

Universidade do Minho
Licenciatura em Engenharia Informática

Aprendizagem e Decisão Inteligentes Grupo 26

André Silva - A87958

Armando Silva - A87949

Joana Oliveira - A87956

João Nunes - A87972

Maio 2022



Conteúdo

| | | |
|----------|---|-----------|
| 1 | Introdução | 3 |
| 2 | Apresentação e Exploração dos <i>Datasets</i> | 4 |
| 2.1 | Estimar o valor da habitação numa região dos EUA | 4 |
| 2.2 | Prever a temperatura do solo no mundo, por região | 5 |
| 3 | Preparação dos <i>Datasets</i> | 7 |
| 3.1 | Estimar o valor da habitação numa região dos EUA | 7 |
| 3.1.1 | Remoção de colunas | 7 |
| 3.1.2 | Identificação dos estados americanos | 8 |
| 3.1.3 | Identificação das regiões americanas | 9 |
| 3.1.4 | Normalização | 10 |
| 3.2 | Prever a temperatura do solo no mundo, por região | 11 |
| 3.2.1 | Preparação inicial | 11 |
| 3.2.2 | Extração da data | 12 |
| 3.2.3 | Regiões do mundo e continentes | 13 |
| 3.2.4 | Intervalos da temperatura | 14 |
| 3.2.5 | Remoção de colunas | 15 |
| 4 | Modelos Desenvolvidos | 17 |
| 4.1 | Estimar o valor da habitação numa região dos EUA | 17 |
| 4.2 | Prever a temperatura do solo no mundo, por região | 18 |
| 5 | Resultados Finais e Análise Crítica | 19 |
| 5.1 | Estimar o valor da habitação numa região dos EUA | 19 |
| 5.2 | Prever a temperatura do solo no mundo, por região | 21 |
| 6 | Sugestões e Recomendações | 23 |
| 6.1 | Estimar o valor da habitação numa região dos EUA | 23 |
| 6.2 | Prever a temperatura do solo no mundo, por região | 23 |
| 7 | Conclusão | 24 |

Lista de Figuras

| | | |
|----|---|----|
| 1 | <i>Box Plots</i> usados para observar a existência de <i>outliers</i> | 5 |
| 2 | <i>Rank Correlation</i> | 5 |
| 3 | Existência de <i>outliers</i> e análise da correlação | 6 |
| 4 | Algumas estatísticas sobre os dois principais atributos | 6 |
| 5 | Preparação dos dados do <i>dataset</i> recebido | 7 |
| 6 | Nodo <i>String Manipulation</i> com a expressão regular | 8 |
| 7 | Nodo <i>Rule Engine</i> com as regras dos estados | 8 |
| 8 | Nodo <i>Rule Engine</i> com as regras das regiões | 9 |
| 9 | Cores associadas a cada região | 9 |
| 10 | Tabela final do <i>dataset</i> recebido | 10 |
| 11 | <i>Pie Charts</i> usados para mostrar algumas estatísticas do <i>dataset</i> . . . | 11 |
| 12 | Preparação dos dados do <i>dataset</i> escolhido | 11 |
| 13 | <i>Metanode</i> com os nodos utilizados | 12 |
| 14 | <i>Metanode</i> com os nodos responsáveis pelas regras das regiões | 13 |
| 15 | Nodos utilizados para a adição das colunas <i>Region</i> e <i>Continent</i> | 14 |
| 16 | Cores atribuídas às classificações | 14 |
| 17 | Definições do nodo <i>Auto-Binner</i> | 15 |
| 18 | Tabela final do <i>dataset</i> escolhido | 15 |
| 19 | Histograma com o número de ocorrências de cada continente | 16 |
| 20 | <i>Pie Chart</i> com a temperatura média de cada continente | 16 |
| 21 | <i>Feature Selection Loop</i> do <i>dataset</i> recebido | 19 |
| 22 | Atributos selecionados do <i>dataset</i> recebido | 19 |
| 23 | Resultado obtido com <i>Linear Regression</i> | 20 |
| 24 | <i>Scatter Plot</i> do resultado obtido do <i>dataset</i> recebido | 20 |
| 25 | <i>Feature Selection Loop</i> do <i>dataset</i> escolhido | 21 |
| 26 | Atributos selecionados do <i>dataset</i> escolhido | 21 |
| 27 | Matriz de confusão | 22 |
| 28 | Resultado do modelo do <i>dataset</i> escolhido | 22 |
| 29 | Estatísticas finais obtidas do <i>dataset</i> escolhido | 22 |

1 Introdução

No âmbito da unidade curricular de Aprendizagem e Decisão Inteligentes, foi nos proposta a exploração, modelação e análise de dois *datasets*. O primeiro, fornecido pelos professores, com o objetivo de estimar o valor da habitação numa região dos Estados Unidos da América. O segundo, escolhido pelo grupo, focado na previsão da temperatura do solo em todas as regiões do mundo.

Para o primeiro *dataset*, é necessário alcançar o objetivo utilizando o algoritmo de regressão linear, ou seja, tentar estimar com a maior precisão possível, o valor da habitação numa região dos EUA, sendo este um problema com supervisão de regressão.

Já para o segundo *dataset*, para abrangermos diferentes preparações de dados e algoritmos, decidimos prever a classificação da temperatura do solo, ou seja, se esta será *Very Cold*, *Cold*, *Cool*, *Warm*, *Hot* ou *Very Hot*, em todas as regiões do mundo, sendo este, um problema com supervisão de classificação.

De forma a obtermos uma melhor compreensão, implementação e desenvolvimento dos *datasets* e, ainda, de ajudar no planeamento dos mesmos, o grupo decidiu optar pela metodologia CRISP-DM. Esta consiste em seis etapas: estudar o negócio, estudar os dados, preparar os dados, modelar, avaliar e desenvolver.

O presente relatório exhibe, explica e justifica os *datasets* escolhidos, todas as decisões tomadas e todo o processo desde o entendimento do *dataset* até à avaliação do modelo final. No final, é realizada uma análise crítica dos resultados obtidos e algumas sugestões são escritas.

2 Apresentação e Exploração dos *Datasets*

A primeira etapa para a realização deste trabalho prático consistiu no entendimento e na exploração dos *datasets*, pois estes podem conter dados incompletos, errados ou incoerentes. Esta etapa ajuda assim a entender a preparação que é necessária para atingir o objetivo de cada um deles e, ainda, a perceber qual o método de avaliação do modelo mais adequado.

Começamos por analisar cada atributo dos *datasets*, através dos nodos de estatística disponíveis no KNIME, e a partir daqui retirar algumas conclusões sobre a preparação de dados necessária.

Apresentamos a seguir cada *dataset*, explicando cada atributo, os nodos utilizados para a exploração dos mesmos e o objetivo final pretendido.

2.1 Estimar o valor da habitação numa região dos EUA

Dataset fornecido pelos professores, que tem como objetivo estimar o valor da habitação numa região dos Estados Unidos da América e com supervisão de regressão. Inicialmente, o *dataset* contém os seguintes **atributos**:

- Avg. Area Income, média de renda dos residentes de uma cidade;
- Avg. Area House Age, média de idade das casas na mesma cidade;
- Avg. Area Number of Rooms, média de número de divisões de uma casa na mesma cidade;
- Avg. Area Number of Bedrooms, média de número de quartos de uma casa na mesma cidade;
- Area Population, população da mesma cidade;
- Price (objetivo), preço a que a casa foi vendida;
- Address, endereço da casa.

Para a exploração de todas as entradas do *dataset*, utilizamos os nodos *Statistics*, *Data Explorer*, *Box Plot*, *Linear* e *Rank Correlation*. Através destes nodos, descobrimos que existem *outliers*, que os valores do *standard deviation* e do *skewness* são válidos e que existe uma correlação baixa, perto de zero, entre o atributo objetivo, *Price*, e o atributo *Avg. Area Number of Bedrooms*.

A seguir apresentamos algumas imagens que mostram alguns resultados da exploração de dados feita.

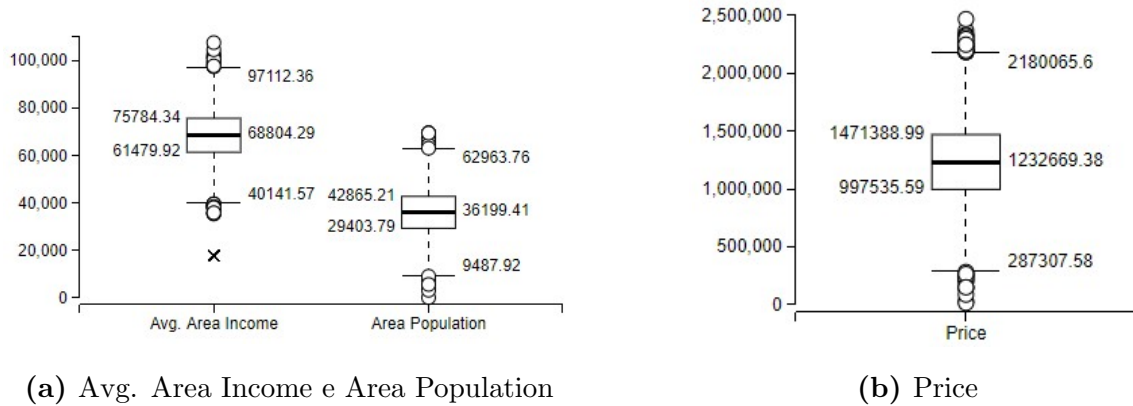


Figura 1. *Box Plots* usados para observar a existência de *outliers*

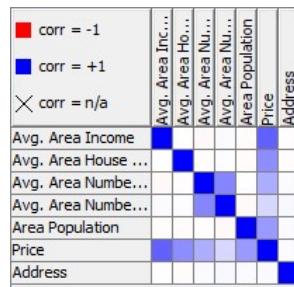


Figura 2. *Rank Correlation*

2.2 Prever a temperatura do solo no mundo, por região

Dataset escolhido pelo grupo através da plataforma data.world com o objetivo de prever a temperatura do solo no mundo dividido em regiões. Este é um *dataset* com supervisão de classificação. Inicialmente este *dataset* tem poucos atributos, pelo que um dos principais focos na preparação de dados foi aumentar substancialmente a quantidade de informação. Apesar disso, continua a ser importante apresentar os **atributos iniciais**:

- dt, data em que foi recolhida a temperatura;
- AverageTemperature (objetivo), temperatura média;
- AverageTemperatureUncertainty, incerteza do valor da temperatura média (quanto mais alto, menos preciso é o valor);
- Country, país onde foi recolhida a temperatura.

Para a exploração deste *dataset*, utilizamos os mesmos nodos já referidos na exploração do *dataset* anterior. Descobrimos, assim, que existem *missing values*, *outli-*

ers e maus valores de *standard deviation* e *skewness*. Isto acontece por haver poucos atributos, e também a existência de algumas entradas que não pertencem ao estudo que queremos fazer (por exemplo, o nome de um continente em vez de um país).

Exibimos a seguir, algumas imagens que obtivemos durante a exploração de dados realizada para este *dataset*.

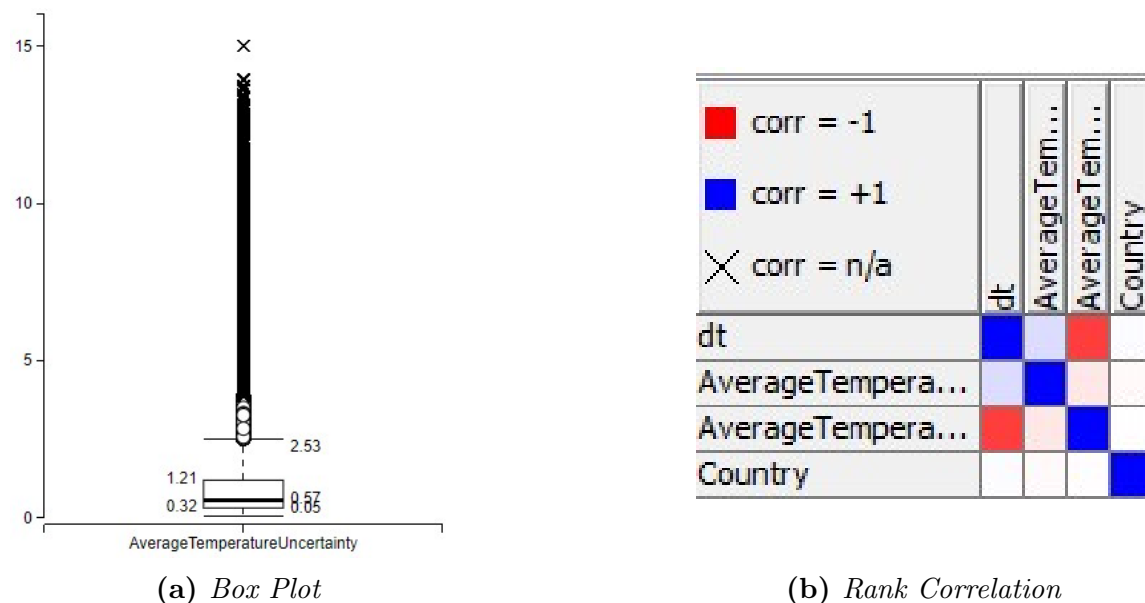


Figura 3. Existência de *outliers* e análise da correlação

| Column | Min | Mean | Median | Max | Std. Dev. | Skewness | Kurtosis | No. Missing |
|-------------------------------|---------|---------|--------|--------|-----------|----------|----------|-------------|
| AverageTemperature | -37,658 | 17,1934 | ? | 38,842 | 10,954 | -1,1143 | 1,0688 | 32 651 |
| AverageTemperatureUncertainty | 0,052 | 1,0191 | ? | 15,003 | 1,2019 | 3,1426 | 13,869 | 31 912 |

Figura 4. Algumas estatísticas sobre os dois principais atributos

3 Preparação dos *Datasets*

Depois de uma completa análise realizada aos dois *datasets*, passamos para a preparação dos dados com o objetivo de eliminarmos qualquer informação não pretendida. Assim conseguimos colocar os dados da melhor forma possível, de modo a atingirmos os melhores resultados finais.

Começamos por fazer uma limpeza nos dados, tratando, por exemplo, dos *missing values* e *outliers*. Depois é realizado um aumento na informação de acordo com o pretendido do problema. Desta forma, os *datasets* ficam prontos para a avaliação final do modelo desenhado.

Exibimos, a seguir, todo o tratamento de dados efetuado em cada um dos *datasets*.

3.1 Estimar o valor da habitação numa região dos EUA

O tratamento dos *missing values* não foi necessário, devido à inexistência dos mesmos. Apesar de existirem *outliers* o grupo assumiu que todos estes são valores válidos no contexto do problema, ou seja, nenhum dos valores fora do normal foi causado por erros nas recolhas de dados ou na inserção no ficheiro. Desta forma, não foram retiradas nenhuma linha do *dataset*, mantendo o pequeno número de linhas inicial, 5000, intacto. Não havia entradas duplicadas e por isso este tratamento também não foi utilizado.

Apresentamos a seguir uma imagem da preparação dos dados efetuada neste *dataset*.

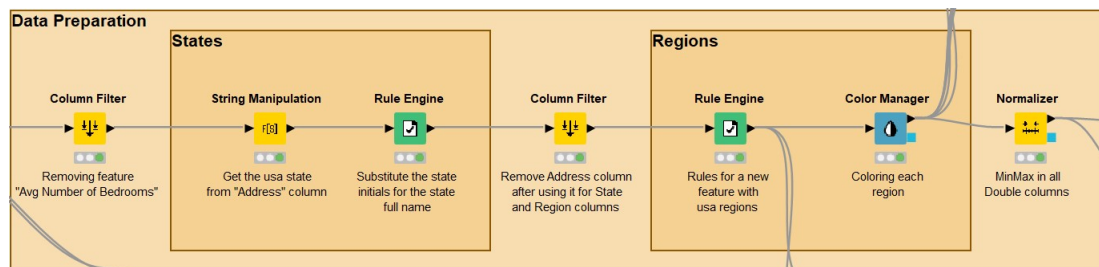


Figura 5. Preparação dos dados do *dataset* recebido

Nas próximas subsubsecções iremos apresentar de uma forma mais detalhada a preparação que foi realizada para este *dataset*.

3.1.1 Remoção de colunas

Na secção anterior, observamos que a correlação entre o atributo objetivo *Price* e o *Avg. Area Number of Bedrooms* é extremamente baixa, precisamente 0.1582, e como tal, não vai influenciar na obtenção do resultado final. Assim, o grupo decidiu remover esta coluna através do nodo *Column Filter*, primeiro nodo da figura 5.

3.1.2 Identificação dos estados americanos

Para obtermos uma visualização gráfica mais interessante, com alguns dados relevantes para serem apresentados e para adicionar mais informação a este *dataset*, decidimos, através do atributo *Address*, descobrir o estado onde a casa se encontra e criar uma nova coluna com essa nova informação.

Para isso, tivemos de criar uma expressão regular que, recebendo um endereço, reconhece as duas letras que identificam o estado. Aplicamos essa expressão regular e criamos a nova coluna através do nodo *String Manipulation*. De seguida, com o nodo *Rule Engine* geramos regras que associam as iniciais dos estados da nova coluna para o respetivo nome completo. Acrescentamos ainda, que a coluna *Address* foi removida por já ter sido utilizada para a criação de informação mais útil para os algoritmos que utilizaremos para a obtenção do resultado pretendido.

Apresentamos a seguir algumas imagens que acrescentam algum contexto ao que foi explicado.

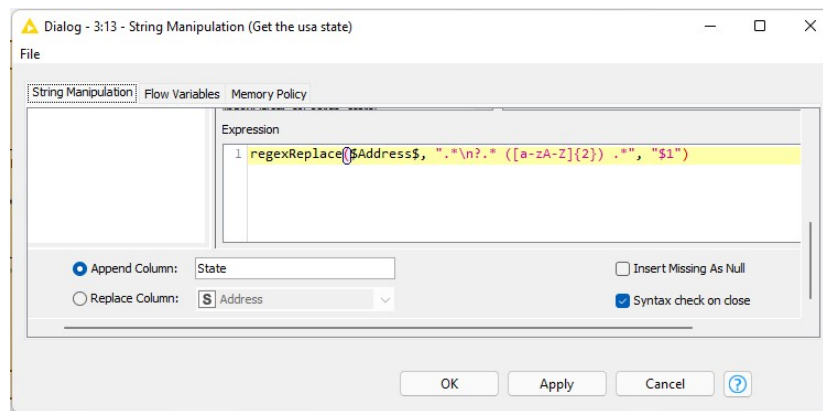


Figura 6. Nodo *String Manipulation* com a expressão regular

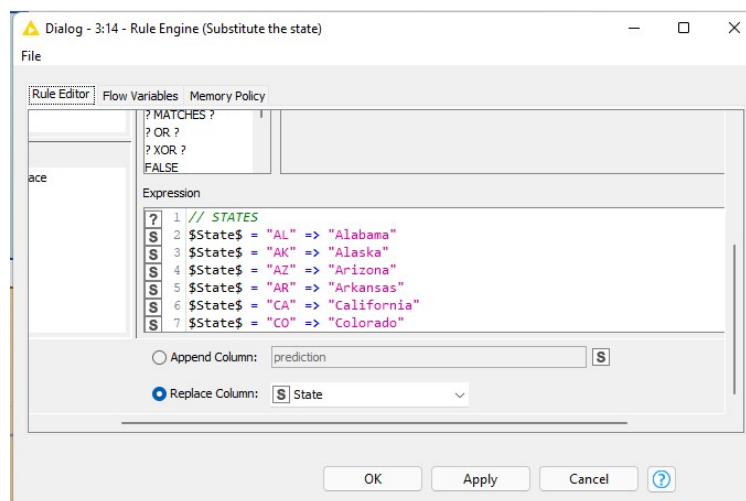


Figura 7. Nodo *Rule Engine* com as regras dos estados

3.1.3 Identificação das regiões americanas

De forma a completarmos ainda mais a informação do nosso *dataset* e acrescentar mais alternativas para a criação dos gráficos estatísticos, o grupo decidiu incluir uma nova coluna com a região do estado nos Estados Unidos da América. Utilizamos o nodo *Rule Engine* com várias regras que conectam os estados a certas regiões. De salientar que alguns estados não se encontram em nenhuma região, portanto foram associados ao valor *Others*. Um exemplo deste caso é o estado *U.S. Armed Forces - Pacific* que foi associado a esse valor.

Acrescentamos também mediante o nodo *Color Manager* algumas cores para cada região com o objetivo de tornar a visualização das estatísticas e da tabela mais fácil e agradável.

Exibimos agora, uma imagem com as regras criadas para a coluna das regiões e as cores atribuídas a cada região.

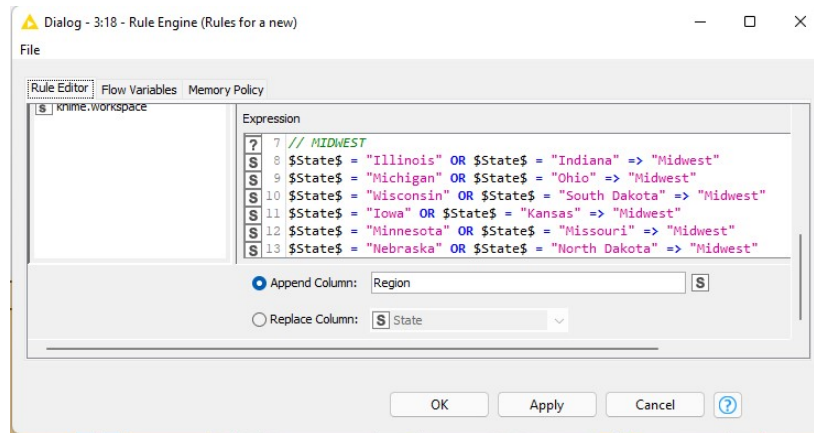


Figura 8. Nodo *Rule Engine* com as regras das regiões

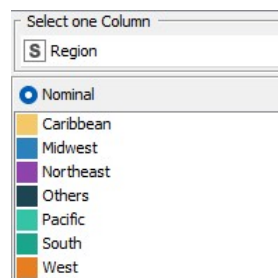


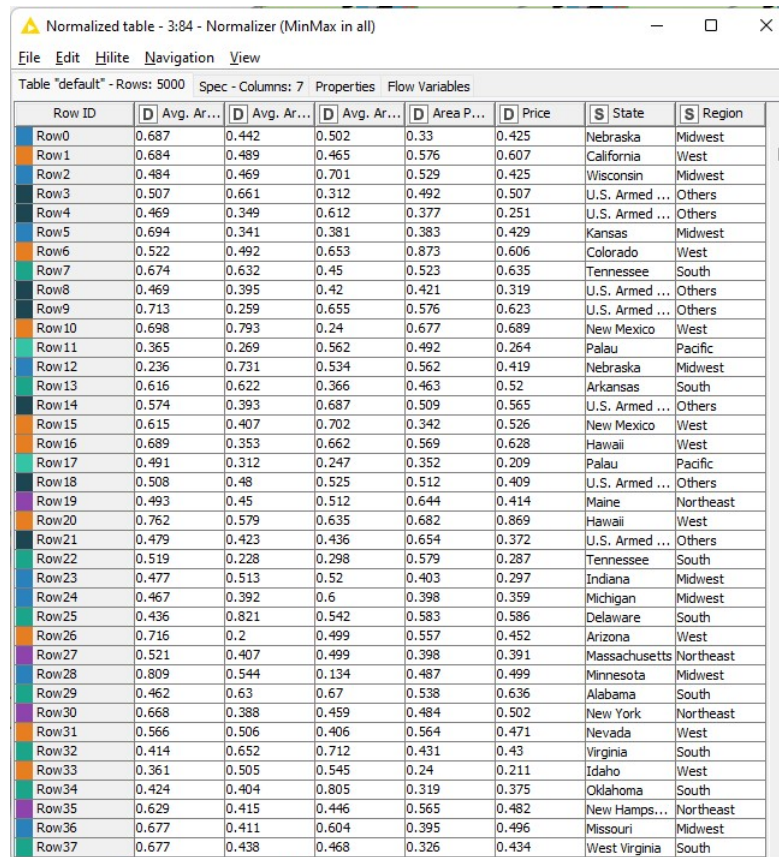
Figura 9. Cores associadas a cada região

3.1.4 Normalização

Todos os atributos deste *dataset*, sem contar com o *Address*, são do tipo *Double*, e por isso, o grupo decidiu diminuir a escala de todos estes atributos. Para isso, utilizamos o nodo *Normalizer*, segundo da figura 5, com uma normalização Min-Max em que o mínimo é 0.0 e o máximo é 1.0.

Acrescentamos ainda que, a normalização do atributo objetivo *Price* fez com que fosse possível calcular os erros no resultado final, que de outra forma não seria possível, pois originavam números extremamente altos devido aos valores iniciais do atributo serem grandes.

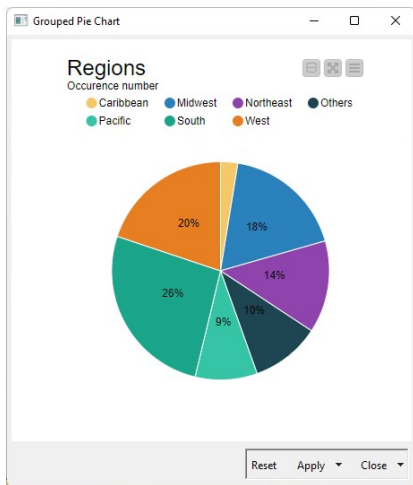
Apresentamos agora um excerto da tabela com toda a preparação completa.



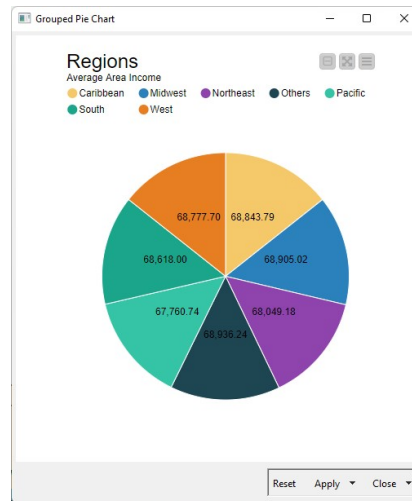
| Row ID | D Avg. Ar... | D Avg. Ar... | D Avg. Ar... | D Area P... | D Price | S State | S Region |
|--------|--------------|--------------|--------------|-------------|---------|----------------|-----------|
| Row0 | 0.687 | 0.442 | 0.502 | 0.33 | 0.425 | Nebraska | Midwest |
| Row1 | 0.684 | 0.489 | 0.465 | 0.576 | 0.607 | California | West |
| Row2 | 0.484 | 0.469 | 0.701 | 0.529 | 0.425 | Wisconsin | Midwest |
| Row3 | 0.507 | 0.661 | 0.312 | 0.492 | 0.507 | U.S. Armed ... | Others |
| Row4 | 0.469 | 0.349 | 0.612 | 0.377 | 0.251 | U.S. Armed ... | Others |
| Row5 | 0.694 | 0.341 | 0.381 | 0.383 | 0.429 | Kansas | Midwest |
| Row6 | 0.522 | 0.492 | 0.653 | 0.873 | 0.606 | Colorado | West |
| Row7 | 0.674 | 0.632 | 0.45 | 0.523 | 0.635 | Tennessee | South |
| Row8 | 0.469 | 0.395 | 0.42 | 0.421 | 0.319 | U.S. Armed ... | Others |
| Row9 | 0.713 | 0.259 | 0.655 | 0.576 | 0.623 | U.S. Armed ... | Others |
| Row10 | 0.698 | 0.793 | 0.24 | 0.677 | 0.689 | New Mexico | West |
| Row11 | 0.365 | 0.269 | 0.562 | 0.492 | 0.264 | Palau | Pacific |
| Row12 | 0.236 | 0.731 | 0.534 | 0.562 | 0.419 | Nebraska | Midwest |
| Row13 | 0.616 | 0.622 | 0.366 | 0.463 | 0.52 | Arkansas | South |
| Row14 | 0.574 | 0.393 | 0.687 | 0.509 | 0.565 | U.S. Armed ... | Others |
| Row15 | 0.615 | 0.407 | 0.702 | 0.342 | 0.526 | New Mexico | West |
| Row16 | 0.689 | 0.353 | 0.662 | 0.569 | 0.628 | Hawaii | West |
| Row17 | 0.491 | 0.312 | 0.247 | 0.352 | 0.209 | Palau | Pacific |
| Row18 | 0.508 | 0.48 | 0.525 | 0.512 | 0.409 | U.S. Armed ... | Others |
| Row19 | 0.493 | 0.45 | 0.512 | 0.644 | 0.414 | Maine | Northeast |
| Row20 | 0.762 | 0.579 | 0.635 | 0.682 | 0.869 | Hawaii | West |
| Row21 | 0.479 | 0.423 | 0.436 | 0.654 | 0.372 | U.S. Armed ... | Others |
| Row22 | 0.519 | 0.228 | 0.298 | 0.579 | 0.287 | Tennessee | South |
| Row23 | 0.477 | 0.513 | 0.52 | 0.403 | 0.297 | Indiana | Midwest |
| Row24 | 0.467 | 0.392 | 0.6 | 0.398 | 0.359 | Michigan | Midwest |
| Row25 | 0.436 | 0.821 | 0.542 | 0.583 | 0.586 | Delaware | South |
| Row26 | 0.716 | 0.2 | 0.499 | 0.557 | 0.452 | Arizona | West |
| Row27 | 0.521 | 0.407 | 0.499 | 0.398 | 0.391 | Massachusetts | Northeast |
| Row28 | 0.809 | 0.544 | 0.134 | 0.487 | 0.499 | Minnesota | Midwest |
| Row29 | 0.462 | 0.63 | 0.67 | 0.538 | 0.636 | Alabama | South |
| Row30 | 0.668 | 0.388 | 0.459 | 0.484 | 0.502 | New York | Northeast |
| Row31 | 0.566 | 0.506 | 0.406 | 0.564 | 0.471 | Nevada | West |
| Row32 | 0.414 | 0.652 | 0.712 | 0.431 | 0.43 | Virginia | South |
| Row33 | 0.361 | 0.505 | 0.545 | 0.24 | 0.211 | Idaho | West |
| Row34 | 0.424 | 0.404 | 0.805 | 0.319 | 0.375 | Oklahoma | South |
| Row35 | 0.629 | 0.415 | 0.446 | 0.565 | 0.482 | New Hamps... | Northeast |
| Row36 | 0.677 | 0.411 | 0.604 | 0.395 | 0.496 | Missouri | Midwest |
| Row37 | 0.677 | 0.438 | 0.468 | 0.326 | 0.434 | West Virginia | South |

Figura 10. Tabela final do *dataset* recebido

Adicionalmente, achamos pertinente colocar imagens de alguns gráficos estatísticos, criados pelo grupo, de forma a entendermos melhor a preparação de dados efetuada. De salientar que estes gráficos foram criados antes da normalização dos valores.



(a) Percentagem de regiões



(b) Média salarial anual por região

Figura 11. *Pie Charts* usados para mostrar algumas estatísticas do *dataset*

3.2 Prever a temperatura do solo no mundo, por região

O tratamento das entradas duplicadas não foi necessário por não existirem no nosso *dataset*. Este inicialmente contava com perto de 580000 linhas, o que nos deu uma margem maior para o tratamento de dados não pretendidos.

Apresentamos a seguir, uma imagem da preparação dos dados efetuada no *dataset* escolhido pelo grupo.

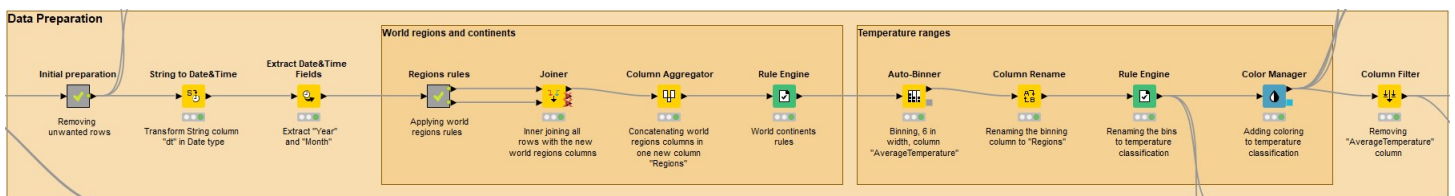


Figura 12. Preparação dos dados do *dataset* escolhido

Nas próximas subsubsecções iremos apresentar de uma forma mais detalhada a preparação que foi realizada para o *dataset* escolhido pelo grupo.

3.2.1 Preparação inicial

Começamos com o tratamento dos *missing values* que existiam neste *dataset*. Devido ao grande número, mais de 30000 linhas, que continham valores em falta, o grupo decidiu eliminar todas estas entradas. Outro fator que contribuiu para o grupo ter tomado esta decisão foi que, caso tivessemos decidido inserir valores nestas entradas, estas poderiam influenciar na classificação final da temperatura, pois consideramos o valor da temperatura um atributo bastante volátil.

Durante a exploração do *dataset* encontramos algumas entradas que não faziam parte do problema que nos propusemos. Por exemplo, caso o país fosse um continente como 'Europa', esta entrada seria eliminada. Outras entradas que identificamos como estranhas foram aquelas que, para o mesmo país, apareciam com nomes escritos de formas diferentes, por exemplo 'Denmark' e 'Denmark (Europe)'. Estas estavam nas mesmas datas e o grupo decidiu apenas manter uma destas entradas para cada país. Depois de uma pesquisa sobre a diferença entre estas, entendemos que a melhor opção seria eliminar as linhas sem, por exemplo '... (Europe)'. Através de uma expressão regular escrita no nodo *Row Filter* conseguimos removê-las e ainda modificar o seu nome retirando, por exemplo, '(Europe)'.

Para o tratamento dos *outliers* o grupo decidiu remover as linhas que continham *outliers* no atributo *AverageTemperatureUncertainty*. Esta decisão foi tomada pois este atributo representa a incerteza da temperatura medida, e um valor alto significa que o valor da temperatura não pode ser considerado por ser impreciso.

Assim, foram eliminadas cerca de 100000 entradas indesejadas. Apresentamos agora o *Metanode* criado com estes nodos de tratamento de dados.

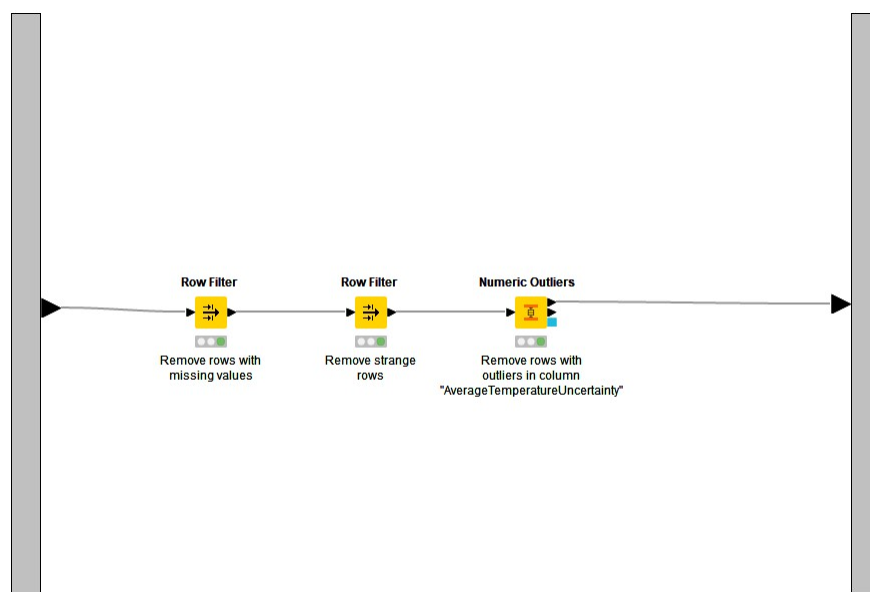


Figura 13. *Metanode* com os nodos utilizados

3.2.2 Extração da data

O primeiro atributo deste *dataset* é a data de registo da temperatura, com o tipo *String*. Para tornarmos este atributo mais informativo, decidimos primeiramente convertê-lo para o tipo *Date*, o que nos possibilitou a extração dos seus campos. Para tal, extraímos os campos *Month* e *Year*. Aumentamos assim a informação total que este atributo nos fornece, adicionando duas novas colunas.

3.2.3 Regiões do mundo e continentes

Como já foi mencionado na secção anterior, o *dataset* escolhido pelo grupo inicialmente é pobre em informação, uma vez que apenas tem 4 atributos. Por essa razão, o grupo considerou que uma das tarefas mais importantes desta fase era o acréscimo do máximo de informação possível.

Para tal o grupo decidiu pegar em cada país e associá-lo a uma região do mundo. Por exemplo, a China pertence à região *Eastern Asia* e a Angola a *Middle Africa*. Para concretizarmos isto foi necessária a implementação de regras para cada continente, através de 5 nodos *Rule Engine*. As regiões implementadas, por continente, foram as seguintes:

- **África:** *Eastern, Middle, Northern, Southern e Western*;
- **América:** *Caribbean, Central, Northern e Southern*;
- **Ásia:** *Central, Eastern, Southeastern, Southern e Western*;
- **Europa:** *Eastern, Northern, Southern e Western*;
- **Oceânia:** *Australia and New Zealand, Melanesia, Micronesia e Polynesia*.

A seguir, encontra-se uma imagem com os respetivos nodos mencionados.

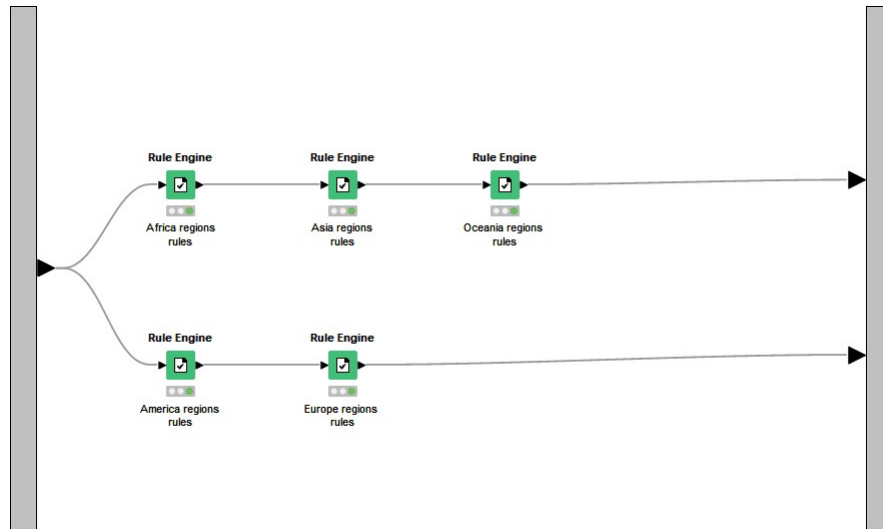


Figura 14. *Metanode* com os nodos responsáveis pelas regras das regiões

Cada nodo vai criar uma nova coluna, no entanto, o objetivo do grupo era juntar todas estas em apenas uma, com o nome de *Region*. Para tal foi necessária a utilização do *Joiner* com as definições, *Inner join* e *Merge join columns*. E, ainda, do *Column Aggregator*, que concatenou todas as colunas em apenas uma, como era pretendido.

De modo a completar ainda mais esta parte, o grupo resolveu adicionar mais uma nova coluna, *Continent*. Para isso, foram criadas novas regras que associam cada região ao respetivo continente, mais uma vez através do nodo *Rule Engine*.

Apresentamos a seguir, a parte da preparação dos dados responsável por esta implementação.

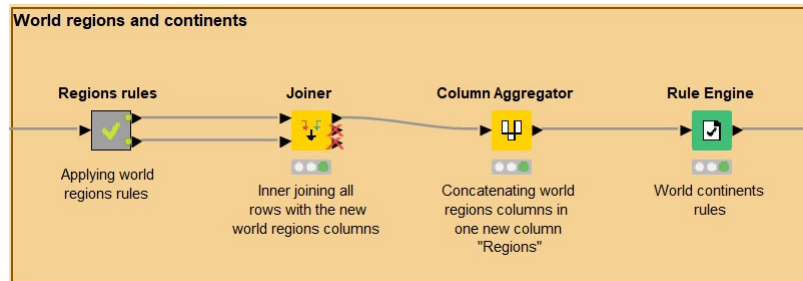


Figura 15. Nodos utilizados para a adição das colunas *Region* e *Continent*

3.2.4 Intervalos da temperatura

Para implementarmos as classificações neste problema, o grupo optou pela utilização do nodo *Auto-Binner* que dividiu o atributo *AverageTemperature* em 6 blocos iguais. Desta forma conseguimos obter diferentes classificações para diferentes intervalos de temperatura, sendo essas: **Very Cold**, **Cold**, **Cool**, **Warm**, **Hot** e **Very Hot**.

Importante realçar que foram realizados vários testes que nos ajudaram a decidir qual o método que o *Auto-Binner* utiliza para a divisão dos blocos. Foram testados os seguintes métodos: largura, frequência e intervalos definidos pelo grupo. No fim, obtivemos a melhor precisão com a utilização da divisão em largura.

Depois de termos os *Bins* definidos, utilizamos o nodo *Column Rename* e *Rule Engine* para renomearmos o nome da coluna para *Temperature Classification* e os respetivos *Bins* para os nomes já mencionados. Adicionamos ainda cores para cada classificação.

Apresentamos agora algumas imagens que demonstram o que foi explicado.

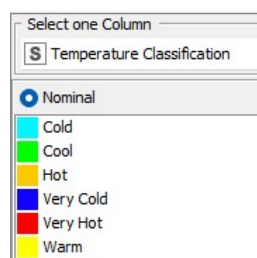


Figura 16. Cores atribuídas às classificações

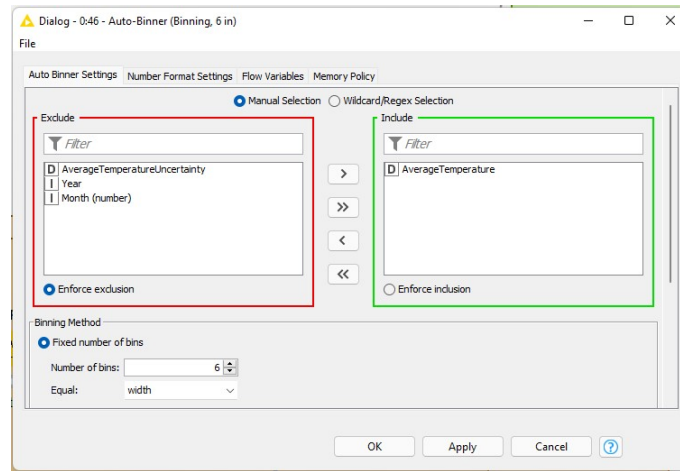


Figura 17. Definições do nodo *Auto-Binner*

3.2.5 Remoção de colunas

Depois de toda esta preparação de dados, o grupo começou a obter precisões de 100%. Isto acontecia devido ao facto das classificações das temperaturas terem sido obtidas através do atributo *AverageTemperature*, pelo que qualquer algoritmo na validação do modelo conseguia facilmente acertar na classificação dada. Para isso não acontecer, o grupo decidiu remover este atributo da tabela.

Terminamos assim a preparação dos dados deste *dataset*. Resumindo, foram adicionadas 5 novas colunas e apenas removida 1, aumentando o total de atributos de 4 para 8. A seguir, apresentamos uma imagem com a tabela final obtida.

| Row ID | dt | D] Averag... | S] Country | I] Year | I] Month (...) | S] Region | S] Continent | S] Temper... |
|-----------|------------|--------------|------------|---------|----------------|-----------------|--------------|--------------|
| Row341039 | 2012-07-01 | 0.342 | Poland | 2012 | 7 | Eastern Europe | Europe | Hot |
| Row341040 | 2012-08-01 | 0.329 | Poland | 2012 | 8 | Eastern Europe | Europe | Hot |
| Row341041 | 2012-09-01 | 0.203 | Poland | 2012 | 9 | Eastern Europe | Europe | Hot |
| Row341042 | 2012-10-01 | 0.193 | Poland | 2012 | 10 | Eastern Europe | Europe | Warm |
| Row341043 | 2012-11-01 | 0.286 | Poland | 2012 | 11 | Eastern Europe | Europe | Warm |
| Row341044 | 2012-12-01 | 0.348 | Poland | 2012 | 12 | Eastern Europe | Europe | Cool |
| Row341045 | 2013-01-01 | 0.452 | Poland | 2013 | 1 | Eastern Europe | Europe | Cool |
| Row341046 | 2013-02-01 | 0.388 | Poland | 2013 | 2 | Eastern Europe | Europe | Cool |
| Row341047 | 2013-03-01 | 0.269 | Poland | 2013 | 3 | Eastern Europe | Europe | Cool |
| Row341048 | 2013-04-01 | 0.279 | Poland | 2013 | 4 | Eastern Europe | Europe | Warm |
| Row341049 | 2013-05-01 | 0.232 | Poland | 2013 | 5 | Eastern Europe | Europe | Hot |
| Row341050 | 2013-06-01 | 0.209 | Poland | 2013 | 6 | Eastern Europe | Europe | Hot |
| Row341051 | 2013-07-01 | 0.417 | Poland | 2013 | 7 | Eastern Europe | Europe | Hot |
| Row341052 | 2013-08-01 | 0.34 | Poland | 2013 | 8 | Eastern Europe | Europe | Hot |
| Row341053 | 1753-05-01 | 2.336 | Portugal | 1753 | 5 | Southern Europe | Europe | Hot |
| Row341054 | 1753-07-01 | 2.345 | Portugal | 1753 | 7 | Southern Europe | Europe | Hot |
| Row341055 | 1753-09-01 | 2.005 | Portugal | 1753 | 9 | Southern Europe | Europe | Hot |
| Row341056 | 1753-10-01 | 1.999 | Portugal | 1753 | 10 | Southern Europe | Europe | Hot |
| Row341057 | 1754-06-01 | 2.138 | Portugal | 1754 | 6 | Southern Europe | Europe | Hot |
| Row341058 | 1754-10-01 | 2.492 | Portugal | 1754 | 10 | Southern Europe | Europe | Hot |
| Row341059 | 1754-11-01 | 2.259 | Portugal | 1754 | 11 | Southern Europe | Europe | Warm |
| Row341060 | 1755-07-01 | 2.222 | Portugal | 1755 | 7 | Southern Europe | Europe | Hot |
| Row341061 | 1755-09-01 | 1.949 | Portugal | 1755 | 9 | Southern Europe | Europe | Hot |
| Row341062 | 1756-07-01 | 2.377 | Portugal | 1756 | 7 | Southern Europe | Europe | Hot |
| Row341063 | 1757-01-01 | 2.5 | Portugal | 1757 | 1 | Southern Europe | Europe | Warm |
| Row341064 | 1758-01-01 | 2.283 | Portugal | 1758 | 1 | Southern Europe | Europe | Warm |
| Row341065 | 1761-03-01 | 2.118 | Portugal | 1761 | 3 | Southern Europe | Europe | Warm |
| Row341066 | 1761-06-01 | 2.506 | Portugal | 1761 | 6 | Southern Europe | Europe | Hot |
| Row341067 | 1761-11-01 | 2.398 | Portugal | 1761 | 11 | Southern Europe | Europe | Warm |
| Row341068 | 1761-12-01 | 2.476 | Portugal | 1761 | 12 | Southern Europe | Europe | Warm |
| Row341069 | 1762-01-01 | 2.321 | Portugal | 1762 | 1 | Southern Europe | Europe | Warm |
| Row341070 | 1762-11-01 | 2.498 | Portugal | 1762 | 11 | Southern Europe | Europe | Warm |
| Row341071 | 1763-04-01 | 2.177 | Portugal | 1763 | 4 | Southern Europe | Europe | Warm |
| Row341072 | 1763-07-01 | 2.414 | Portugal | 1763 | 7 | Southern Europe | Europe | Hot |
| Row341073 | 1763-12-01 | 2.335 | Portugal | 1763 | 12 | Southern Europe | Europe | Warm |
| Row341074 | 1764-08-01 | 2.158 | Portugal | 1764 | 8 | Southern Europe | Europe | Hot |
| Row341075 | 1765-04-01 | 2.136 | Portugal | 1765 | 4 | Southern Europe | Europe | Warm |
| Row341076 | 1765-05-01 | 2.466 | Portugal | 1765 | 5 | Southern Europe | Europe | Hot |

Figura 18. Tabela final do *dataset* escolhido

Adicionalmente achamos importante mostrar alguns gráficos estatísticos de forma a observarmos mais facilmente a preparação de dados efetuada.

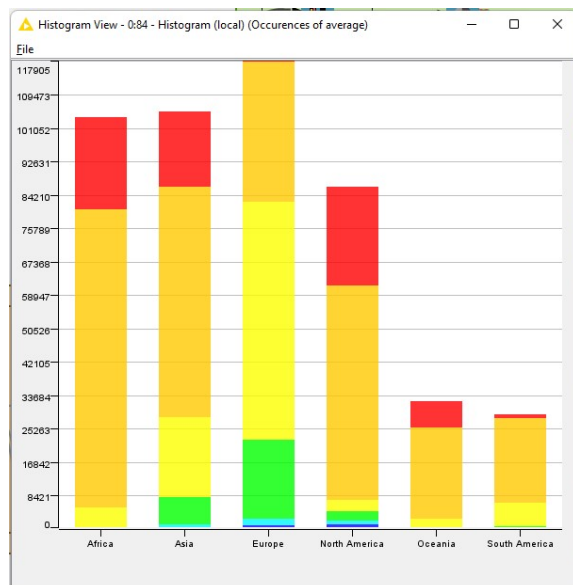


Figura 19. Histograma com o número de ocorrências de cada continente

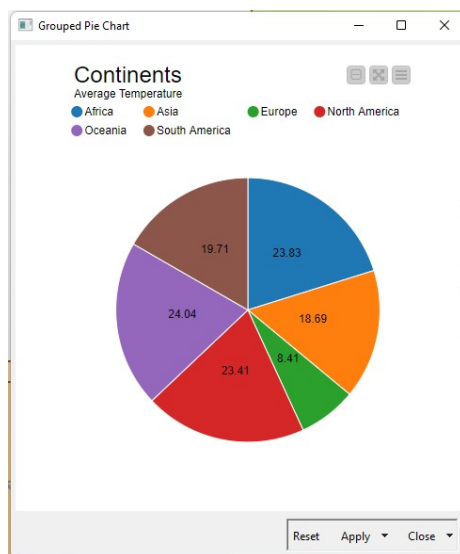


Figura 20. *Pie Chart* com a temperatura média de cada continente

4 Modelos Desenvolvidos

Já com a preparação de dados de ambos os *datasets* pronta, passamos para o desenvolvimento de modelos de aprendizagem. Desta forma, conseguimos testar todo o tratamento realizado e explicado na secção anterior, dando uma resposta ao problema proposto para cada um dos *datasets*.

De salientar que, para ambos os *datasets*, foram utilizados vários algoritmos de acordo com o problema em questão. Nesta secção iremos apresentar os algoritmos testados, quais as suas características e os parâmetros de treino utilizados.

4.1 Estimar o valor da habitação numa região dos EUA

Depois de termos o *dataset* preparado, passamos para a realização dos testes com os seguintes algoritmos: *Linear Regression*, *Regression Tree* e *Random Forest*. Relembramos que este problema é do tipo com supervisão de regressão e para a partição dos dados utilizamos o *loop* de nodos *X-Partitioner*, com a regra do polegar, e o respetivo *X-Aggregator*.

Linear Regression: Primeiro algoritmo escolhido e que utiliza uma equação para estimar uma variável, dadas outras variáveis com outros valores. Os nodos que utilizamos para este algoritmo foram o *Linear Regression Learner* e o *Regression Predictor*. De salientar que este foi o algoritmo que obtivemos melhores resultados, com um *R-Square* de 0.918, e portanto foi o utilizado para o procedimento *Feature Selection* e também para a validação e obtenção do resultado final. Tópicos que vão ser melhor explicados na secção seguinte.

Regression Tree: Algoritmo que utiliza uma árvore de decisão, mas que está adaptada para a regressão. Através de um processo iterativo, esta divide continuamente os dados em segmentos mais pequenos até encontrar as folhas terminais. Implementamos com os nodos *Simple Regression Tree Learner* e *Simple Regression Tree Predictor*. Com este algoritmo o grupo obteve um *R-Square* de 0.555. Valor significamente mais baixo comparando-o com o primeiro algoritmo testado.

Random Forest: Algoritmo do tipo *ensemble* que constrói múltiplas árvores de decisão. Para problemas de regressão, esta devolve a média prevista pelas diferentes árvores. Implementação realizada com os nodos *Random Forest Learner (Regression)* e *Random Forest Predictor (Regression)*. Para o nosso *dataset*, obtivemos um *R-Square* de 0.889. Valor alto, mas inferior à regressão linear.

4.2 Prever a temperatura do solo no mundo, por região

Com o *dataset* preparado, começamos a realização dos testes com os algoritmos: *Decision Tree*, *Random Forest*, *Tree Ensemble* e *Logistics Regression*. Relembramos que este problema é do tipo com supervisão de classificação. Para a partição dos dados utilizamos o *loop* de nodos *X-Partitioner*, com a regra do polegar, e o respetivo *X-Aggregator*.

***Decision Tree*:** Algoritmo que pode também ser considerado um grafo hierarquizado, em que cada ramo representa a seleção entre um conjunto de alternativas e as folhas representam uma decisão. Para a implementação utilizamos os nodos *Decision Tree Learner* e *Decision Tree Predictor*. No nosso *dataset* obtivemos uma precisão de 77.88%, ligeiramente mais baixo comparado com o melhor resultado atingido.

***Random Forest*:** Algoritmo do tipo *ensemble* que constrói múltiplas árvores de decisão. Para problemas de classificação, este devolve a classificação mais comum entre todas as árvores. Implementação realizada com os nodos *Random Forest Learner* e *Random Forest Predictor*. Com este algoritmo obtivemos uma precisão de 78.14%, valor mais alto alcançado. E como tal, este vai ser o algoritmo utilizado para o *Feature Selection* e a validação do modelo final. Tudo isto vai ser explicado na próxima secção.

***Tree Ensemble*:** Algoritmo que utiliza árvores de decisão com o objetivo de juntar um grupo de árvores de decisão com resultados fracos e formar resultados mais fortes. Para esta implementação utilizamos os nodos *Tree Ensemble Learner* e *Tree Ensemble Predictor*. Com este algoritmo obtivemos uma precisão de 78.10%, valor ligeiramente mais baixo que o melhor resultado que obtivemos.

***Logistics Regression*:** Utiliza uma função logística para produzir um modelo com previsões de valores a partir de várias outras variáveis. Os nodos implementados foram o *Logistic Regression Learner* e o *Logistic Regression Predictor*. Este algoritmo obteve uma precisão de 40.89%. Comparado a outros valores atingidos, este é o mais baixo por uma diferença bastante substancial.

5 Resultados Finais e Análise Crítica

Com todos os diferentes modelos de aprendizagem testados, apresentamos agora o algoritmo final escolhido, com algumas imagens da implementação, e o respetivo resultado para ambos os *datasets*. Apresentamos ainda, uma análise crítica a esses mesmos resultados.

5.1 Estimar o valor da habitação numa região dos EUA

Depois de sabermos qual o melhor algoritmo para o nosso problema, é preciso saber quais os atributos necessários para obtermos o melhor resultado possível. Este procedimento tem o nome de *Feature Selection*.

Implementamos dentro do *Feature Selection Loop*, o nodo *X-Partitioner*, seguindo a regra do polegar, com o respetivo *X-Aggregator*. De salientar que, no nodo *Feature Selection Loop Start* optamos pela estratégia *Forward Feature Selection*. Relembramos ainda, que utilizamos o algoritmo *Linear Regression*. No fim do *loop* no nodo *Feature Selection Loop End* optamos por otimizar a métrica *R-Square* (R^2).

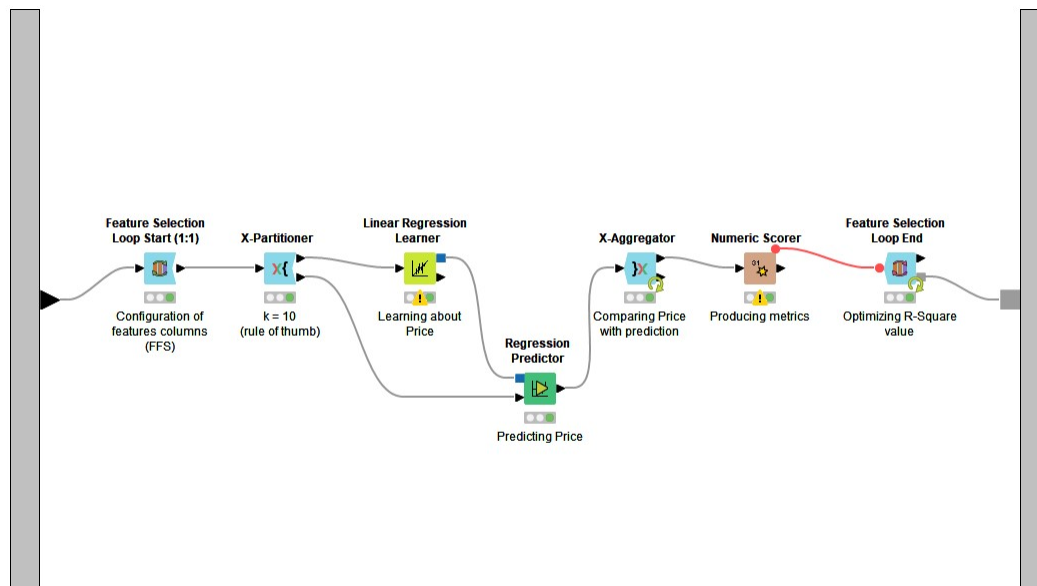


Figura 21. *Feature Selection Loop* do *dataset* recebido

| Optimization Criterion: <i>The score is being maximized.</i> | | |
|--|-----------------|-----------------------------|
| R^2 | Nr. of features | |
| 0,918 | 4 | D Avg. Area Income |
| 0,918 | 5 | D Avg. Area House Age |
| 0,798 | 3 | D Avg. Area Number of Rooms |
| 0,615 | 2 | D Area Population |
| 0,408 | 1 | S Price |
| | | S State |
| | | S Region |

Figura 22. Atributos selecionados do *dataset* recebido

Com os atributos selecionados, passamos para a obtenção do resultado final. Da mesma forma que no procedimento explicado acima, utilizamos o *X-Partitioner* e o *X-Aggregator* para a partição dos dados. Por fim, através do nodo *Numeric Scorer* obtivemos uma tabela com todas as métricas calculadas e através do nodo *Scatter Plot* um gráfico que compara os resultados previstos com os corretos.

Exibimos a seguir a tabela dos resultados finais e o gráfico.

| | |
|---------------------------------|-------|
| R ² : | 0,918 |
| Mean absolute error: | 0,033 |
| Mean squared error: | 0,002 |
| Root mean squared error: | 0,041 |
| Mean signed difference: | 0 |
| Mean absolute percentage error: | NaN |
| Adjusted R ² : | 0,918 |

Figura 23. Resultado obtido com *Linear Regression*

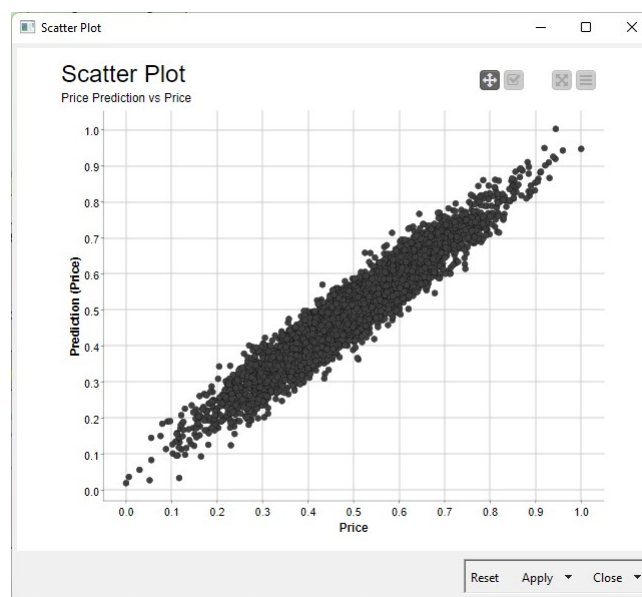


Figura 24. *Scatter Plot* do resultado obtido do *dataset* recebido

Com os resultados obtidos, o grupo considera que a preparação de dados efetuada foi ótima, conseguindo obter valores extremamente positivos, por exemplo, o valor do *R-Square* atingido foi 0.918 bastante perto do desejado 1. Mesmo os valores nas métricas de erro obtidas são extremamente baixos. Apesar disso, o valor do *R-Square* não é 1, e portanto, algo poderia ter sido melhorado durante a preparação para melhorar o resultado obtido.

5.2 Prever a temperatura do solo no mundo, por região

Tal como no *dataset* anterior, depois de descobrirmos o melhor algoritmo para o nosso problema, passamos para o procedimento *Feature Selection*. No *loop* deste método, no nodo *Feature Selection Loop Start*, decidimos optar pela estratégia *Backward Feature Elimination*, experimentando assim duas estratégias diferentes. Utilizamos, mais uma vez, os nodos de partição de dados *X-Partitioner* e o respetivo *X-Aggregator*, mas desta vez com um $K = 2$, pois, uma vez que o nosso *dataset* tem muitas linhas, o esforço computacional seria gigante se colocássemos em prática a regra do polegar ($K = 10$). No fim deste *loop*, no nodo *Feature Selection Loop End*, optamos por otimizar a métrica *Accuracy*.

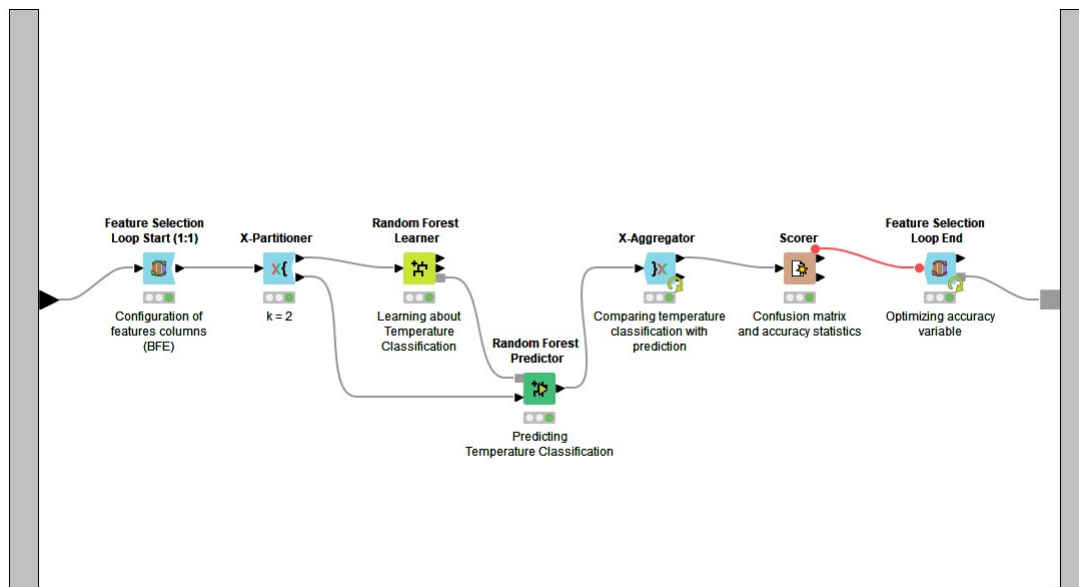


Figura 25. *Feature Selection Loop* do *dataset* escolhido

| Optimization Criterion: <i>The score is being maximized.</i> | | |
|--|-----------------|-------------------------------|
| Accuracy | Nr. of features | dt |
| 0,777 | 4 | AverageTemperatureUncertainty |
| 0,777 | 6 | Country |
| 0,777 | 5 | Year |
| 0,774 | 7 | Month (number) |
| 0,748 | 3 | Region |
| 0,741 | 2 | Continent |
| 0,619 | 1 | Temperature Classification |

Figura 26. Atributos selecionados do *dataset* escolhido

Já com os melhores atributos selecionados, passamos para a obtenção do resultado final. Utilizamos mais uma vez o *X-Partitioner*, desta vez com a regra do polegar, e o *X-Aggregator*. Para concluirmos, falta apenas visualizar o resultado. Para tal utilizamos o nodo *Scorer (JavaScript)*, com o qual adquirimos uma matriz de confusão com todas as precisões para as diferentes classificações e também a precisão final.

Exibimos a seguir a matriz de confusão com os resultados finais e algumas tabelas adicionais com algumas estatísticas.

Confusion Matrix

Scorer View

Confusion Matrix

| | Cold (Pr... | Cool (Pr... | Hot (Pre... | Very Col... | Very Hot... | Warm (P... | |
|-------------|-------------|-------------|-------------|-------------|-------------|------------|--------|
| Cold (Ac... | 945 | 1694 | 240 | 208 | 54 | 523 | 25.79% |
| Cool (Ac... | 368 | 17148 | 2261 | 117 | 133 | 10036 | 57.04% |
| Hot (Act... | 89 | 1131 | 237104 | 67 | 15298 | 13325 | 88.80% |
| Very Col... | 199 | 545 | 29 | 713 | 66 | 27 | 45.16% |
| Very Hot... | 46 | 168 | 28319 | 77 | 45376 | 486 | 60.93% |
| Warm (A... | 110 | 6202 | 21194 | 19 | 328 | 68167 | 70.99% |
| | 53.78% | 63.78% | 82.00% | 59.37% | 74.08% | 73.64% | |

Overall Statistics

| Overall Accuracy | Overall Error | Cohen's kappa (κ) | Correctly Classified | Incorrectly Classified |
|------------------|---------------|----------------------------|----------------------|------------------------|
| 78.14% | 21.86% | 0.630 | 369453 | 103359 |

Reset Apply Close

Figura 27. Matriz de confusão

| Row ID | I True Positives | I False Positives | I True Negatives | I False Negatives | D Recall | D Precision | D Sensitivity | D Specificity | D F-meas... |
|-----------|------------------|-------------------|------------------|-------------------|----------|-------------|---------------|---------------|-------------|
| Cold | 945 | 812 | 468336 | 2719 | 0.258 | 0.538 | 0.258 | 0.998 | 0.349 |
| Cool | 17148 | 9740 | 433009 | 12915 | 0.57 | 0.638 | 0.57 | 0.978 | 0.602 |
| Hot | 237104 | 52043 | 153755 | 29910 | 0.888 | 0.82 | 0.888 | 0.747 | 0.853 |
| Very Cold | 713 | 488 | 470745 | 866 | 0.452 | 0.594 | 0.452 | 0.999 | 0.513 |
| Very Hot | 45376 | 15879 | 382461 | 29096 | 0.609 | 0.741 | 0.609 | 0.96 | 0.669 |
| Warm | 68167 | 24397 | 352395 | 27853 | 0.71 | 0.736 | 0.71 | 0.935 | 0.723 |

Figura 28. Resultado do modelo do *dataset* escolhido

| Row ID | D Overall Accuracy | D Overall Error | D Cohen's kappa | I Correctly Classified | I Incorrectly Classified |
|---------|--------------------|-----------------|-----------------|------------------------|--------------------------|
| Overall | 0.781 | 0.219 | 0.63 | 369453 | 103359 |

Figura 29. Estatísticas finais obtidas do *dataset* escolhido

Já com os resultados obtidos, e apesar de terem sido abaixo dos 80%, o grupo considera o resultado bastante positivo, devido ao facto do nosso *dataset* inicialmente ter muito pouca informação e ser constituído por perto de 580000 linhas. Assim, o resultado obtido reforça que a preparação de dados realizada foi apropriada ao problema.

6 Sugestões e Recomendações

Por último, decidimos realizar uma análise dos resultados obtidos no contexto do problema de cada um dos *datasets* e aos modelos finais desenvolvidos. Assim, conseguimos obter algumas sugestões e recomendações sobre como obter os melhores resultados para os problemas em questão.

6.1 Estimar o valor da habitação numa região dos EUA

No problema recebido propunha-se a estimação do valor da habitação numa região dos Estados Unidos da América, como tal o objetivo é obter um valor de *R-Square* perto de 1.

Durante toda a realização deste *dataset* e de termos feito diferentes preparações de dados e testado diferentes algoritmos, o grupo obteve os melhores resultados quando removeu a coluna *Avg. Area Number of Bedrooms* e normalizou todos os valores do tipo *Double*. Desta forma, a métrica *R-Square* foi maximizada e os erros calculados foram mínimos. Das novas colunas adicionadas, apenas o atributo *State* foi utilizado para o cálculo do resultado final e como tal, a coluna *Region* não necessita de ser construída.

Para os algoritmos, o grupo sugere a utilização da *Linear Regression* para a *Feature Selection* e para a validação do modelo, pois foi o que obteve os melhores resultados.

6.2 Prever a temperatura do solo no mundo, por região

Para o problema que escolhemos é proposta a previsão da temperatura do solo no mundo, por região. Sendo assim, o objetivo é obter a previsão mais alta possível.

Ao longo de toda a realização do problema do *dataset* escolhido, e depois de várias alternativas examinadas para a preparação de dados e para a sua validação, o grupo atingiu os melhores resultados quando removeu todos os *missing values* e removeu os *outliers* do atributo *AverageTemperatureUncertainty*. Todas as colunas que foram adicionadas foram fundamentais para aumentar a quantidade de informação que o *dataset* continha, mas a partir do *Feature Selection*, deduzimos que o atributo *Continent* não é necessário, e este apenas é sugerido para obter algumas estatísticas interessantes que o *dataset* consegue produzir.

Em termos de algoritmos, o grupo recomenda a utilização do *Random Forest* tanto para a *Feature Selection* e para a validação do modelo. No entanto, o algoritmo *Decision Tree* tem uma precisão ligeiramente inferior e é computacionalmente mais leve, sendo este também um bom algoritmo para ser utilizado.

7 Conclusão

Dado por concluído o trabalho prático de ADI, consideramos importante realçar todos os pontos positivos e negativos, e ainda, efetuar uma análise crítica final do trabalho realizado.

A extensa preparação de dados feita em ambos os *datasets* destaca o bom entendimento dos problemas com o objetivo de melhorar o resultado final em mente. Destacamos também a boa organização dos modelos desenvolvidos com todos os nodos comentados. O grupo ter optado por desenvolver um modelo, para o *dataset* escolhido, com um problema diferente do recebido pelo enunciado, é também um ponto positivo do trabalho realizado.

Apesar de termos optado por um problema diferente para cada um dos *datasets*, o grupo poderia ter escolhido um sistema de aprendizagem diferente, especialmente as redes neurais ou mesmo aprendizagens sem supervisão com a utilização dos *clusters*.

O grupo quer ainda salientar, que este projeto ajudou a consolidar toda a matéria lecionada e que foi utilizada durante o desenvolvimento do mesmo, desde a análise inicial até à escolha dos algoritmos de validação dos modelos.

Por fim, o grupo considera que o trabalho realizado é bastante positivo, pois cumpre com todos os requisitos propostos no enunciado.