Universidade do Minho
Escola de Engenharia
Licenciatura em Engenharia Informática

Unidade Curricular de
Aprendizagem e Decisão Inteligentes

Ano Letivo de 2021/2022

Relatório
Conceção de modelos de aprendizagem

Grupo 28

Guilherme Gonçalves	a88280
Laura Rodrigues	a93169
Rita Gomes	a87960

6 de Maio de 2022

Índice

1 Introdução

O presente relatório serve de suporte ao trabalho realizado no âmbito da unidade curricular de Aprendizagem e Decisão Inteligentes.

Para a elaboração deste trabalho foi utilizada a ferramenta KNIME, que possui uma ampla coleção de funções de transformações de features e algoritmos de Machine Learning, além de oferecer ferramentas para uma visualização rápida e amigável dos resultados numa interface gráfica, tornando o entendimento de dados e o design dos fluxos de trabalho mais práticos, intuitivos e acessíveis.

O objetivo deste trabalho consiste em analisar, explorar e procurar extrair conhecimento de dois datasets. Um destes datasets foi fornecido pelos docentes da UC e a escolha do segundo dataset ficou ao cargo do grupo. Tendo isto em conta, o dataset que nos foi atribuído intitula-se de *wine quality classification* atendendo ao facto de o número do nosso grupo ser o 28. Para a escolha do segundo dataset, o grupo consultou várias fontes para encontrar um que fosse interessante. Após analisar vários datasets, o grupo optou por escolher um que fazia a análise dos 120 anos dos jogos olímpicos. Esta escolha baseou-se no facto de este ser um dataset que contém *missing values*, algo que não acontecia no dataset da qualidade do vinho, pelo que achamos por bem que seria relevante abordar o tratamento deste tipo de valores. Para além disso, observamos que o dataset tinha diferentes tipos de informação relevante para fazer uma análise completa dos dados.

2 Formalização dos problemas

2.1 Dataset *wine quality classification*

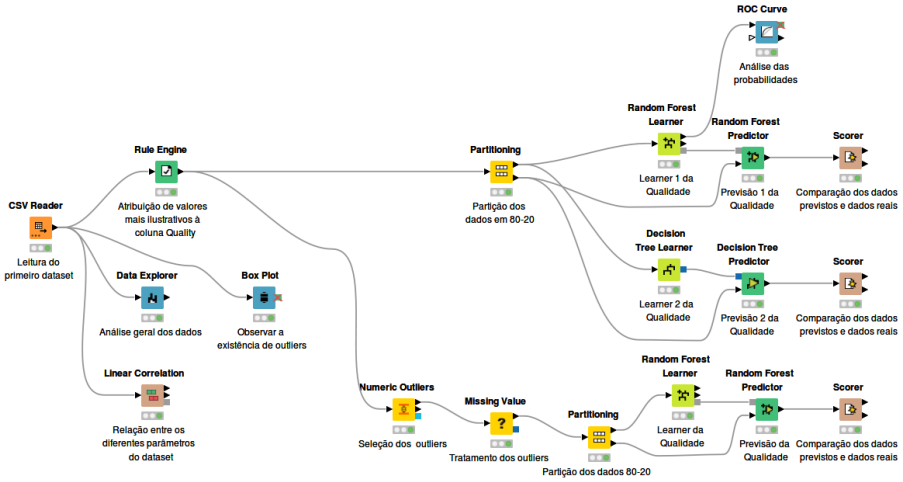


Figura 2.1: dataset:wine quality classification

2.1.1 Análise dos dados

A análise deste dataset tinha como objetivo determinar a qualidade de um vinho tinto produzido em Portugal. Ao correlacionar a qualidade com cada um dos parâmetros das outras colunas e observar quais destes têm influência nesta é possível prever o que permite obter um vinho de bom ou má qualidade.

Iniciámos a análise do problema carregando o ficheiro wine_quality_classification.csv para o worflow do Knime usando o novo CSV Reader. Após ter o dataset disponível passámos à sua análise.

- **Data Explorer:** Análise geral dos dados. Este nodo permitiu-nos compreender melhor o tipo de informação presente no dataset e assim trabalhá-lo objetivamente. Para além disso, graças a este temos outra perspetiva dos dados gerais, dado que obtivemos as médias, mínimos e máximos dos diferentes parâmetros. Por exemplo, a média da qualidade, mostra-nos que há mais vinhos maus no dataset do que bons, uma vez que 0.136 está mais próximo de 0 (mau).

Data Explorer View

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
fixed acidity		4.600	15.900	8.320	1.741	3.031	0.983
volatile acidity		0.120	1.580	0.528	0.179	0.032	0.672
citric acid		0	1	0.271	0.195	0.038	0.318
residual sugar		0.900	15.500	2.539	1.410	1.988	4.541
chlorides		0.012	0.611	0.087	0.047	0.002	5.680
free sulfur dioxide		1	72	15.875	10.460	109.415	1.251
total sulfur dioxide		6	289	46.468	32.895	1082.102	1.516
density		0.990	1.004	0.997	0.002	0.000	0.071
pH		2.740	4.010	3.311	0.154	0.024	0.194
sulphates		0.330	2	0.658	0.170	0.029	2.429
alcohol		8.400	14.900	10.423	1.066	1.136	0.861
quality		0	1	0.136	0.343	0.117	2.129

Figura 2.2: Interactive View: Data Explorer View

- **Linear Correlation:** Analisar a relação entre os diferentes parâmetros do dataset permitiu-nos perceber quais as características de um vinho que mais influenciam a qualidade deste.

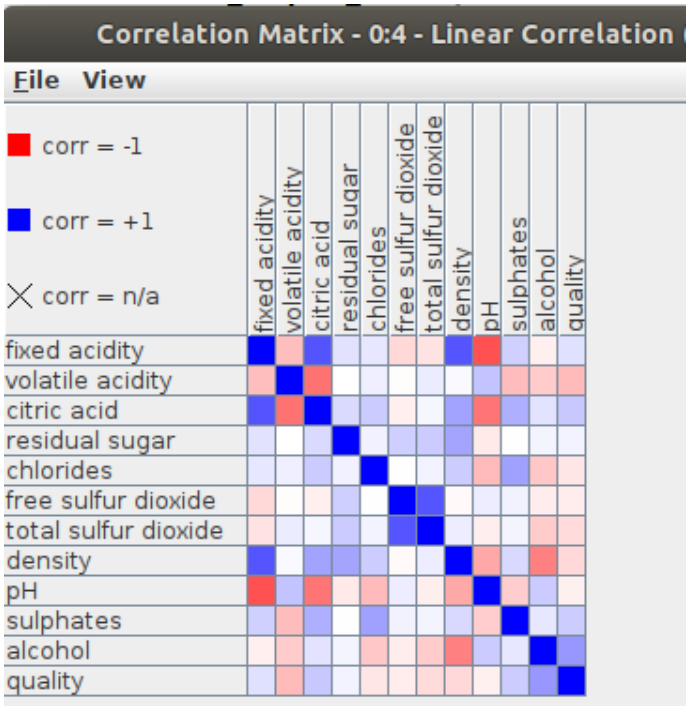


Figura 2.3: Correlation matrix

- **Box Plot:** Observar a existencia de outliers, de modo a saber se haveria ou não a necessidade de os tratar.

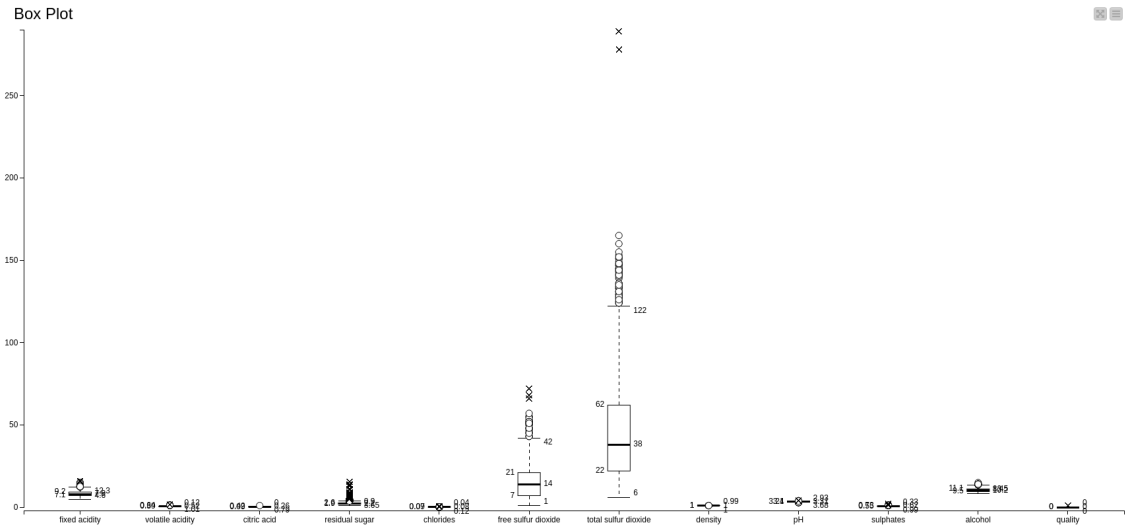


Figura 2.4: Interactive View: Box Plot

2.1.2 Tratamento e análise dos dados

- **Rule Engine:** Atribuição de valores mais ilustrativos à coluna Quality que facilitaram o uso de nodos de aprendizagem/previsão.

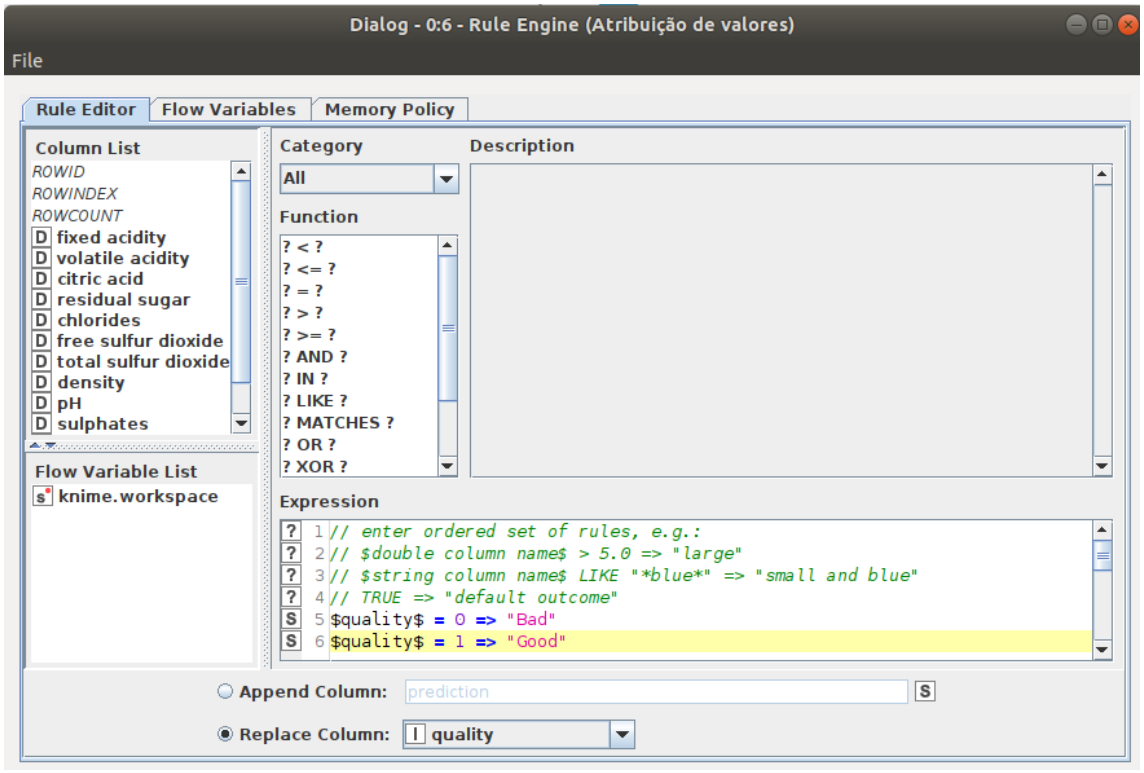


Figura 2.5: Rules Engine

De forma a perceber qual seria a melhor forma de analisar e fazer uma previsão da qualidade deste dataset experimentámos fazê-lo de três maneiras, fazendo sentido compará-las duas a duas.

Análises 1 e 2:

Optámos por uma maneira simples e, no entanto, eficiente de o fazer. Criámos uma partição de forma a podermos comparar os dados previstos com outros já existentes no dataset e assim fazer uma previsão adequada.

- **Partitioning:** Partição dos dados em 80-20.

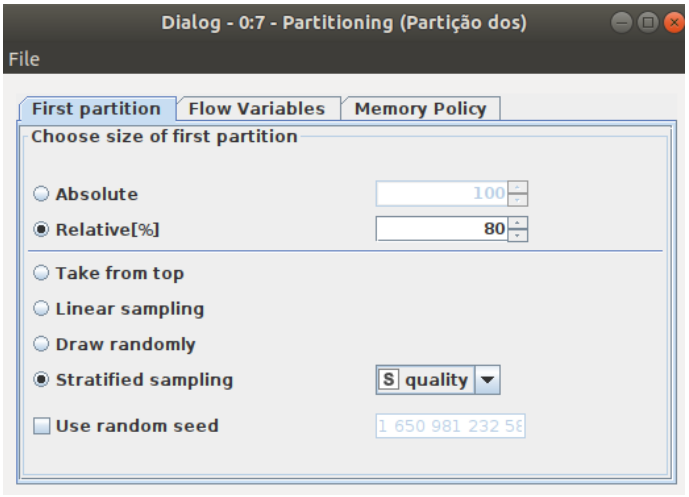


Figura 2.6: Partitioning

- **Learners e Predictors:** O uso de dois learners diferentes permitiu-nos descobrir qual dos dois métodos de aprendizagem usados seria capaz de melhor prever a qualidade dos vinhos com base nos parâmetros diferentes do dataset. Assim escolhemos os dois que melhor se adaptavam aos tipos de dados que temos: *Random Forest Learner* e *Decision Tree Learner*.

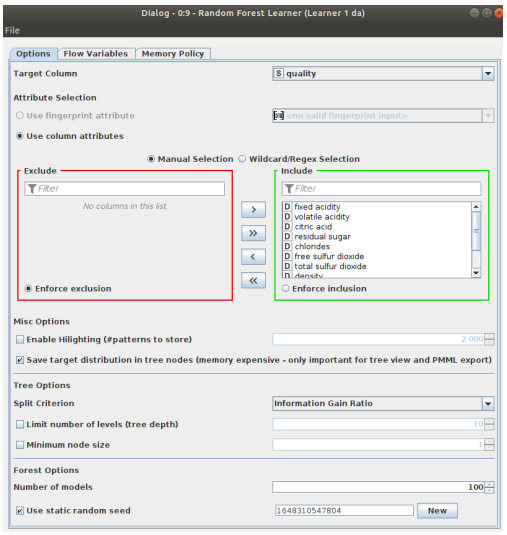


Figura 2.7: Random Forest Learner

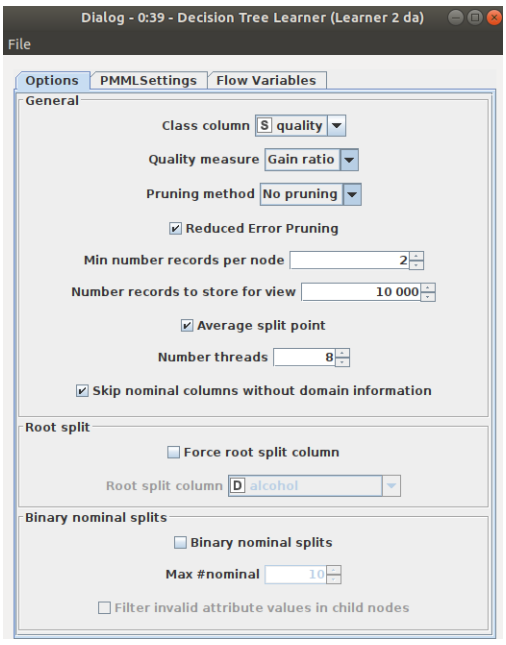


Figura 2.8: Decision Tree Learner

Em cada uma das análises usámos os respectivos Predictors: *Random Forest Predictor* e *Decision Tree Predictor*. Para cada um destes foram passadas como input duas

amostras: a que veio do learner e a outra partição criada anteriormente. Associado ao primeiro learner temos um gráfico **Roc Curve** que nos permitiu perceber quais eram o parâmetros que tinham mais peso num vinho de boa qualidade e quais os que, por sua vez, influenciariam a que um vinho fosse mau.

- **Roc Curve**: Tal como podemos observar o parâmetro que mais influencia um vinho bom é a quantidade de álcool, enquanto que um vinho mau é mais influenciado pela acidez deste.

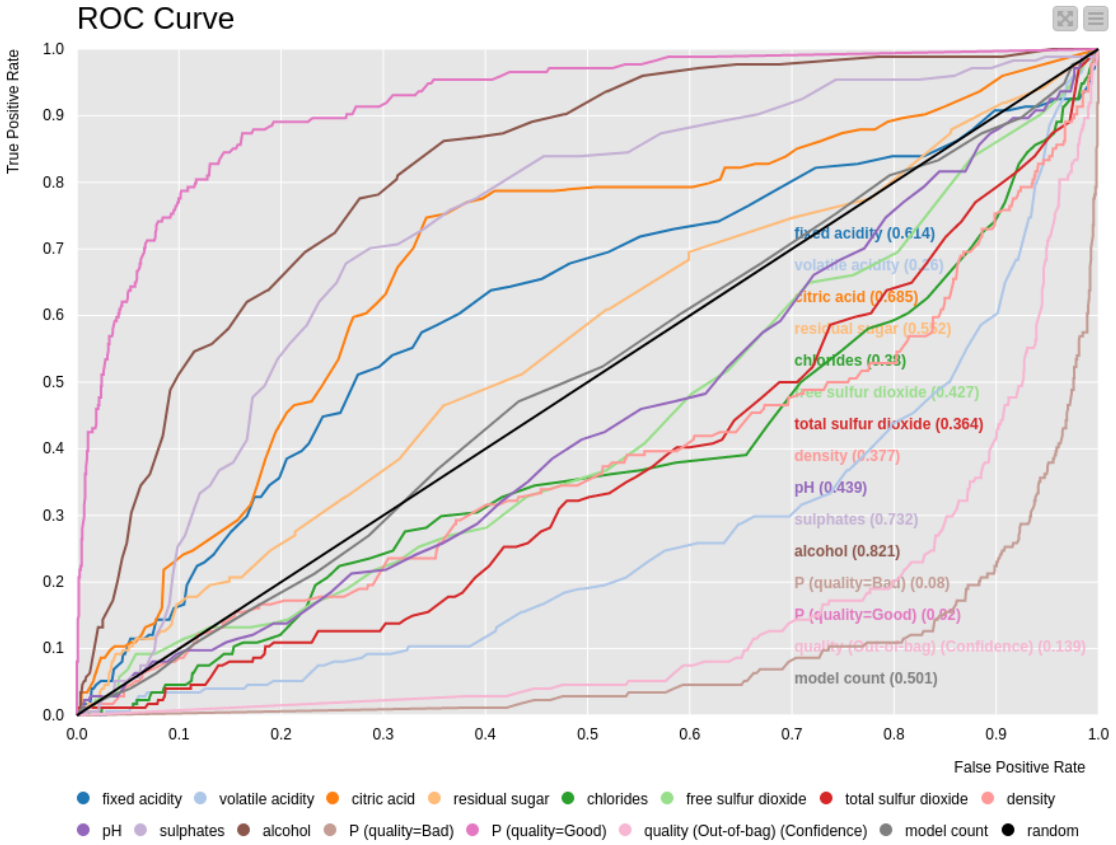


Figura 2.9: Resultado do nodo Roc Curve

- **Scorers**: Comparando os valores obtidos no final das duas análises, pudemos observar que, para este dataset, o método que obteve melhor resultados foi o de *Random Forest* (à esquerda). Concluímos, assim, que este era o melhor método de análise, uma vez que ambos passaram pelo mesmo tratamento de dados e tiveram a mesma semente de partição.

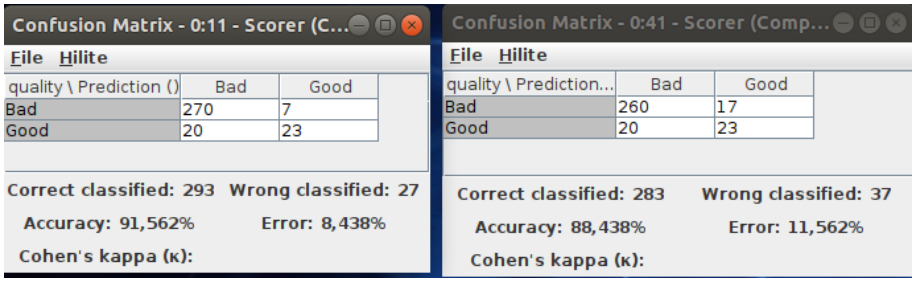


Figura 2.10: Matrizes de Confusão

Análises 1 e 3:

Após descobrir qual o melhor learner a usar, passámos ao tratamento dos **Outliers**. Tal como observado anteriormente no nodo Box Plot, existem outliers no nosso dataset, mas não sabíamos a influência destes no resultado final que estávamos a obter.

- **Numeric Outliers:** Começámos por selecionar estes valores e substituí-los por missing values, de forma a podermos depois tratá-los de forma adequada.

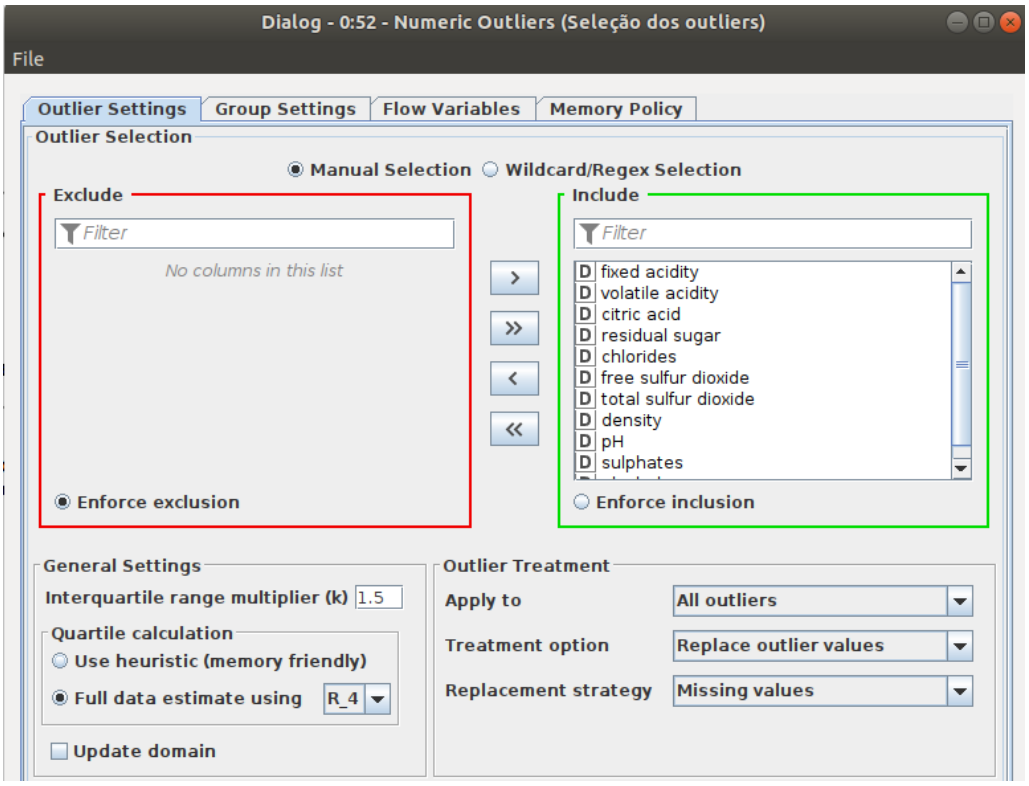


Figura 2.11: Numeric Outliers

- **Missing Values:** Tratamentos dos valores dos outiers, agora substituídos por missing values. Para cada um dos parâmetros dos vinhos, optámos por substituí-los pela respectiva média. Dado não existirem outliers na coluna da qualidade, o tratamento aplicados aos parâmetros nominais (que no caso é apenas Quality) é irrelevante.

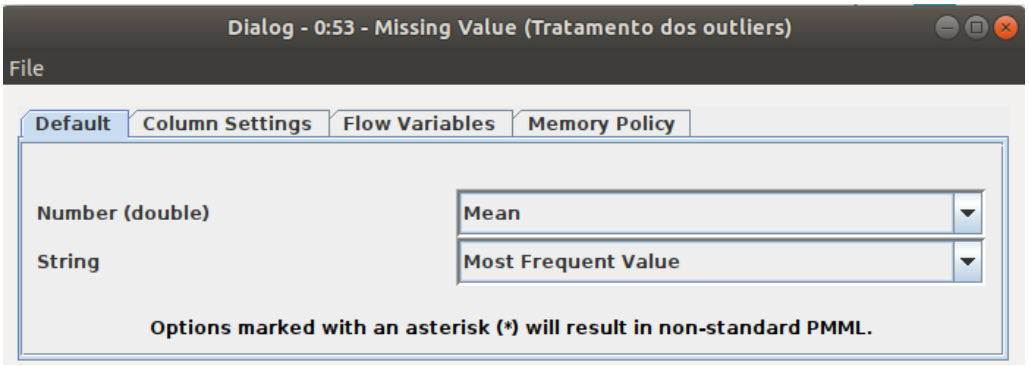


Figura 2.12: Missing Values

Após aplicar este tratamento de dados os nodos que se seguiram foram iguais à análise 1, de modo a garantir uma comparação justa e correta dos resultados, incluindo as mesmas Sementes usadas para gerar as partições e as usadas nos learners.

- **Scorers**: Comparando os valores obtidos no final das duas análises, podemos observar que, para este dataset, recorrer ao tratamento de Outliers prejudica a previsão do dado Quality (apesar da diferença ser muito pequena - 1,562%).

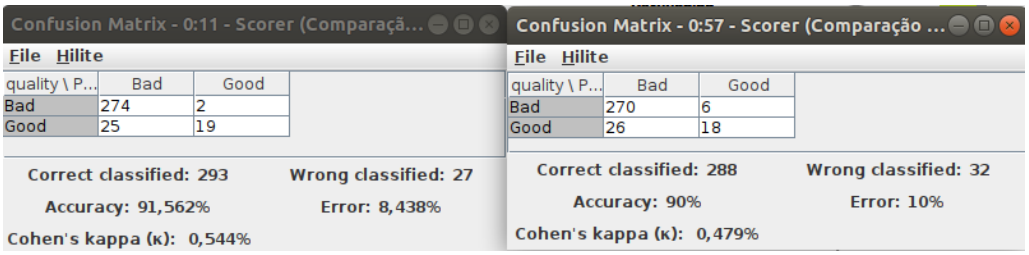


Figura 2.13: Missing Values

2.2 Dataset *120 years of Olympic history: athletes and results*

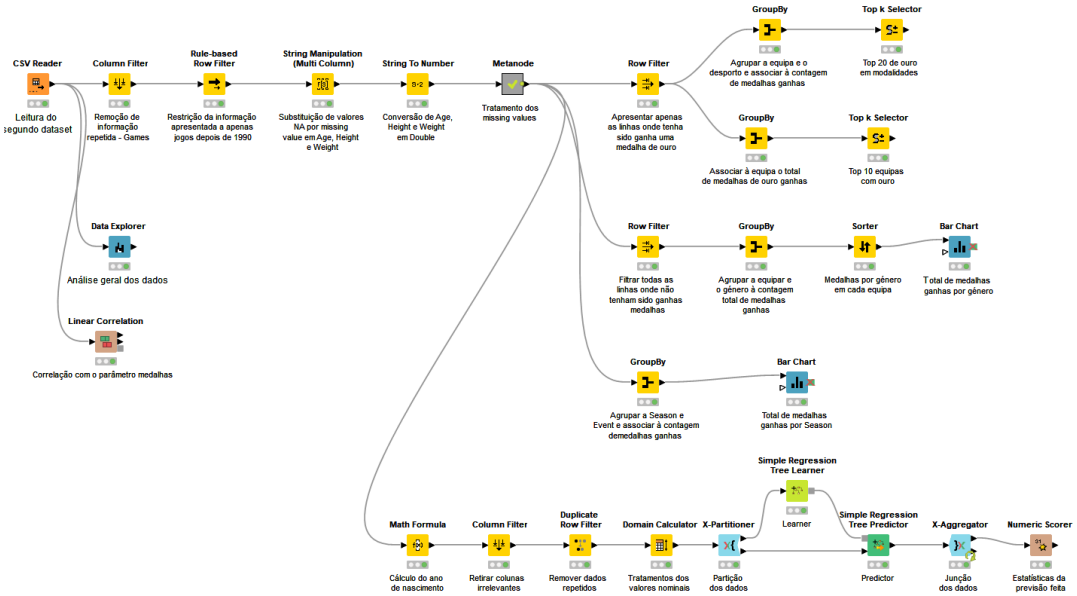


Figura 2.14: 120 years of Olympic history: athletes and results

Retirado de: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

2.2.1 Análise dos dados

A análise deste dataset passou por vários objetivos, alguns propostos pelo criador do dataset na secção *Inspiration* da página do *Kaggle* e outros por nós. Para além disso, tentámos completar alguns dos que foram propostos. É de notar que algumas das análises feitas não faziam 100% sentido, algo que será explicado mais aprofundadamente adiante.

- Analisar a participação e performance de mulheres ao longo dos anos (tendo sido aproveitado para analisar a discrepância de perfomance entre homens e mulheres).
- Analisar a participação e performance de diferentes nações ao longo dos anos para o qual optámos por selecionar as 10 melhores equipas, com base no maior número de medalhas de ouro.
- Analisar a participação e performance em diferentes estações (jogos de Inverno e Verão).

- Analisar a participação e performance em diferentes desportos para a qual foram selecionas as 20 modalidades e respetivos países com maior número de medalhas de ouro.

- prever as alturas com base em diferentes parâmetros relevantes

Uma vez definidos os objetivos, procedemos assim à análise do dataset:

- **Data Explorer:** Análise geral dos dados. Este nodo permitiu-nos compreender melhor o tipo de informação presente no dataset e assim trabalhá-lo objetivamente. Para além disso, graças a este temos outra perspetiva dos dados gerais, dado que obtivemos as média, mínimo e máximo do parâmetro numérico Year (apesar do dataset ser referente aos anos 1896 - 2016, a maior parte dos eventos registados aconteceram a partir do ano 1990.

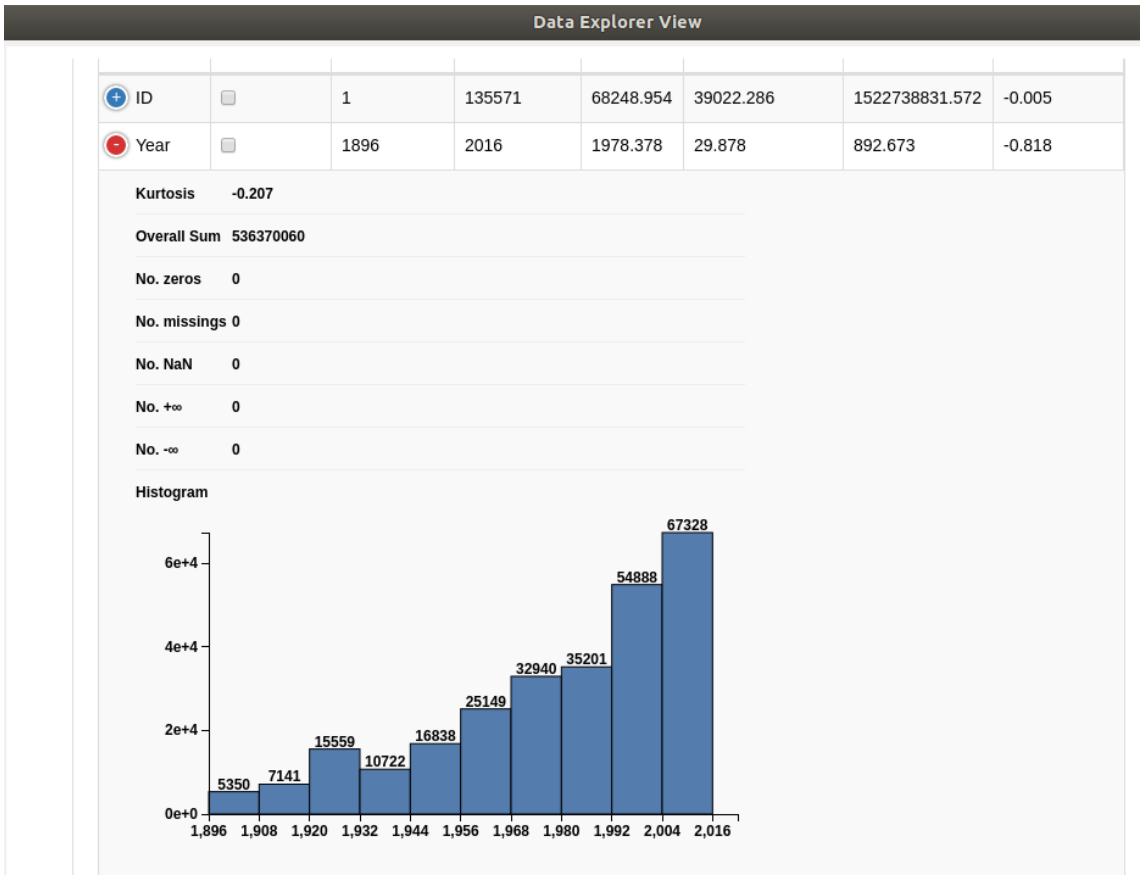


Figura 2.15: Interactive View: Data Explorer View - Numeric

Por outro lado, relativamente aos dados nominais, pudemos observar a grande dimensão de equipas, países e modalidades, bem como perceber em que parâmetros havia a presenção de missing values, representados neste dataset por *NA*.

Figura 2.16: Interactive View: Data Explorer View - Nominal

- **Linear Correlation:** Analisar a relação entre os diferentes parâmetros do data-set permitiu-nos perceber que, infelizmente, não havia grande dependência entre dados.

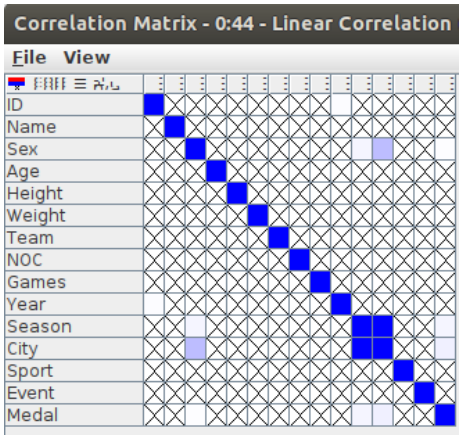


Figura 2.17: Correlation matrix

2.2.2 Tratamento dos dados

Finda toda a análise que achámos necessária passámos ao tratamento dos dados.

- **Column Filter:** Remoção de informação repetida uma vez que a coluna *Games* era equivalente às colunas *Year* e *Season*.

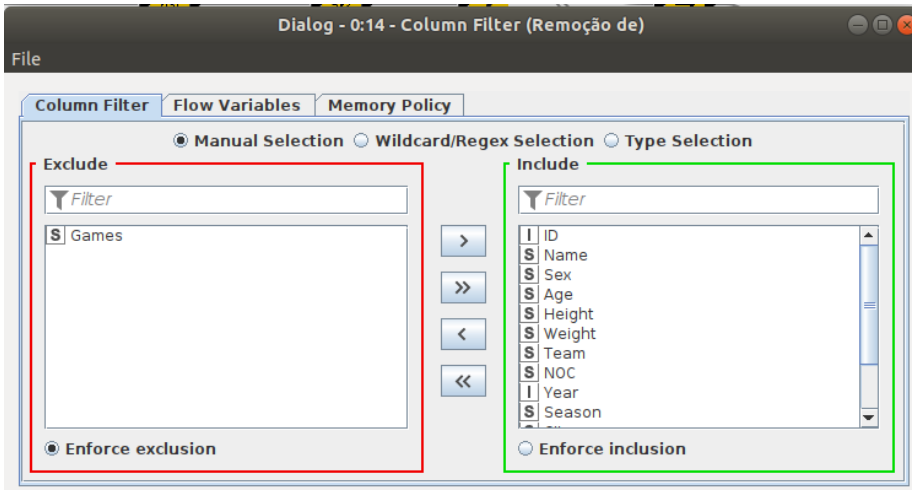


Figura 2.18: Column Filter

- **Rule based Row Filter:** Dada a extensão do dataset e uma vez que a maioria da informação se concentrava a partir dos anos 90, optámos por restringir a informação apresentada a apenas jogos depois de 1990.

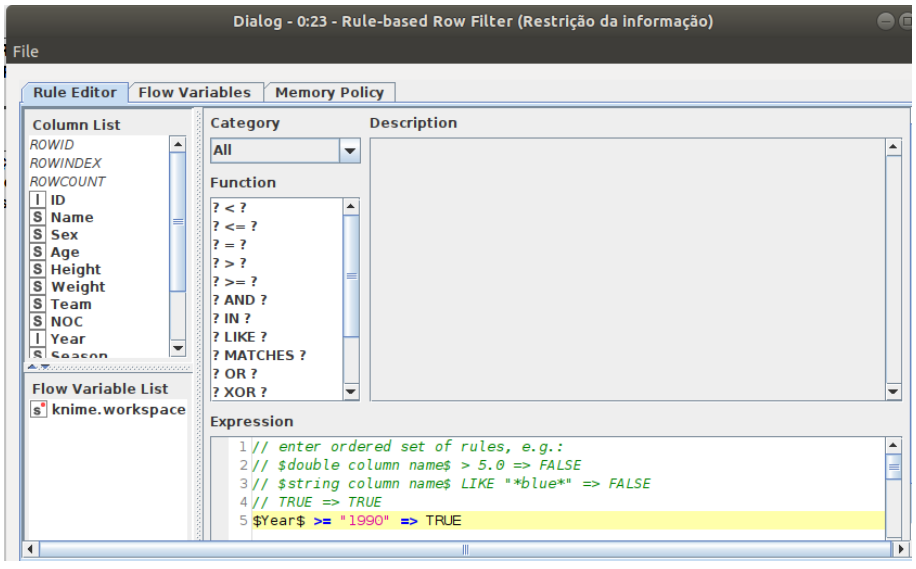


Figura 2.19: Rule based Row Filter

- **String Manipulation (Multicolumn):** Dado termos observado a presença de missing values no dataset, houve a necessidade de os tratar. No entanto, apenas os dos parâmetros Age, Height e Weight foram interpretados como tal, dado que os NA presentes na coluna *Medal* seriam sinónimo a não ter ganho medalhas (informação essa que era bastante importante). Assim, substituímos os valores NA por missing value em de modo a facilitar a sua seleção e tratamento. Para isto foi usada a expressão:

  $$CURRENTCOLUMN$$.equals("NA") ? null:$$CURRENTCOLUMN$$

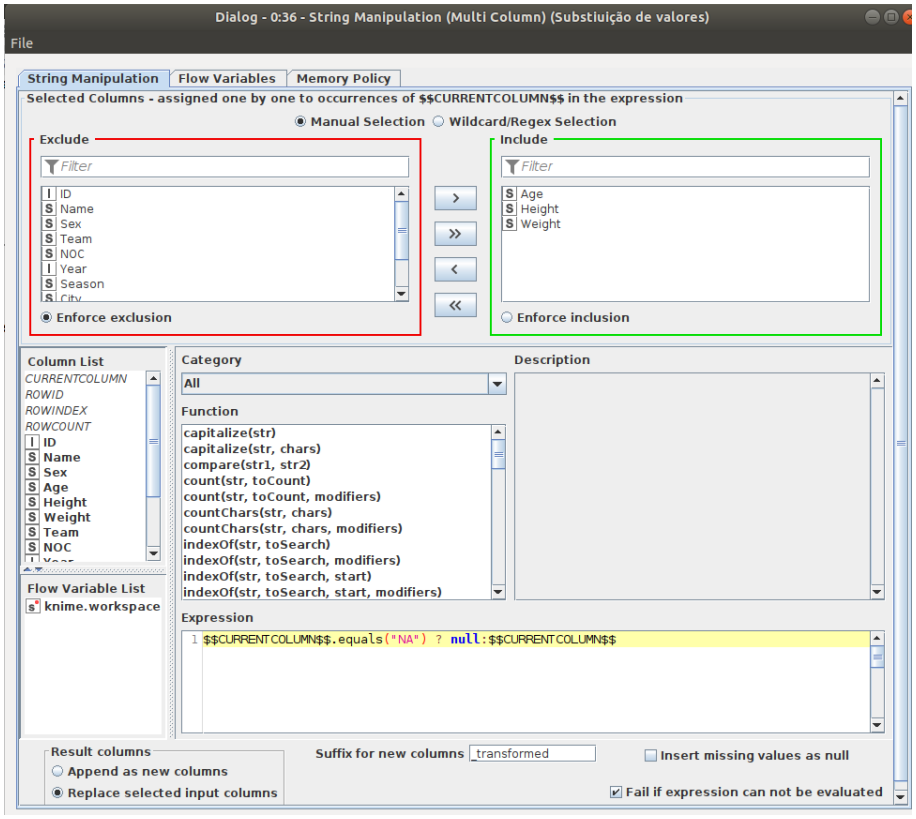


Figura 2.20: String Manipulation (Multicolumn)

- **String to Number:** Antes de passar para o tratamento dos missing values e de modo a facilitar esta ação (uma vez que gostaríamos de os tratar como valores numéricos), recorremos à conversão de Age, Height e Weight em Double.

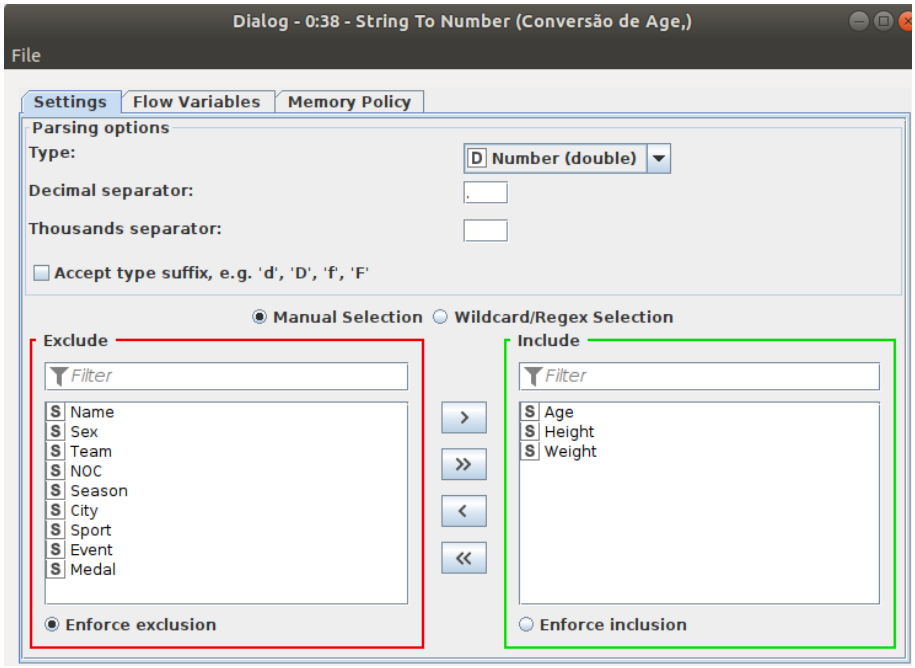


Figura 2.21: String to Number

- **Missing Value:** Finalmente, procedemos ao tratamento dos missing values. Apesar das várias operações possíveis, a que fazia mais sentido tendo em conta os dados em causa foi a de Interpolação Linear em que foi feita a interpolação entre os valores anterior e seguinte. Para tal fazer sentido optámos por ordenar cada um dos parâmetros antes de aplicar a respetiva transformação. No fim reestabelecemos a ordem do dataset original, reordenando-o pelo id. Por fim agrupámos estes nodos num metanodo de modo a facilitar a visualização do processo.

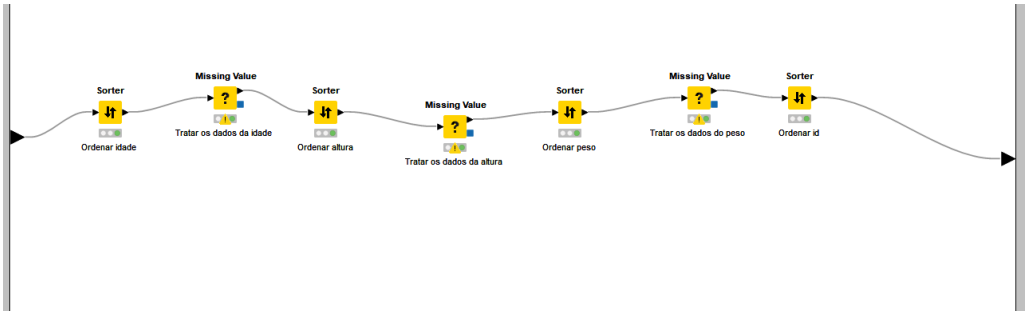


Figura 2.22: Missing Value

Análise 1: Top 20 em modalidades

- **Row Filter:** Apresentar apenas as linhas onde tenha sido ganha uma medalha de ouro de modo a facilitar a análise para o objetivo pretendido.

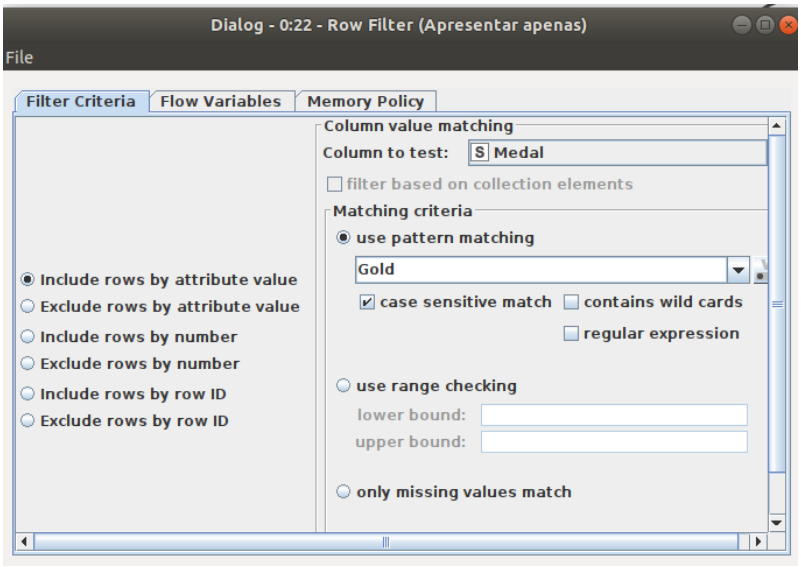


Figura 2.23: Row Filter - Gold

- **GroupBy:** Agrupar a equipa e o desporto e associar à contagem de medalhas ganhas.

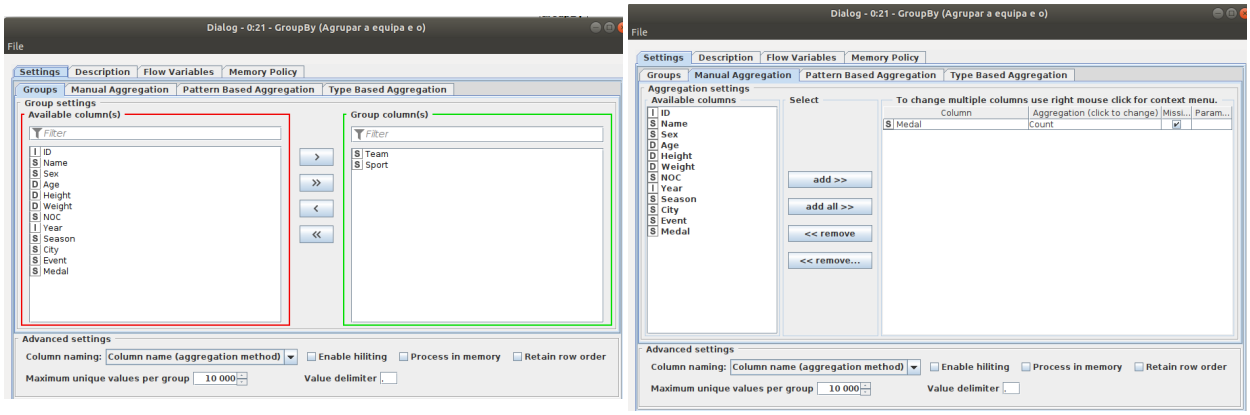
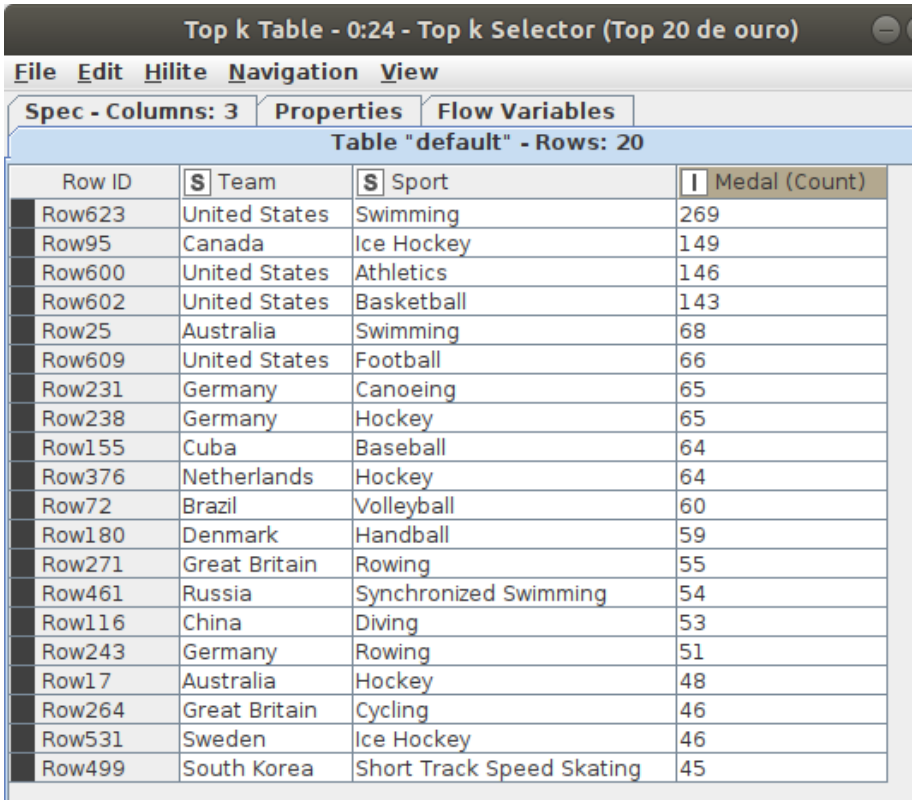


Figura 2.24: GroupBy

- **Top k Selector:** Participação e performance em diferentes desportos dos 20 melhores países (com maior número de medalhas de ouro).



Row ID	S Team	S Sport	I Medal (Count)
Row623	United States	Swimming	269
Row95	Canada	Ice Hockey	149
Row600	United States	Athletics	146
Row602	United States	Basketball	143
Row25	Australia	Swimming	68
Row609	United States	Football	66
Row231	Germany	Canoeing	65
Row238	Germany	Hockey	65
Row155	Cuba	Baseball	64
Row376	Netherlands	Hockey	64
Row72	Brazil	Volleyball	60
Row180	Denmark	Handball	59
Row271	Great Britain	Rowing	55
Row461	Russia	Synchronized Swimming	54
Row116	China	Diving	53
Row243	Germany	Rowing	51
Row17	Australia	Hockey	48
Row264	Great Britain	Cycling	46
Row531	Sweden	Ice Hockey	46
Row499	South Korea	Short Track Speed Skating	45

Figura 2.25: Top k Selector

Análise 2: Top 10 equipas

A semelhança da análise 1 foi feito um Row Filter de maneira a selecionar as linhas em que tinha sido ganha uma medalha de ouro (figura 2.23).

- **GroupBy:** Associar à equipa o total de medalhas de ouro ganhas.
- **Top k Selector:** Top 10 de equipas que ganharam mais medalhas de ouro.

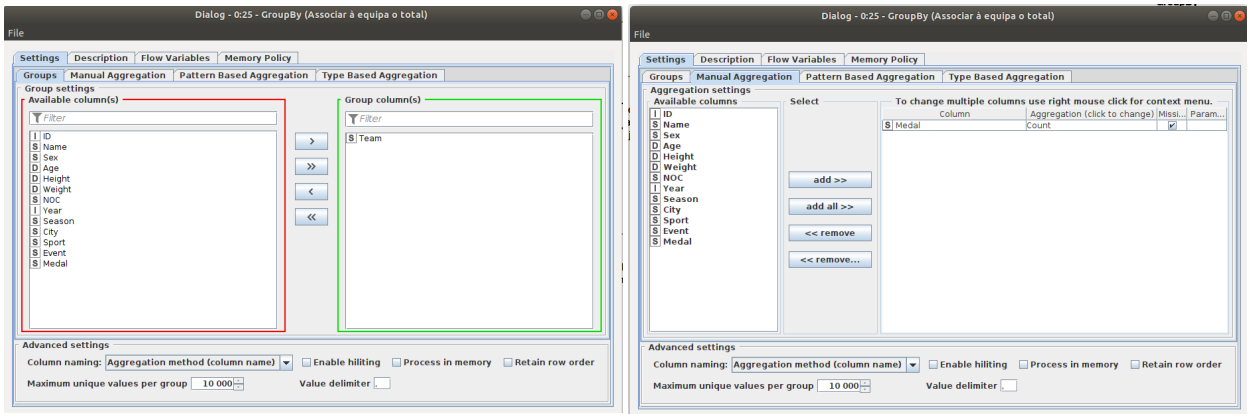


Figura 2.26: GroupBy

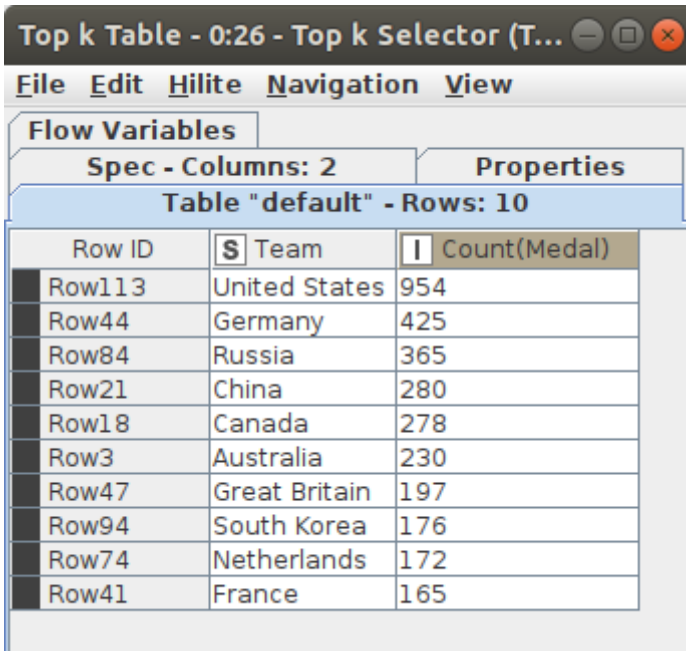


Figura 2.27: Top k Selector

Análise 3: Total de medalhas ganhas por género

- **Row Filter:** Filtrar todas as linhas onde não tenham sido ganhas medalhas de modo a facilitar a visualização de todas as entradas em que tenha sido ganha alguma medalha (Bronze, Prata ou Ouro).

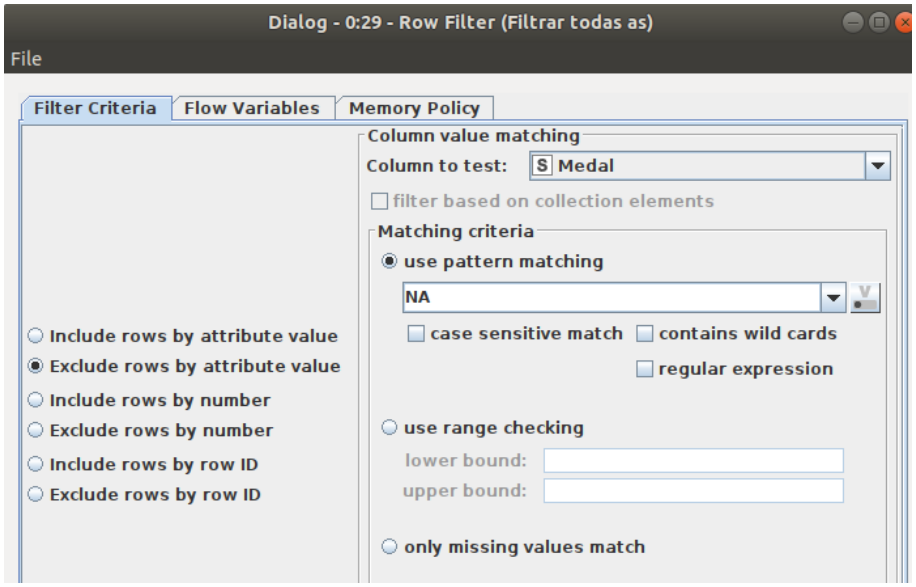


Figura 2.28: Row Filter

- **GroupBy:** Agrupar a equipar e o género à contagem total de medalhas ganhas.

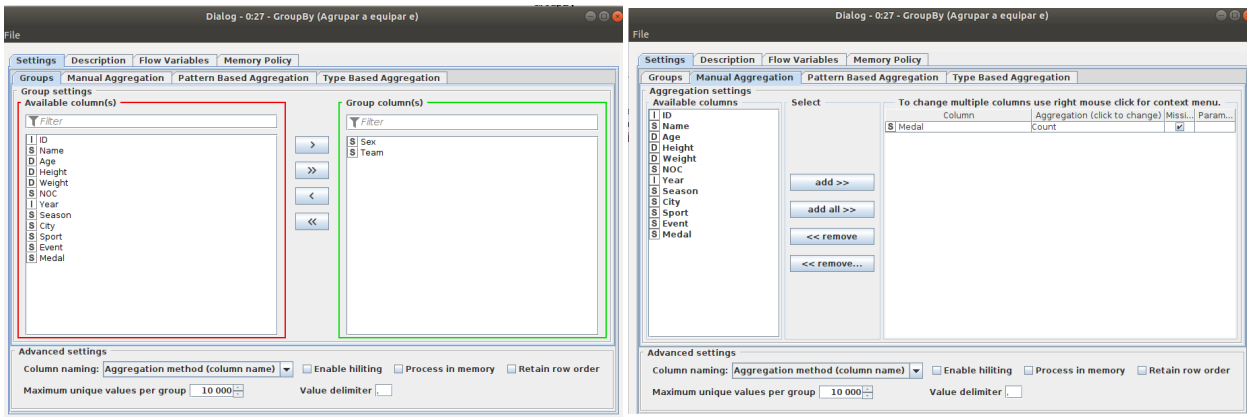
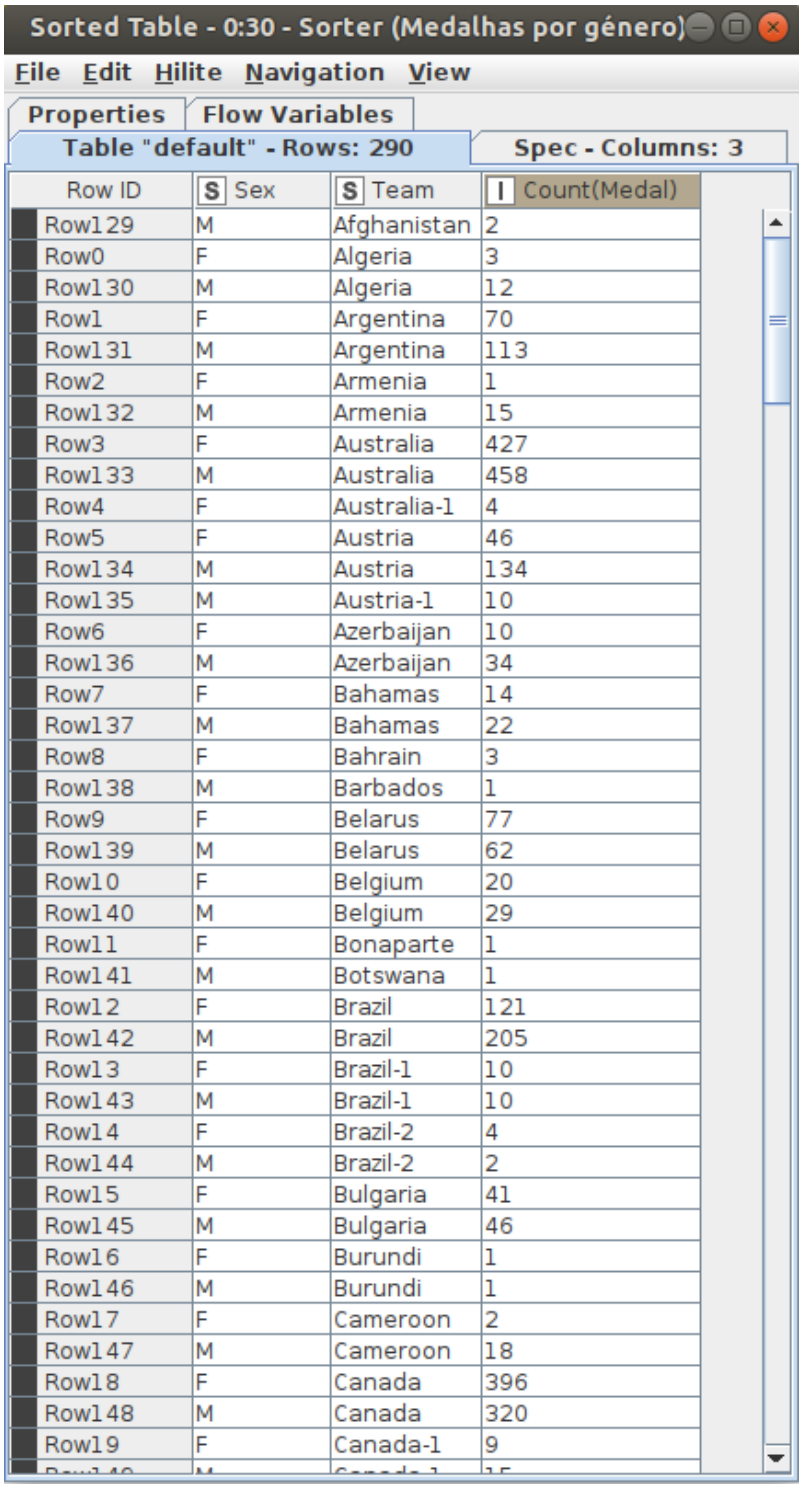


Figura 2.29: GroupBy

- **Sorter:** Visualização das medalhas ganhas por género em cada equipa de forma ordenada.



Row ID	S Sex	S Team	I Count(Medal)
Row129	M	Afghanistan	2
Row0	F	Algeria	3
Row130	M	Algeria	12
Row1	F	Argentina	70
Row131	M	Argentina	113
Row2	F	Armenia	1
Row132	M	Armenia	15
Row3	F	Australia	427
Row133	M	Australia	458
Row4	F	Australia-1	4
Row5	F	Austria	46
Row134	M	Austria	134
Row135	M	Austria-1	10
Row6	F	Azerbaijan	10
Row136	M	Azerbaijan	34
Row7	F	Bahamas	14
Row137	M	Bahamas	22
Row8	F	Bahrain	3
Row138	M	Barbados	1
Row9	F	Belarus	77
Row139	M	Belarus	62
Row10	F	Belgium	20
Row140	M	Belgium	29
Row11	F	Bonaparte	1
Row141	M	Botswana	1
Row12	F	Brazil	121
Row142	M	Brazil	205
Row13	F	Brazil-1	10
Row143	M	Brazil-1	10
Row14	F	Brazil-2	4
Row144	M	Brazil-2	2
Row15	F	Bulgaria	41
Row145	M	Bulgaria	46
Row16	F	Burundi	1
Row146	M	Burundi	1
Row17	F	Cameroon	2
Row147	M	Cameroon	18
Row18	F	Canada	396
Row148	M	Canada	320
Row19	F	Canada-1	9

Figura 2.30: Sorter

- **Bar Chart:** Visualização das medalhas ganhas por género na totalidade.

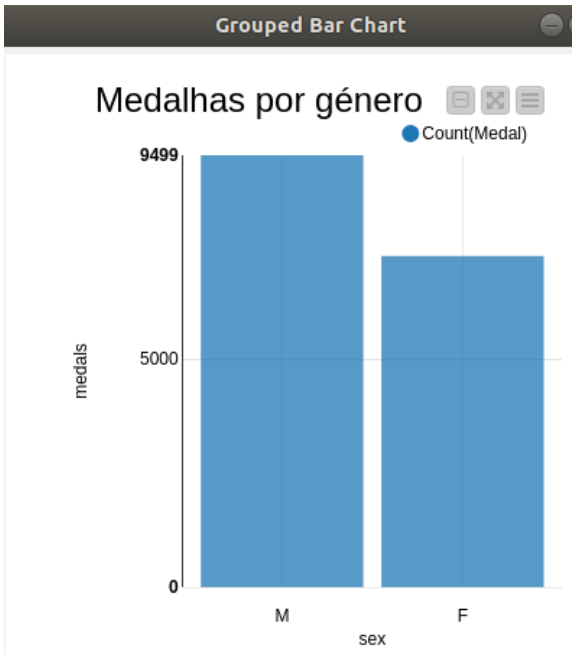


Figura 2.31: Bar Chart

Análise 4: Total de medalhas ganhas por Season

- **Groupby:** Agrupar a Season e Event e associar à contagem de medalhas ganhas.

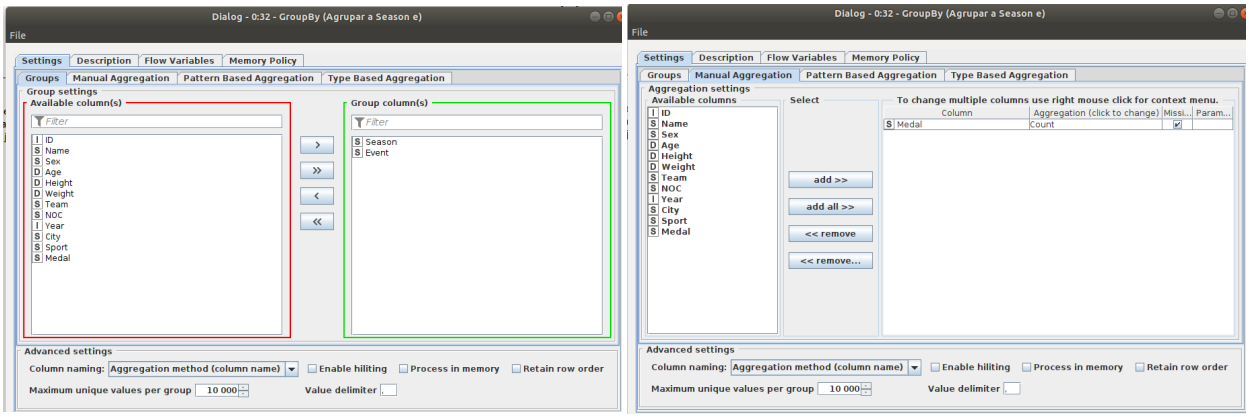


Figura 2.32: GroupBy

- **Bar Chart:** Visualização das medalhas ganhas por season na qual se selecionou a soma de todas as medalhas ganhas numa season.

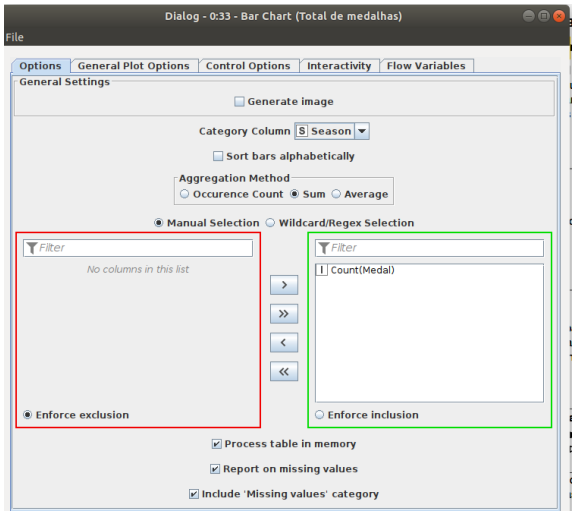


Figura 2.33: Configurações

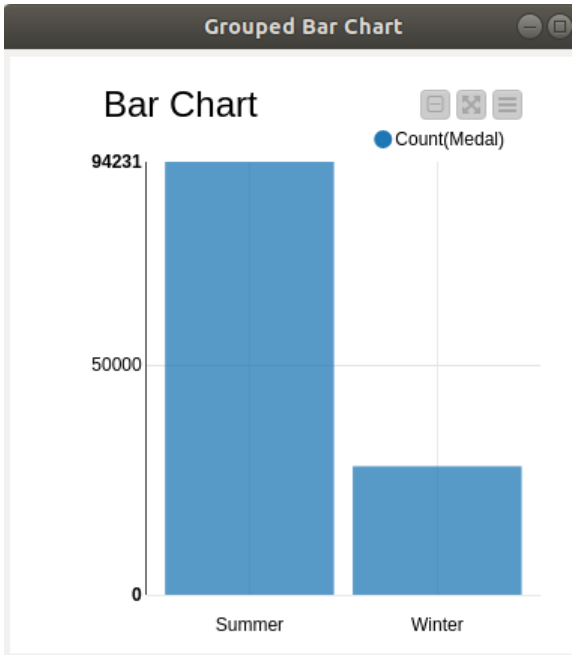
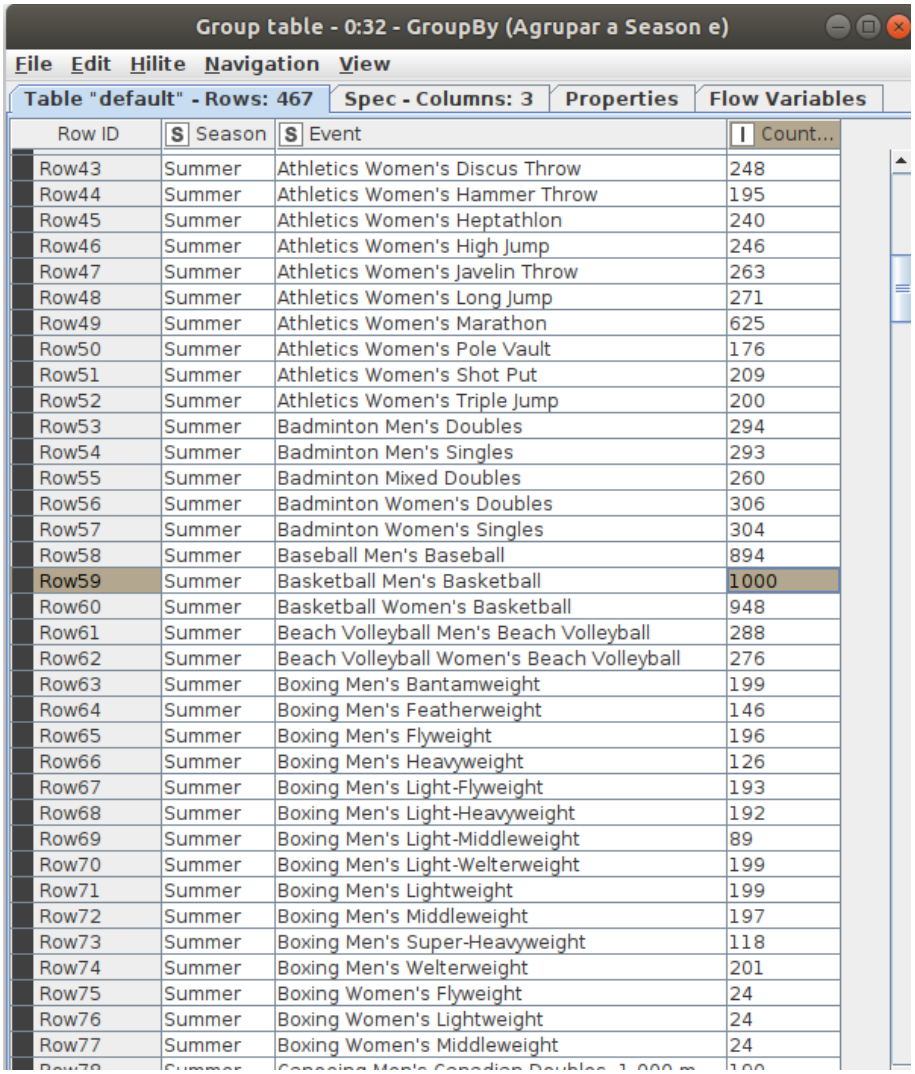


Figura 2.34: Resultados finais

Apesar destas análises terem sido propostas achamos que não faça 100% sentido comparar alguns destes valores. Por exemplo, tal como podemos observar na tabela gerada na análise 4, estamos a comparar o total de medalhas ganhas em desportos individuais

e de equipa. No entando, não é considerada uma medalha por equipa mas sim por cada jogador desta.



Row ID	Season	Event	Count...
Row43	Summer	Athletics Women's Discus Throw	248
Row44	Summer	Athletics Women's Hammer Throw	195
Row45	Summer	Athletics Women's Heptathlon	240
Row46	Summer	Athletics Women's High Jump	246
Row47	Summer	Athletics Women's Javelin Throw	263
Row48	Summer	Athletics Women's Long Jump	271
Row49	Summer	Athletics Women's Marathon	625
Row50	Summer	Athletics Women's Pole Vault	176
Row51	Summer	Athletics Women's Shot Put	209
Row52	Summer	Athletics Women's Triple Jump	200
Row53	Summer	Badminton Men's Doubles	294
Row54	Summer	Badminton Men's Singles	293
Row55	Summer	Badminton Mixed Doubles	260
Row56	Summer	Badminton Women's Doubles	306
Row57	Summer	Badminton Women's Singles	304
Row58	Summer	Baseball Men's Baseball	894
Row59	Summer	Basketball Men's Basketball	1000
Row60	Summer	Basketball Women's Basketball	948
Row61	Summer	Beach Volleyball Men's Beach Volleyball	288
Row62	Summer	Beach Volleyball Women's Beach Volleyball	276
Row63	Summer	Boxing Men's Bantamweight	199
Row64	Summer	Boxing Men's Featherweight	146
Row65	Summer	Boxing Men's Flyweight	196
Row66	Summer	Boxing Men's Heavyweight	126
Row67	Summer	Boxing Men's Light-Flyweight	193
Row68	Summer	Boxing Men's Light-Heavyweight	192
Row69	Summer	Boxing Men's Light-Middleweight	89
Row70	Summer	Boxing Men's Light-Welterweight	199
Row71	Summer	Boxing Men's Lightweight	199
Row72	Summer	Boxing Men's Middleweight	197
Row73	Summer	Boxing Men's Super-Heavyweight	118
Row74	Summer	Boxing Men's Welterweight	201
Row75	Summer	Boxing Women's Flyweight	24
Row76	Summer	Boxing Women's Lightweight	24
Row77	Summer	Boxing Women's Middleweight	24

Figura 2.35: Medalhas por Season

Análise 5: Previsão de alturas

- **Math Formula:** Cálculo do ano de nascimento de maneira a haver uma constante entre diferentes jogadores. Denomina-se *YearOfBirth*.

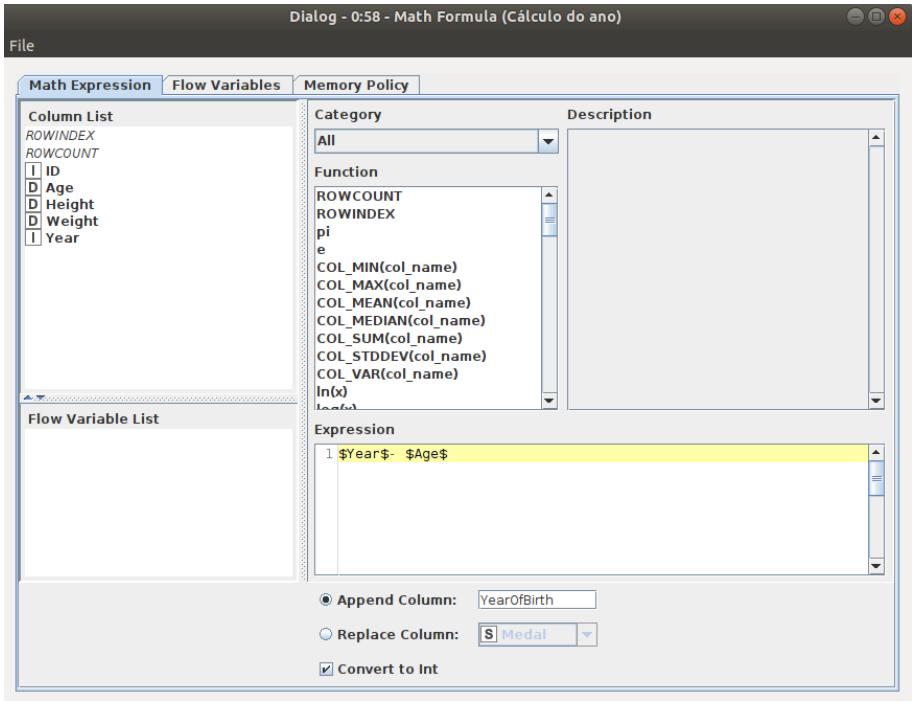


Figura 2.36: Math Formula

- **Column Filter:** Limitar às colunas que fazem sentido para prever a altura. O objetivo inicial seria usar um metanodo *Forward Feature Selection* para prever as colunas mais adequadas, no entanto houve alguma dificuldade em configurar este metanodo dado recorrer a um learner que necessitava de processar parâmetros nominais. Assim, optámos por selecionar as colunas que nos pareceram influênciar esta previsão.

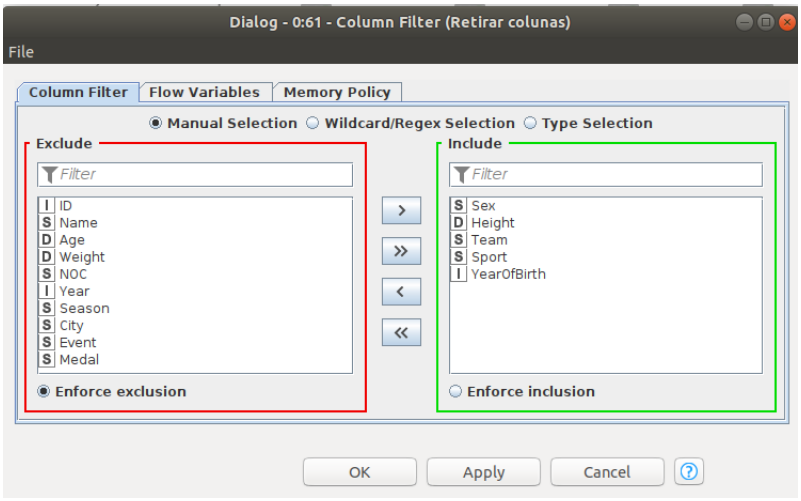


Figura 2.37: Column Filter

- **Duplicate Row Filter:** Remover dados repetidos de modo a limpar os dados desnecessários e assim não influenciar de forma negativa a previsão.

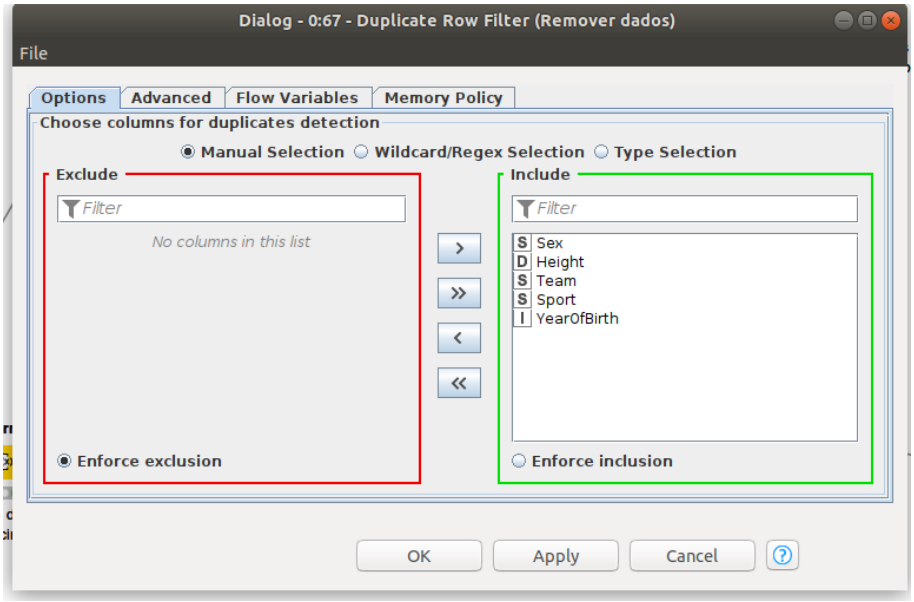


Figura 2.38: Duplicate Row Filter

- **Domain Calculator:** Tratamentos dos valores nominais de modo a poder usar o learner escolhido.

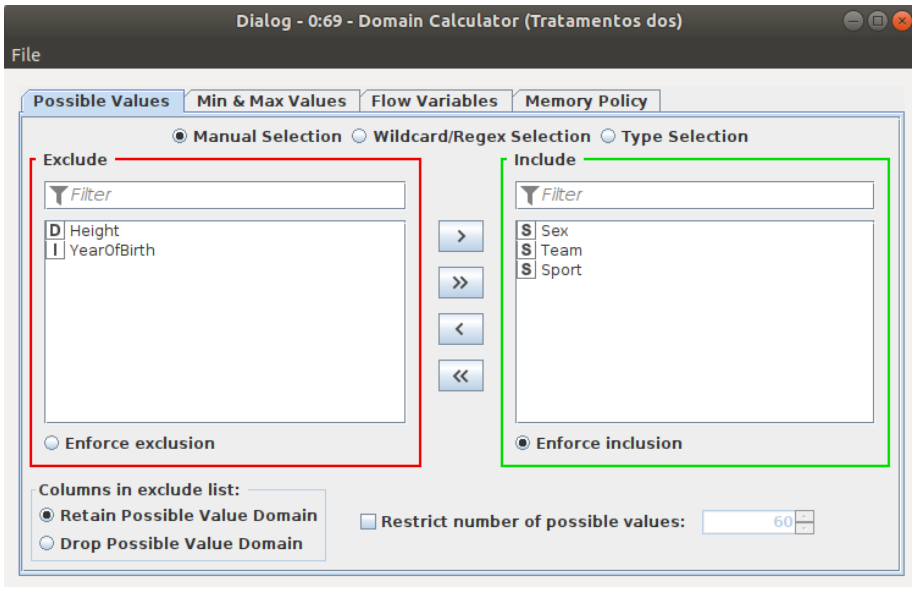


Figura 2.39: Domain Calculator

- **X-Partitioner:** Partição dos dados usando uma semente constante escolhida pelo grupo.

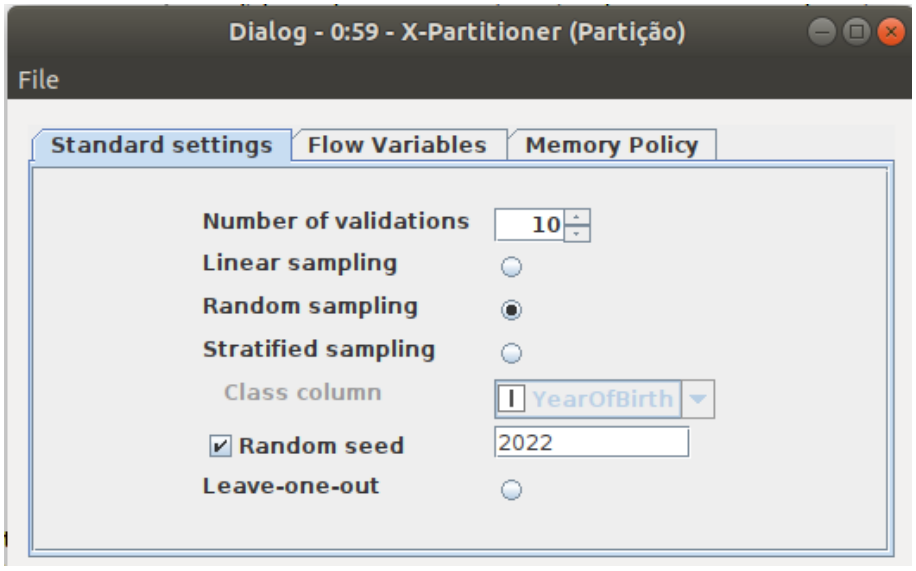


Figura 2.40: X-Partitioner

- **Learner e Predictor:** Optámos pelos nodos *Simple Regression Tree Learner* e *Simple Regression Tree Predictor* dado permitirem prever um valor numérico através de outros de diferente tipo. Escolhemos, aqui, prever a altura de um desportista com base nas colunas selecionadas alteriormente.

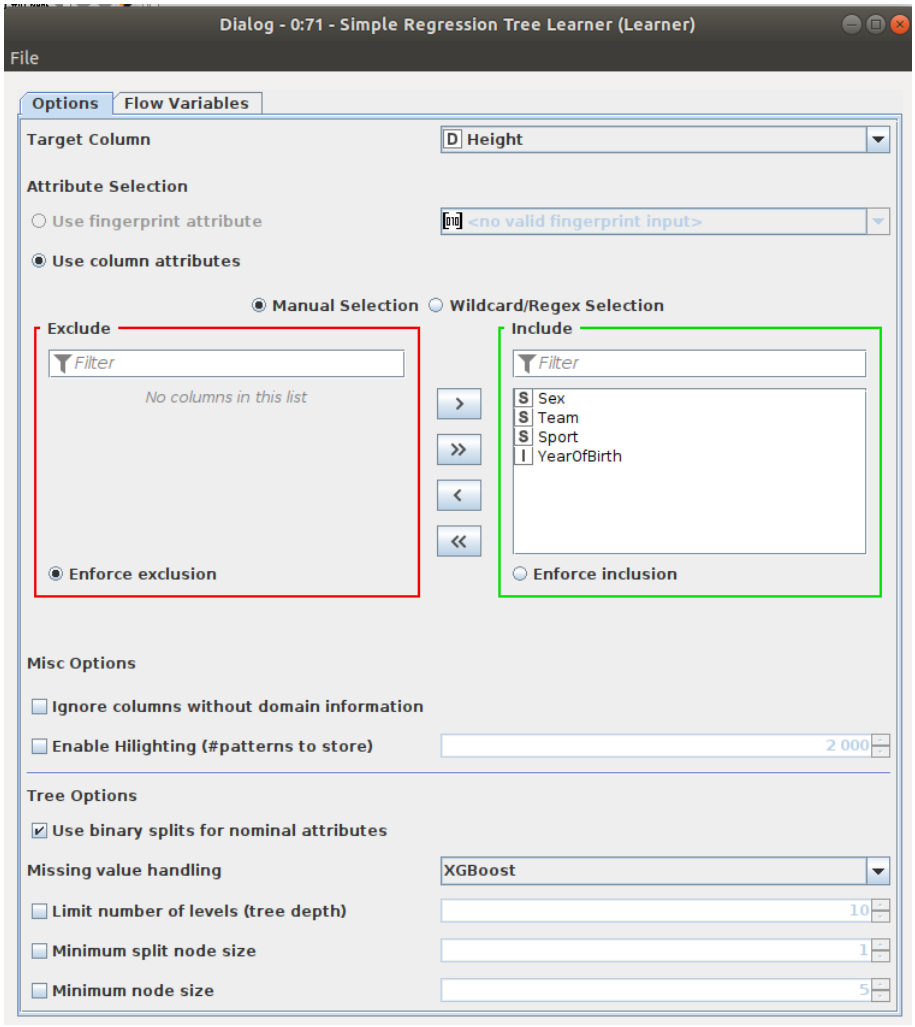


Figura 2.41: Simple Regression Tree Learner

- **X-Aggregator:** Junção dos dados de previsão das alturas com os restantes

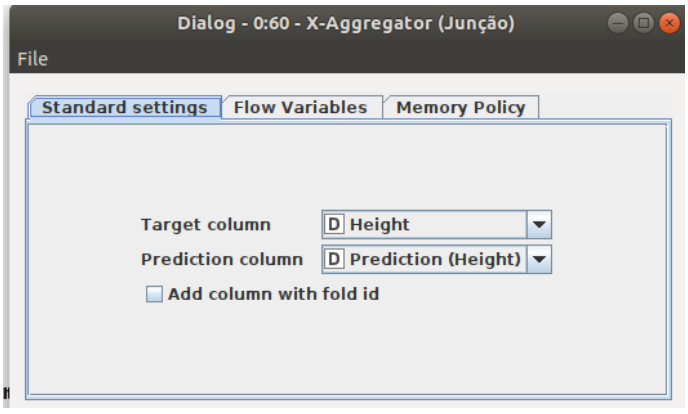


Figura 2.42: X-Aggregator

2.2.3 Apresentação dos resultados

- **Tabela de agregação:** Comparação entre as alturas previstas e as reais.

Prediction table - 3:60 - X-Aggregator (Junção)

File Edit Hilite Navigation View

Table "default" - Rows: 61586 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	S Sex	D Height	S Team	S Sport	I YearOf...	D Predicti...
Row1	M	170	China	Judo	1989	174.5
Row6	F	185	Netherlands	Speed Skating	1967	166
Row10	M	188	United States	Cross Count...	1961	183
Row32	F	159	Finland	Sailing	1966	164
Row160	M	133	Chad	Judo	1961	133
Row186	M	165	United States	Wrestling	1978	172.5
Row199	M	190	France	Handball	1970	188.333
Row235	M	175	Iraq	Football	1982	185
Row242	M	165	Egypt	Cycling	1978	179.667
Row246	M	173	Qatar	Weightlifting	1976	133
Row286	F	191	United States	Softball	1985	167
Row299	M	203	Egypt	Volleyball	1988	194
Row436	M	194	Egypt	Athletics	1989	191
Row446	M	170	Algeria	Athletics	1969	178.5
Row449	M	170	Iran	Wrestling	1990	188
Row451	M	176	Belgium	Athletics	1989	178.5
Row453	M	178	Australia	Athletics	1978	184.667
Row470	M	169	Australia	Wrestling	1981	175
Row504	M	175	Sudan	Swimming	1981	166.5
Row536	M	133	United Arab ...	Shooting	1957	184
Row542	M	185	Iraq	Football	1983	170
Row578	F	133	Singapore	Badminton	1971	158
Row584	M	164	Azerbaijan	Wrestling	1969	164
Row588	M	190	Uzbekistan	Wrestling	1990	165
Row596	F	162	Egypt	Weightlifting	1992	160
Row605	M	181	Kuwait	Athletics	1969	182
Row610	M	178	Uzbekistan	Boxing	1982	168
Row611	M	183	Kyrgyzstan	Taekwondo	1989	160
Row613	M	172	Uzbekistan	Wrestling	1990	165
Row615	M	172	Russia	Boxing	1990	180
Row651	M	175	Japan	Cycling	1969	175.25
Row742	M	133	Algeria	Handball	1971	181
Row749	F	160	Sri Lanka	Shooting	1970	170
Row774	F	159	Nigeria	Athletics	1988	177
Row851	M	171	Morocco	Football	1980	177.75
Row978	M	182	Russia-2	Bobsleigh	1983	187
Row980	F	163	Guyana	Athletics	1997	163

Figura 2.43: Tabela de agregação

- **Numeric Scorer:** Estatisticas da previsão feita

Figura 2.44: Numeric Scorers

3 Conclusão

Com a realização deste trabalho prático, o grupo conseguiu aprofundar conhecimento relativo à análise e tratamento de dados através da exploração de dois datasets, utilizando os modelos de aprendizagem abordados ao longo do semestre.