sistemas de recomendação, é agora importante entender os *workflows* que realizamos. Assim, para este *dataset*, criamos um único *workflow* que está subdivido em oito momentos.

- **Leitura e construção dos *datasets***
  Nesta parte do *workflow* é realizado a importação e construção do *dataset*.

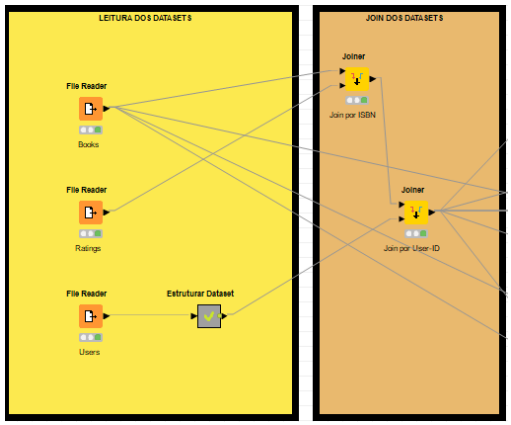


Figura 14 : Workflow - Leitura e construção dos datasets

- **Exploração dos dados**

  Nesta fração do *workflow* é explorado o *dataset*, de modo que, seja possível perceber os dados, e, também, encontrar anomalias presentes.

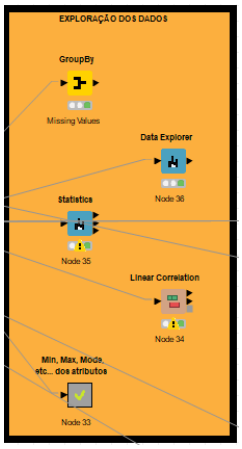


Figura 15 : Workflow – Exploração dos dados

- **Visualização dos dados (Pré-tratamento)**

  Esta parte do *workflow* complementa a parte anterior, isto porque, é uma forma útil e importante de encontrarmos anomalias nos dados, tal como ocorreu. E é, de certa forma, indespensável para a compreensão do contexto do *dataset*.

- **Tratamento dos dados**

Figura 16 : Workflow – Visualização dos dados (Pré-tratamento)

Nesta parte do *workflow* foi realizado todo o tratamento de dados necessário para construir um *dataset* final com todos os atributos necessários e sem anomalias.

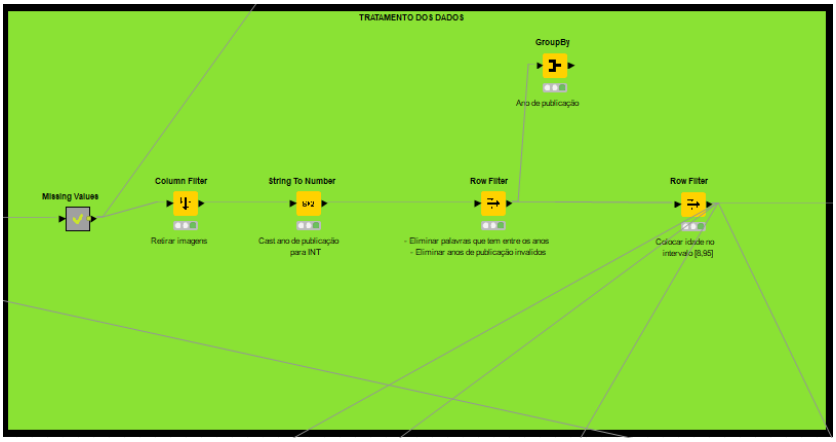


Figura 17 : Workflow – Tratamento dos dados

- **Visualização dos dados (Pós-tratamento)**

  Este troço do *workflow* é fulcral e nele são manipulados os dados para uma extração de conhecimento do *dataset* com sucesso.

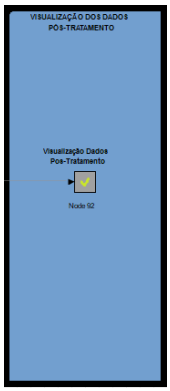


Figura 18 : Workflow – Visualização dos dados (Pós-tratamento)

- **Sistema de recomendação de livros e autores através de regras de associação**

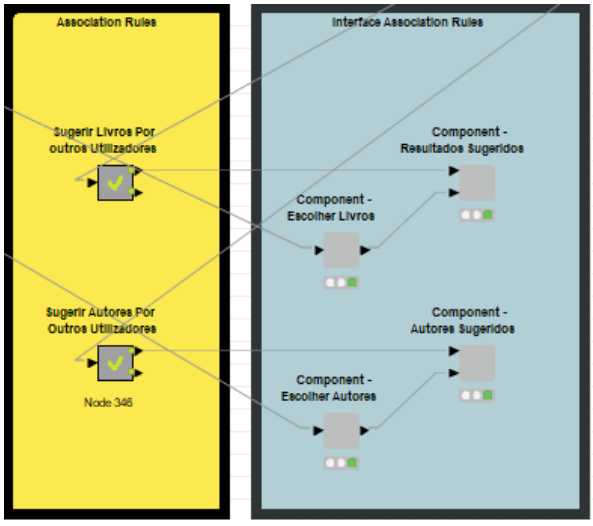


Figura 19 : Workflow – Sistema de recomendação de livros e autores através de regras de associação

- **Sistema de recomendação de livros e autores através da idade dos utilizadores (*clustering*)**

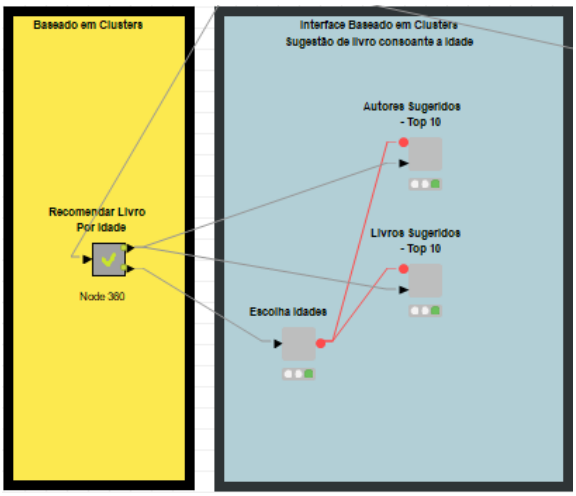


Figura 20 : Workflow - Sistema de recomendação de livros e autores através da idade dos utilizadores (clustering)

- **Sistema de recomendação de livros através de filtragem colaborativa baseada em memória (vizinho mais próximo baseado no utilizador)**

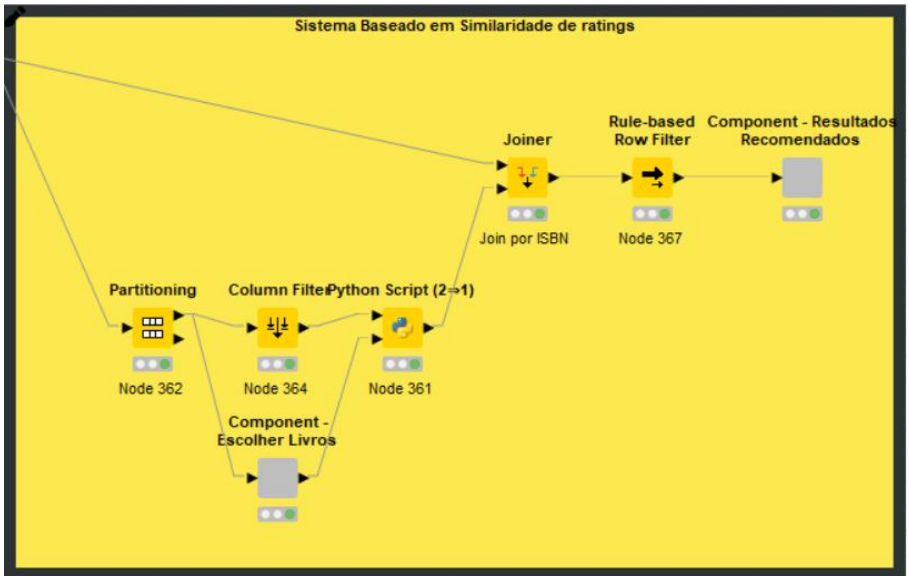


Figura 21 : Workflow - Sistema de recomendação de livros através filtragem colaborativa baseada em memória (vizinho mais próximo baseado no utilizador)

6. Descrição detalhada do Sistema de Recomendação desenvolvido e dos paradigmas e técnicas implementadas

No nosso trabalho implementamos três sistemas recomendação, sendo eles:

- ✓ Sistema de recomendação de livros e autores através de regras de associação
- ✓ Sistema de recomendação de livros e autores através da idade dos utilizadores (*clustering*)

✓ Sistema de recomendação de livros através de filtragem colaborativa baseada em memória (vizinho mais próximo baseado no utilizador)

Sistema de recomendação de livros e autores através de regras de associação

Inicialmente, criamos um *metanode* onde iremos implementar as regras de associação para sugerir os livros aos nossos utilizadores. Como o nosso *dataset* não possui uma *feature* importante, nomeadamente a categoria dos livros, optamos por sugerir aos utilizadores livros que outros utilizadores, que têm livros lidos em comum, leram. Isto porque, se um determinado utilizador tem em comum livros lidos com outro utilizador é porque há probabilidade de esse utilizador gostar do mesmo género de livros.

Começamos por criar uma lista dos livros lidos pelos *Id* dos utilizadores. De seguida, através do nodo *Association Rule Learner* com uma confiança de 25%, conseguimos obter poucos livros para recomendar, devido à falta de dados no nosso *dataset*. Por fim, nesta fase, ordenamos de forma descendente os livros com melhor qualidade da regra de associação.

Através do nodo *Table View*, criamos a interface para receber os *inputs* dos utilizadores, sendo esta utilizada para inserir os livros que o utilizador leu.

Posto isto, comparamos os resultados obtidos pela nossa regra de associação com as escolhas dos utilizadores e colocamos a *true* ou *false*, consoante haja correspondência. Para terminar, realizamos um processo para que o utilizador observe as recomendações de forma ordenada (da melhor para a pior).

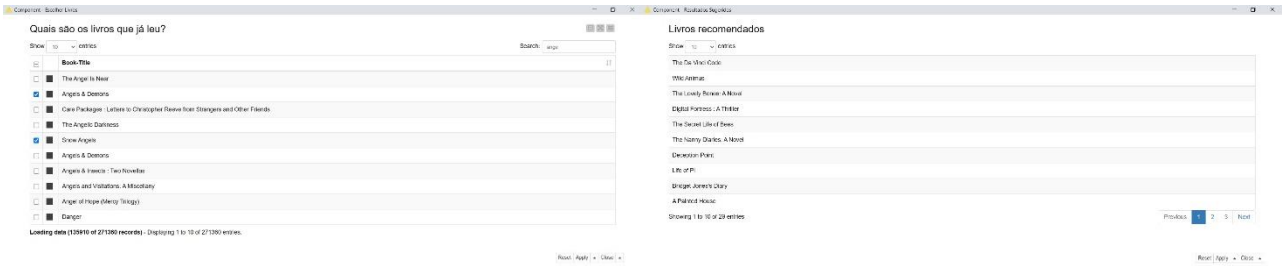


Figura 22 : Interface - Sistema de recomendação de livros através de regras de associação

No sistema de recomendação de autores através de regras de associação realizamos o mesmo processo que para os livros, contudo aplicamos uma confiança maior. Visto que, para esta sugestão o nosso *dataset* possui mais dados e por consequência conseguimos fornecer mais recomendações (confiança 45%).

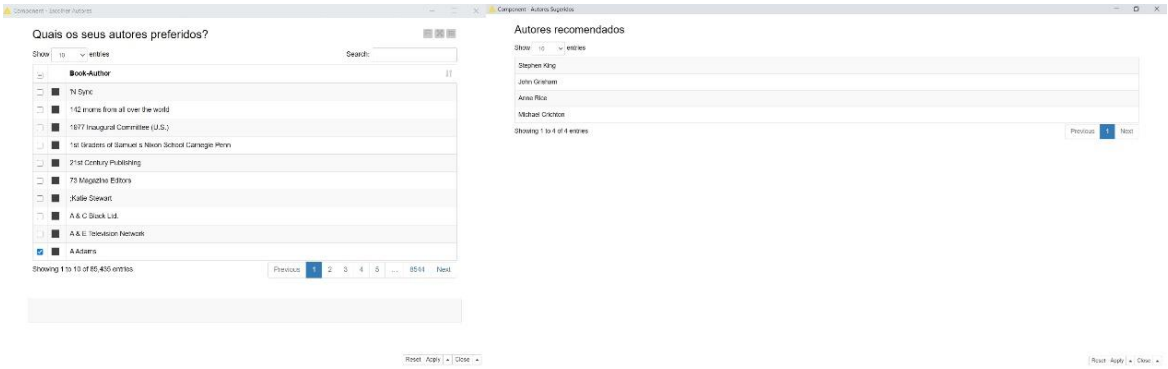


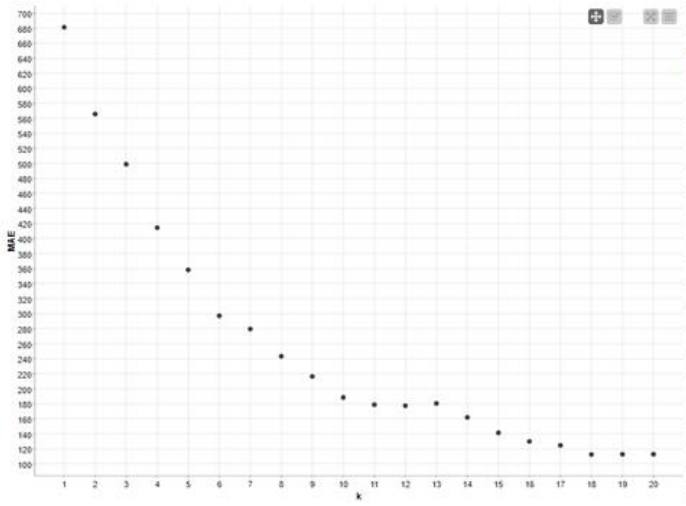
Figura 23 : Interface - Sistema de recomendação de autores através de regras de associação

Sistema de recomendação de livros e autores através da idade dos utilizadores (*clustering*)

Primeiramente, começamos por criar uma divisão das idades em quatro *bins* onde achamos que existe uma maior diferenção do tipo de leitura, por parte dos utilizadores.

- ✓ **[8-13] anos –** Livros lidos por crianças
- ✓ **[14-17] anos –** Livros lidos por adolescentes
- ✓ **[18-25] anos –** Livros lidos por adultos jovens
- ✓ **[26-95] anos –** Livros lidos por adultos

De seguida, adicionamos à nossa tabela inicial quatro colunas com os quatros *bins*, e, para cada utilizador, preenchemos com o valor 1 no *bin* em que está inserido e com o valor 0 nos restantes *bins*. Posteriormente, organizamos uma tabela com os livros, o respetivo *rating* e o número de utilizadores que leram aquele livro, por *bin* de idade. Antes de prosseguirmos para a próxima fase optamos por escolher o modelo *K-Means*, visto que, consideramos ser o melhor modelo para aplicar no nosso trabalho. De seguida, para conseguirmos obter um bom resultado no nosso modelo de *clustering* recorremos ao método do cotovelo e verificamos que o número de *clusters* que otimiza o nosso modelo é o número 2, tal como podemos reparar nas seguintes imagens :



Row ID	D MAE	I k	I Iteration	D MAE_Diff
Row0#0	681.253	1	0	0
Row0#1	565.746	2	1	115.506
Row0#2	498.942	3	2	66.804
Row0#3	414.397	4	3	84.545
Row0#4	358.294	5	4	56.103
Row0#5	297.162	6	5	61.133
Row0#6	279.524	7	6	17.638
Row0#7	243.311	8	7	36.213
Row0#8	216.429	9	8	26.882
Row0#9	188.423	10	9	28.006
Row0#10	178.832	11	10	9.592
Row0#11	177.467	12	11	1.365
Row0#12	180.478	13	12	0
Row0#13	161.918	14	13	18.56
Row0#14	141.42	15	14	20.498
Row0#15	129.892	16	15	11.528
Row0#16	124.682	17	16	5.21
Row0#17	112.543	18	17	12.139
Row0#18	112.857	19	18	0
Row0#19	112.857	20	19	0

Figura 24 : Método do cotovelo – K-Means

É de salientar que houve uma grande discrepância nos elementos constituintes de cada *cluster*, sendo que isto se deve ao dados do nosso *dataset*, como já foi referido anteriormente.

- ✓ ***Cluster* 1:** 96 elementos
- ✓ ***Cluster* 2:** 85512 elementos

Através do nodo *Autocomplete Text Widget*, criamos a interface para receber os *inputs* dos utilizadores, sendo esta utilizada para inserir o *bin* no qual a idade do

utilizador se insere.

De seguida, tendo em conta a idade introduzida pelo utilizador, comparamos quais livros foram lidos por essa idade, através do nodo *Rule Engine*. Previamente, de forma a obtermos melhores recomendações, definimos um *rating* mínimo para os livros a serem recomendados. Na interface dos *outputs* os livros são apresentados de forma decrescente de *rating*.

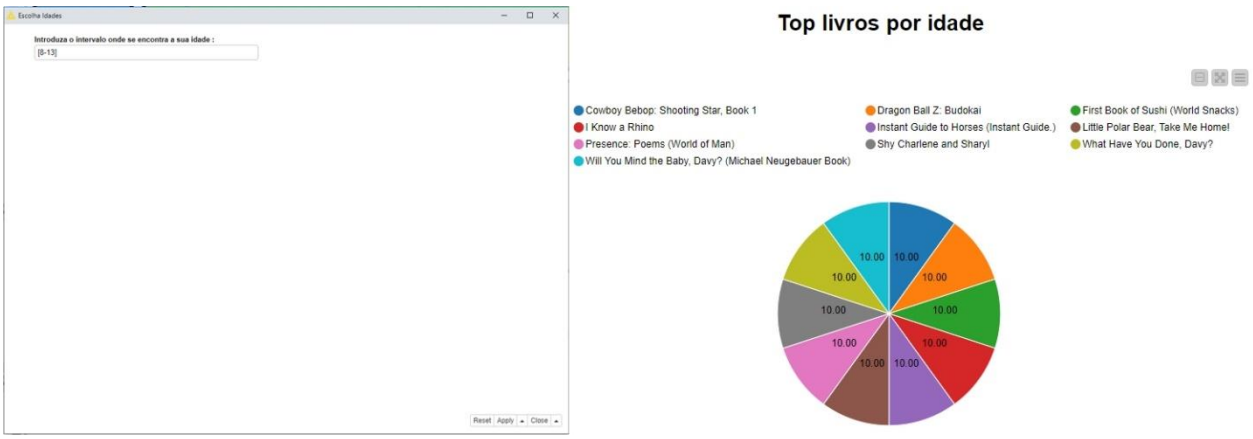


Figura 25 : Interface - Sistema de recomendação de livros através da idade dos utilizadores (clustering)

Foi implementado o mesmo processo de recomendação dos livros, tendo em conta a idade dos utilizadores, mas para a recomendação de autores.

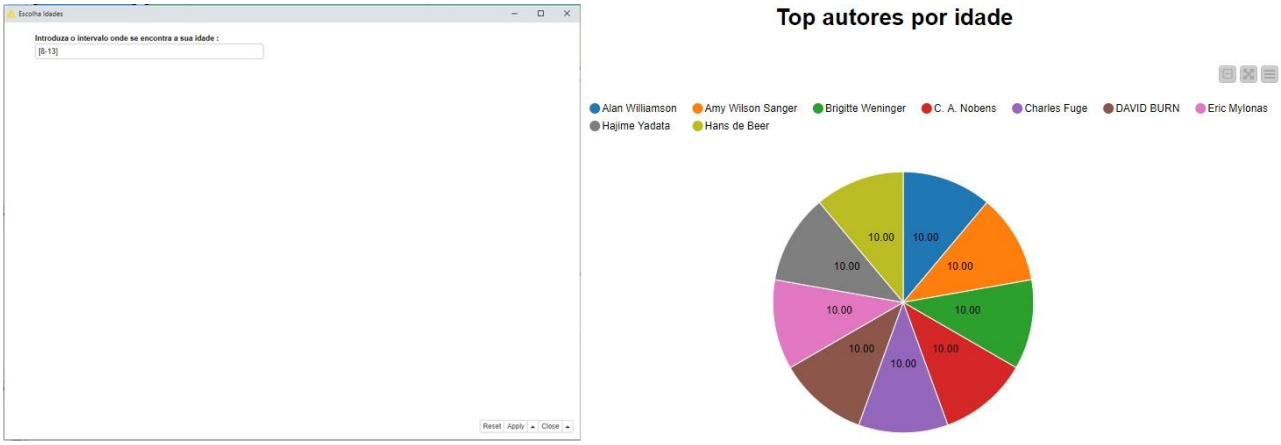
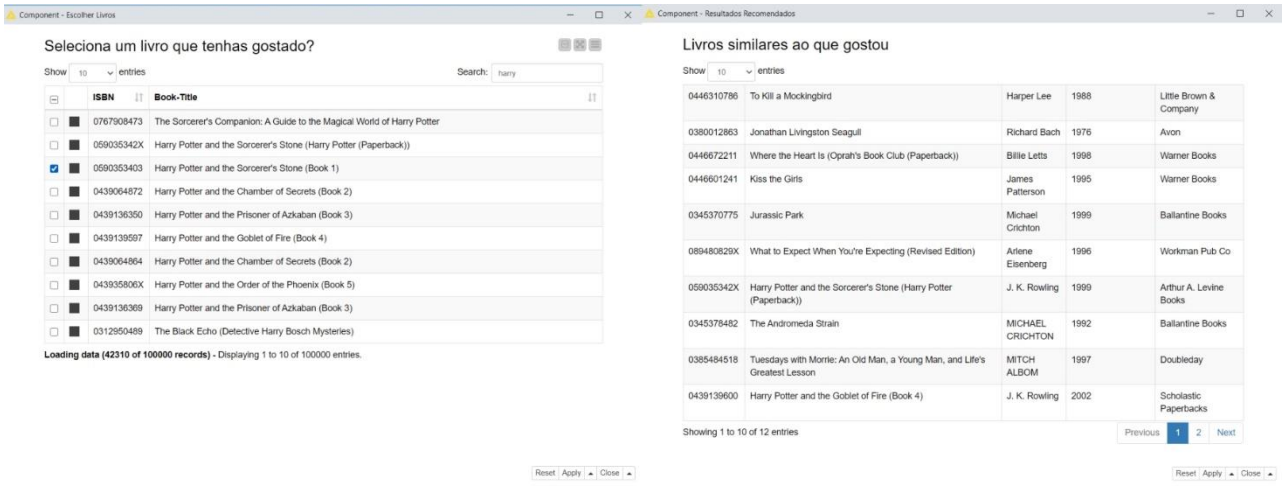


Figura 24 : Interface - Sistema de recomendação de autores através da idade dos utilizadores (clustering)

Sistema de recomendação de livros através de filtragem colaborativa baseada em memória (vizinho mais próximo baseado no utilizador)



7. Sumário dos resultados obtidos;

8. Apresentação de sugestões e recomendações após análise dos resultados obtidos e dos modelos desenvolvidos para melhoria do Sistema de Recomendação.