



Universidade do Minho

MACHINE LEARNING: FUNDAMENTOS E APLICAÇÕES

CONCEÇÃO E IMPLEMENTAÇÃO DE MODELOS DE MACHINE LEARNING BASEADOS EM ÁRVORES



Cristina Mendes **PG42576**



Maria Pires **A86268**



Matilde Silva **PG42584**



Pedro Machado **A83719**

25 DE NOVEMBRO DE 2020

Resumo

O presente relatório foi elaborado no âmbito da disciplina *Sistemas Baseados em Similaridade*, pertencente ao perfil *Machine Learning: Fundamentos e Aplicações*, no qual será exposto todo o processo de tratamento de dados, concepção e otimização de modelos baseados em árvores, com recurso à plataforma *Knime*. Deste modo, primeiro será efetuada uma introdução ao projeto, bem como uma descrição dos seus objetivos. Em segundo lugar, é introduzido o primeiro conjunto de dados que trata a previsão do número de incidentes rodoviários. Seguidamente, é tratado um conjunto de dados sobre o desempenho académico de estudantes. Estas duas secções são compostas pela introdução e explicação dos dados que compõem o conjunto, a sua visualização gráfica, o seu pré-processamento, modelação e, finalmente, otimização. De notar que todas as secções possuem uma análise crítica.

Conteúdo

1	Introdução	5
2	Previsão do número de incidentes rodoviários	6
2.1	<i>Features</i> do dataset	6
2.2	Visualização e análise dos dados	7
2.3	Pré-processamento de dados	11
2.4	Modelação	15
2.5	Otimização	16
2.6	Análise de resultados	17
3	Students' Academic Performance Dataset	18
3.1	<i>Features</i> do conjunto de dados	18
3.2	Visualização e análise de dados	19
3.3	Pré-processamento de dados	23
3.4	Modelação	26
3.5	Otimização	28
3.6	Análise de resultados	30
4	Conclusão	31

Lista de Figuras

2.1	Distribuição das features	8
2.2	Análise de várias <i>features</i> em relação aos incidentes	9
2.3	Tabelas com a média da hora de acordo com as estradas afetadas	9
2.4	Pie chart com as ruas afetadas	10
2.5	Nodos para obter a correlação e a <i>feature selection</i>	12
2.6	Correlação dos atributos do <i>dataset</i>	12
2.7	Uso de um algoritmo genético para a seleção de <i>features</i>	13
2.8	Várias combinações de <i>features</i>	13
2.9	Remoção de algumas <i>features</i>	14
2.10	Aggregação das ruas com menor valor informativo	14
2.11	Transformação da humidade para valores entre 0 e 1	14
2.12	Estado final do <i>dataset</i> após tratamento de dados	15
2.13	Modelo preditivo com <i>Random Forests</i>	15
2.14	Hiper-parâmetros do <i>Tuning</i>	16
2.15	Nodos utilizados para o <i>tuning</i> do modelo de <i>Random Forest Tree</i>	16
2.16	Melhores resultados das combinações do <i>tuning</i>	16
2.17	Exemplo de um dos resultados obtidos	17
3.1	Distribuição das features do conjunto de dados	20
3.2	<i>Box Plot</i> dos dados numéricos do conjunto	21
3.3	Consulta de recursos por turma	21
3.4	<i>Pie Charts</i> para diversas <i>features</i> , por género	22
3.5	<i>Scatter Plot</i> com o registo das mãos levantadas por disciplina	22
3.6	Definições aplicadas ao nodo <i>Column Filter</i>	23
3.7	Normalização de dados numéricos	23
3.8	Conversão da <i>feature Class</i>	24
3.9	Metanodo <i>Análise de Dados</i>	24
3.10	Análise da satisfação dos pais cujos filhos têm local de nascença e nacionalidade diferentes	25
3.11	Análise do resultado dos alunos por nacionalidade e género	25
3.12	Resultado do nodo <i>Backward Feature Elimination</i>	26
3.13	<i>Random Forest</i>	26
3.14	Definições aplicadas ao nodo <i>Random Forest</i>	27
3.15	Matriz de confusão	28
3.16	Metanodo <i>tuning</i>	28
3.17	Configurações obtidas após otimização	28
3.18	Modelação otimizada da <i>Random Forest</i>	29
3.19	Matriz de confusão após otimização	29
3.20	<i>Scorer View</i>	30

3.21 <i>ROC Curve</i>	30
---------------------------------	----

Lista de Tabelas

2.1	Tipos de dados do conjunto	7
3.1	Tipos de dados do conjunto	19

1 | Introdução

O objetivo deste trabalho é a concepção e desenvolvimento de um projeto de *Machine Learning*, com recurso à plataforma *Knime* e modelos baseados em árvores que, na primeira parte, seja capaz de prever a ocorrência incidentes rodoviários na cidade de Braga, e na segunda parte, que seja capaz de prever a satisfação dos pais face à escola onde estudam os seus filhos. Posto isto, serão desenvolvidos dois modelos de aprendizagem através da análise de dois conjuntos de dados distintos.

A estrutura do relatório cumpre as etapas que constituem o processo de extração de conhecimento de um conjunto de dados. Na primeira secção, para cada conjunto de dados, são apresentadas as diferentes *features* que o constituem e são analisados os seus tipos de dados e significado. De seguida, são expostos e analisados diferentes gráficos relativos a estas características. Na terceira parte, é demonstrado o pré-processamento de dados realizado, esclarecendo todos os passos do processo. Finalmente, nas seções de modelação e otimização, são testados modelos de aprendizagem distintos, é descrito o processo de otimização do modelo e apresentado o resultado ótimo. Posto isto, é apresentada uma análise crítica aos resultados obtidos bem como sugestões e recomendações face a estes.

2 | Previsão do número de incidentes rodoviários

Este *dataset* refere-se aos incidentes rodoviários em Braga durante o ano de 2019. É pretendido analisar os dados de forma a obter conclusões e informações relevantes.

Assim, o objetivo final, com o uso deste *dataset*, é a previsão da ocorrência de incidentes. Conseguimos, então, prever para qualquer outro dia qual o grau de incidentes que poderão ocorrer numa dada região.

Com isto tornamos o uso do nosso modelo generalizado para qualquer zona do país e em qualquer dia, sendo que apenas necessita do conjunto de dados similar aos fornecidos.

É necessário também realçar que irá ser usado um dos datasets para treino, enquanto que outro servirá para testar o modelo final, prevendo o target (não presente nestes dados). O *dataset* que foi facultado para a realização deste trabalho tem uma extensão de aproximadamente 1 ano (do dia 15 de Janeiro de 2019 pelas 19h até ao dia 31 de Dezembro de 2019 pelas 23h).

2.1 *Features* do dataset

As *features* presentes são:

- **city_name** - nome da cidade em causa;
- **record_date** - o timestamp associado ao registo;
- **magnitude_of_delay** - magnitude do atraso provocado pelos incidentes que se verificam no record_date correspondente;
- **delay_in_seconds** - atraso, em segundos, provocado pelos incidentes que se verificam no record_date correspondente;
- **affected_roads** - estradas afectadas pelos incidentes que se verificam no record_date correspondente;
- **luminosity** - o nível de luminosidade que se verificava na cidade de Braga;
- **avg_temperature** - valor médio da temperatura para o record_date na cidade de Braga;
- **avg_atm_pressure** - valor médio da pressão atmosférica para o record_date na cidade de Braga;

- **avg_humidity** - valor médio da humidade para o record_date na cidade de Braga;
- **avg_wind_speed** - valor médio da velocidade do vento para o record_date na cidade de Braga;
- **avg_precipitation** - valor médio de precipitação para o record_date na cidade de Braga;
- **avg_rain** - avaliação qualitativa do nível de precipitação para o record_date na cidade de Braga;
- **accidents** - indicação acerca do nível de incidentes rodoviários que se verificam no record_date correspondente na cidade de Braga.

Feature	Tipo
city_name	Nominal
record_date	Date
magnitude_of_delay	Nominal
delay_in_seconds	Numeric
affected_roads	Nominal
luminosity	Nominal
avg_temperature	Numeric
avg_atm_pressure	Numeric
avg_humidity	Numeric
avg_wind_speed	Numeric
avg_precipitation	Numeric
avg_rain	Nominal
accidents	Nominal

Tabela 2.1: Tipos de dados do conjunto

2.2 Visualização e análise dos dados

A primeira etapa para este caso de estudo foi a visualização e análise dos valores das *features* presentes no *dataset*, de maneira a que seja possível identificar padrões e situações anormais (por exemplo, *missing values*) que requerem tratamento, para que não tenham, futuramente, impacto negativo no modelo que estamos a desenvolver.

Primeiramente optamos por utilizar o nodo *Statistics*. Através dele conseguimos observar vários gráficos de barras (Figura 2.1), otimizando assim a leitura destas variáveis. Também conseguimos adquirir conhecimentos sobre os extremos e a média destas. De realçar que a utilização do *statistics* verificou-se que o *dataset* não continha valores em falta.

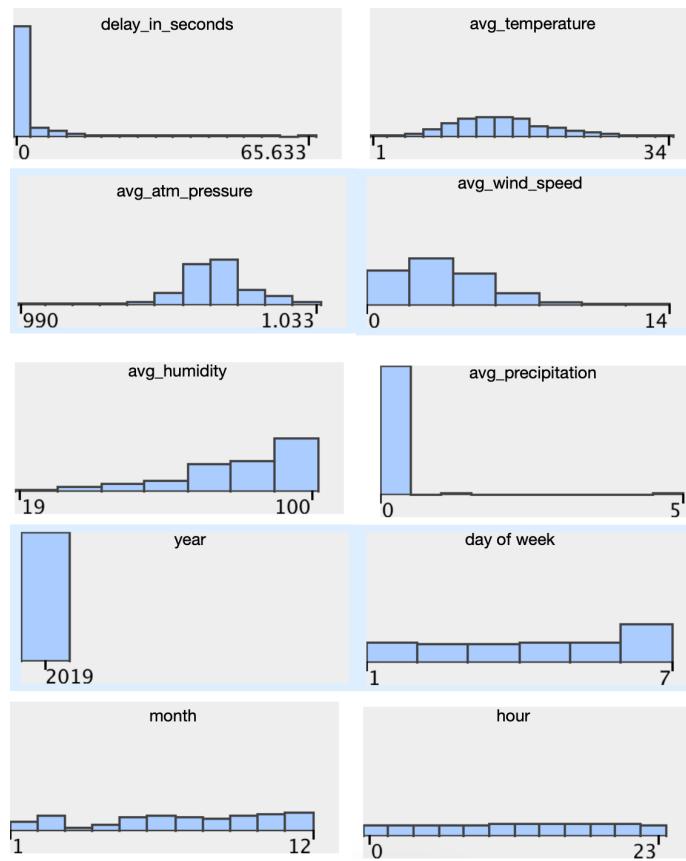


Figura 2.1: Distribuição das features

Através de um *Bar Chart* é possível fazer uma visualização dos dados. A coluna em análise é a dos *accidents* e as variáveis em estudo são *avg_temperature*, *avg_humidity*, *avg_atm_pressure*, *avg_wind_speed* e *avg_precipitation*.

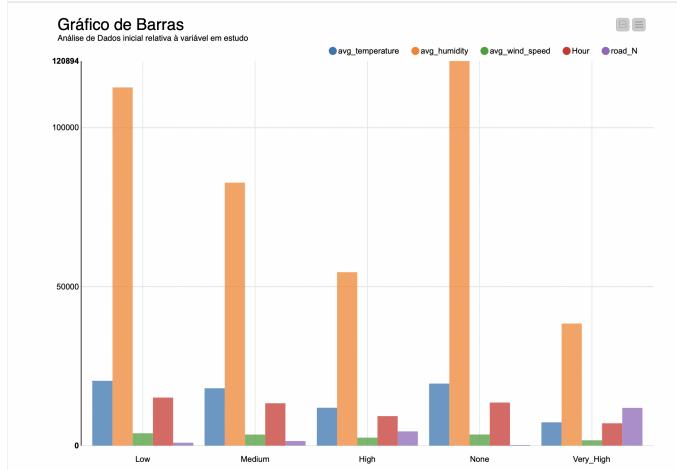


Figura 2.2: Análise de várias *features* em relação aos incidentes

Através de nodos *GroupBy* é possível fazer uma leitura de dados agrupando uma ou mais colunas, geralmente para aplicar algum tipo de função de agregação.

Row ID	road_N	another_roads	Mean(Hour)	Row ID	road_N	another_roads	Mean(Hour)
Row0	0	0	11	Row267	44	2	18
Row1	0	1	12	Row268	44	3	17
Row2	0	2	10	Row269	44	12	17
Row3	0	3	14	Row270	44	13	17
Row4	0	4	9	Row271	45	2	17
Row5	0	5	14	Row272	45	3	17
Row6	0	6	0	Row273	45	4	17
Row7	1	0	13	Row274	45	6	17
Row8	1	1	13	Row275	45	11	17
Row9	1	2	13	Row276	46	5	17
Row10	1	3	13	Row277	46	9	18
Row11	1	4	16	Row278	47	0	17
Row12	1	6	15	Row279	47	4	17
Row13	2	0	10	Row280	47	6	18
Row14	2	1	10	Row281	48	3	18
Row15	2	2	13	Row282	48	6	15
Row16	2	3	13	Row283	49	4	18
Row17	2	4	15	Row284	49	8	17
Row18	2	5	15	Row285	49	11	17
Row19	3	0	14	Row286	50	2	18
Row20	3	1	17	Row287	50	3	19
Row21	3	2	12	Row288	51	9	17
Row22	3	3	14	Row289	52	8	18
Row23	3	4	14	Row290	52	9	18
Row24	3	5	13	Row291	53	3	18
Row25	4	0	15	Row292	53	7	17
Row26	4	1	16	Row293	53	9	18
Row27	4	2	12	Row294	55	8	17
Row28	4	3	14	Row295	55	11	18
Row29	4	4	14	Row296	56	7	19
Row30	4	10	18	Row297	57	1	18
Row31	5	0	15	Row298	59	8	16
Row32	5	1	10	Row299	60	3	18
Row33	5	2	11	Row300	60	4	17
Row34	5	3	13	Row301	60	6	18
Row35	5	4	16	Row302	65	11	17
Row36	5	5	13	Row303	67	3	18

Figura 2.3: Tabelas com a média da hora de acordo com as estradas afetadas

Na figura anterior é possível observar as colunas *road_N* (estradas nacionais), *another_roads* (outras ruas existentes em Braga) e a *hora* dos incidentes que afetam as ruas anteriormente descritas.

Assumindo que a hora de ponta na cidade de Braga é entre as 17h e as

18h, conseguimos aferir que na primeira tabela, quando existem poucas ruas afetadas por causa dos incidentes, o horário da ocorrência geralmente é fora desse intervalo. Já na segunda tabela, encontramos os valores mais elevados de estradas afetadas e curiosamente, corresponde aos horários onde existe maior afluência de carros nas estradas.

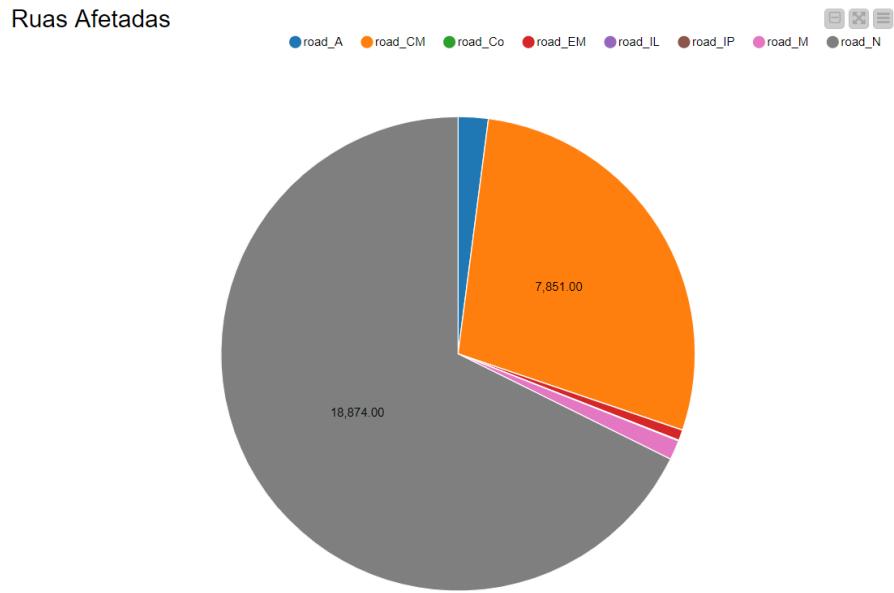


Figura 2.4: Pie chart com as ruas afetadas

Com este gráfico, é possível verificar que estradas nacionais e as municipais (em particular as CM), são as estradas com maior ocorrência no *dataset*. Deste modo, podemos concluir que os incidentes ocorridos afetarão com maior probabilidade este tipo de estradas. Este valor deve-se sobretudo à grande extensão destas estradas, pelo que, um incidente irá provocar várias referências à mesma estrada e, por isso mesmo, a ocorrência superior destas estradas é natural.

2.3 Pré-processamento de dados

Para otimizar o rendimento de um modelo de *Machine Learning* é necessário fazer um tratamento ao conjunto de dados de forma a explorar as *features* mais relevantes para a obtenção de informação.

Este processo revelou-se como sendo um dos mais importantes. Deste modo, necessitou de mais atenção no desenvolvimento do trabalho devido à influência que tem na fase de previsão e análise de resultados. Assim, é pretendido tratarmos e analisarmos os dados de todas as *features* com exceção da *accidents* visto que é a nossa variável *target*.

Observando o *dataset* inicial, é possível aferir que a *feature* *city_name* não possui qualquer valor informativo para o modelo preditivo.

De seguida, e devido ao facto que a *feature* *record_date* possui a informação relativa ao tempo cronológico estar num formato de texto, e para que se possa extraír o conhecimento dela, fizemos uma conversão do formato *String* para o formato *DateTime*. Deste modo, conseguimos extraír os diferentes campos nele contidos, como o ano, o mês, o dia da semana e a hora. Simultaneamente, foi removida a coluna *Year* por ser sempre o mesmo ano (2019) e *record_date*, pois já foi processada.

Após a remoção destas duas *features*, foi necessário trabalhar no *affected_roads* visto que se encontrava com uma *string* de forma muito simplista com a enumeração de todas ruas afetadas, sendo que muitas delas estavam repetidas. Por isso mesmo, foi efetuado um *Java Snippet* com vista a tornar este atributo em vários, cada um específico a um tipo de estrada. Neste *dataset*, eram referidas estradas Nacionais (N), Autoestradas (A), Estradas Municipais (M, EM e CM), Itinerários (IP e IL) e os *County* (referidos no tratamento de dados como Co).

Esta última estrada não foi encontrada durante a nossa pesquisa, pelo que temos dúvidas da validade destas estradas no *dataset*. Contudo, a utilização deste tipo de estradas tornou o modelo mais eficiente, pelo que não as removemos.

Depois da exclusão de *features* devido à existência de apenas uma classe (atributos como *city_name* e *Year*), é necessário realizar *feature selection* de forma a remover os atributos menos relevantes para o modelo preditivo.

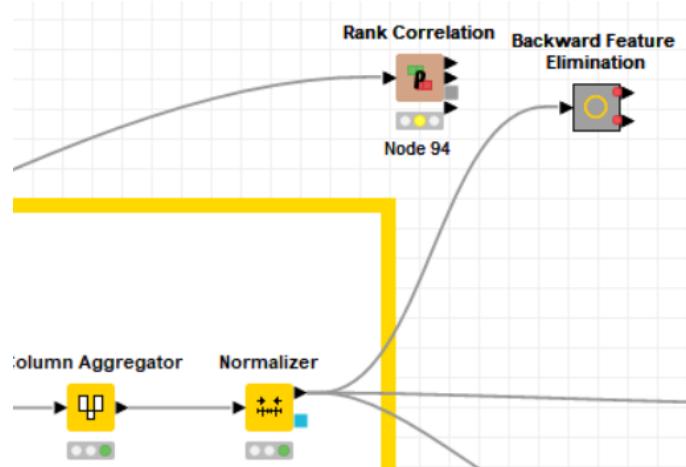


Figura 2.5: Nodos para obter a correlação e a *feature selection*

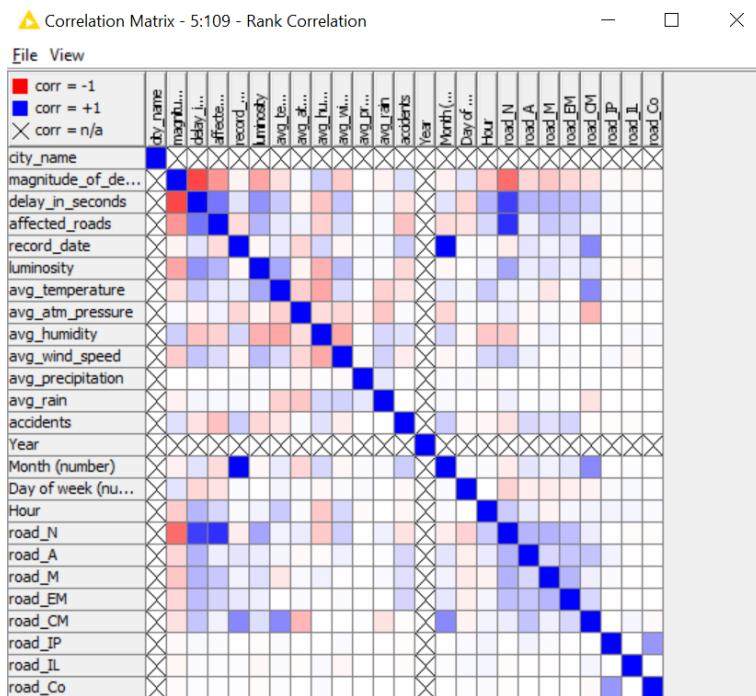


Figura 2.6: Correlação dos atributos do *dataset*

A análise de correlação tem por objetivo medir o grau de relacionamento entre variáveis. Quando a correlação é perto de 1 ou -1 , esta é considerada positiva forte ou negativa forte, respectivamente. Através deste método, verificámos que atributos como a *avg_precipitation*, *avg_atm_pressure* e *avg_rain* têm uma correlação quase nula (são variáveis não-correlacionadas) com a *feature accidents*. Com isto, podemos ponderar em remover estas *features* numa perspetiva de aumentar a *accuracy* do modelo.

Para além do uso de nodos que possibilitam a visualização da correlação dos atributos, podemos usar o *Backward Feature Elimination* de forma a indicar-nos quais as *features* com maior relevância para o modelo. De notar que neste nodo foi usado o **Genetic Algorithm** de forma a otimizar os controladores (*features* do *dataset*).

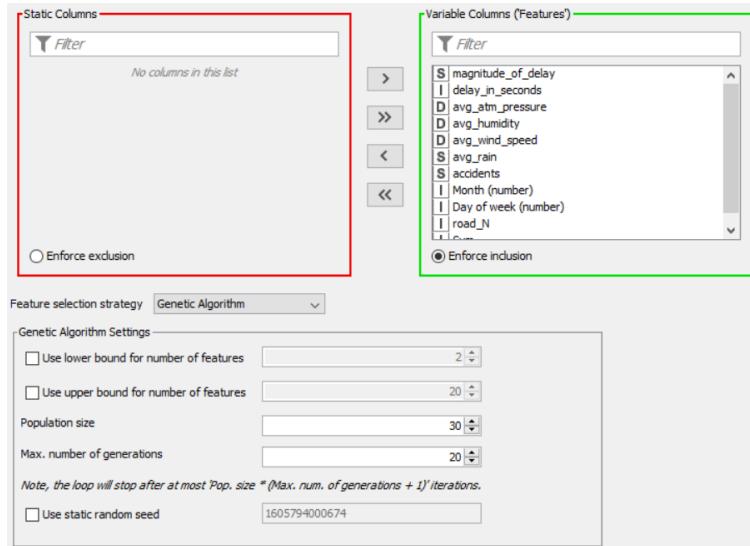


Figura 2.7: Uso de um algoritmo genético para a seleção de *features*

Optimization Criterion: The score is being maximized.		
Accuracy	Nr. of features	
0,891	7	S magnitude_of_delay
0,889	8	I delay_in_seconds
0,886	6	D avg_atm_pressure
0,884	6	D avg_humidity
0,884	6	D avg_wind_speed
0,879	5	S avg_rain
0,878	7	S accidents
0,878	6	I Month (number)
0,878	6	I Day of week (number)
0,878	6	I road_N
0,873	8	I Sum

Figura 2.8: Várias combinações de *features*

Após estes dois métodos de seleção de atributos, filtraram-se então as seguintes *features*¹:

¹De notar que muitas foram removidas devido aos vários testes realizados e devido à performance do modelo obtido.

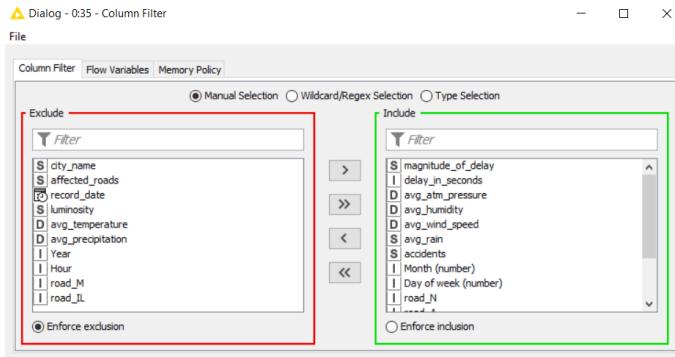


Figura 2.9: Remoção de algumas *features*

Para além da remoção das *features* previamente referidas, mostrou-se necessário agregar o *count* das estradas de forma a aumentar a eficiência destes parâmetros no modelo preditivo. Deste modo, estas *features* foram agrupadas da seguinte forma, visto que, após várias tentativas, demonstrou ser o melhor tratamento.

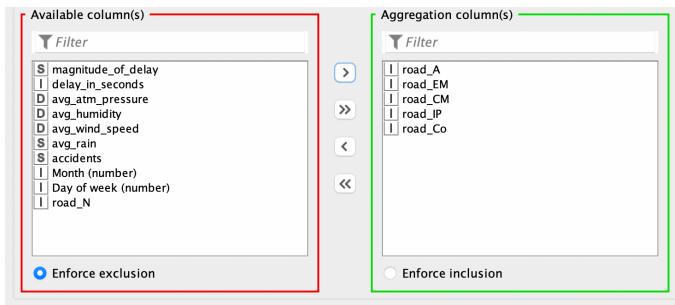


Figura 2.10: Agregação das ruas com menor valor informativo

Após a análise da *feature avg_humidity*, verificou-se que esta consistia na percentagem da humidade no ar na altura que foi registada. Por isso mesmo, usou-se um *Normalizer* com vista a ficar com valores compreendidos entre 0 e 1.

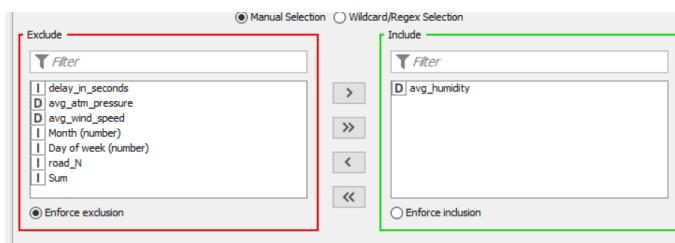


Figura 2.11: Transformação da humidade para valores entre 0 e 1

Desta maneira, o tratamento de dados fica completo e tem-se os dados do *dataset* prontos a serem utilizados pelo modelo preditivo.

Row ID	S magnitude_of_delay	I delay_in_seconds	D avg_mm_pressures	D avg_mm_pressures	D avg_mm_pressures	S avg_mm_pressures	D avg_mm_pressures	S avg_mm_pressures	I accidents	I accidents	I Month (number)	I Day of week (number)	I road_N	I another_roads
Row1	UNDEFINED	401	1.025	0.321	5	Sem Chuva	Medium	2	4	3	0			
Row2	MAJOR	1.027	0.327	0.42	4	Sem Chuva	High	9	2	0	2			
Row3	UNDEFINED	3812	1.013	1	6	Sem Chuva	High	6	2	6	3			
Row4	MAJOR	498	1.014	0.778	6	Sem Chuva	Medium	5	3	4	2			
Row5	MAJOR	13770	1.023	0.738	5	Sem Chuva	Low	5	6	0	2			
Row6	UNDEFINED	0	1.017	0.914	2	Sem Chuva	None	9	1	0	2			
Row7	MAJOR	1264	0	0.667	4	Sem Chuva	Medium	8	4	0	3			
Row8	UNDEFINED	0	1.022	0.792	5	Sem Chuva	High	2	4	2	0			
Row9	MAJOR	23804	1.029	0.654	2	Sem Chuva	Very,High	5	1	47	0			
Row10	MAJOR	867	1.023	1	1	Sem Chuva	Medium	9	2	1	2			
Row11	MAJOR	315	1.017	0.778	2	Sem Chuva	Low	10	4	0	3			
Row12	UNDEFINED	0	1.019	0.654	3	Sem Chuva	None	10	6	0	2			
Row13	UNDEFINED	0	1.020	0.84	5	Sem Chuva	High	10	6	0	2			
Row14	MAJOR	2714	0	0.778	3	Sem Chuva	High	10	4	6	3			
Row15	MAJOR	1634	1.023	0.748	3	Sem Chuva	High	9	4	2	2			
Row16	MAJOR	9298	1.012	0.753	2	Sem Chuva	Very,High	11	5	23	5			
Row17	UNDEFINED	0	1.015	1	1	Sem Chuva	None	7	3	0	2			
Row18	MAJOR	2059	1.021	0.605	2	Sem Chuva	High	10	4	4	4			
Row19	MAJOR	658	1.007	0.84	6	Com Chuva	None	4	3	2	0			
Row20	UNDEFINED	0	1.024	0.614	3	Com Chuva	None	12	1	0	0			
Row21	UNDEFINED	0	1.016	0.444	6	Sem Chuva	Low	5	7	0	0			
Row22	MAJOR	262	1.021	0.605	4	Sem Chuva	None	5	6	0	0			
Row23	UNDEFINED	343	1.023	1	0	Sem Chuva	Medium	7	1	1	2			
Row24	UNDEFINED	0	1.025	0.914	1	Sem Chuva	None	1	1	0	0			
Row25	MAJOR	10998	1.025	0.84	1	Sem Chuva	Very,High	2	3	21	0			
Row26	MAJOR	0	1.020	0.778	6	Sem Chuva	Medium	6	0	0	2			

Figura 2.12: Estado final do *dataset* após tratamento de dados

De notar que este tratamento de dados será completamente igual no *dataset* de teste e treino.

2.4 Modelação

Depois de termos o pré-processamento de dados realizado, é imperativo utilizar modelos de *Machine Learning* de forma a prever a variável *target*, neste caso o grau de incidentes (**accidents**).

Neste trabalho era pretendido a utilização de modelos baseados em árvores como as *Decision Trees* e *Random Forest Trees*.

Em primeiro lugar, foi usado uma *Decision Tree* visto que foram os modelos de *Machine Learning* mais usados nos trabalhos práticos individuais. Contudo, após algumas submissões na plataforma da competição, verificou-se que as *Random Forest* permitiam obter uma *accuracy* superior, tal como era previsto devido ao grande número de árvores que implementa com muita diversidade entre elas.

Em segundo lugar, após a escolha do modelo a ser usado, começou-se a usar uma partição do *dataset* de treino com vista a ter uma noção da *accuracy* do modelo. Desta forma, este será treinado com todos os dados de treino para preverem o grau de incidentes nos dados de teste.

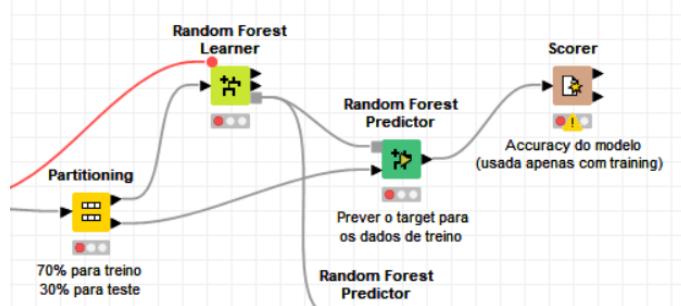


Figura 2.13: Modelo preditivo com *Random Forests*

2.5 Otimização

De forma a otimizar o modelo preditivo, usou-se *tuning* dos hiper-parâmetros. Neste caso, como foram usadas *Random Forest Trees*, os hiper-parâmetros usados incidiram no *Split Criterion*, número mínimo do tamanho dos nodos, número máximo de níveis das árvores e, por último, o número de árvores utilizadas no modelo.

Parameter	Start value	Stop value	Step size	Integer?
number_levels	13	25	1.0	<input checked="" type="checkbox"/>
node_size	2	7	1.0	<input checked="" type="checkbox"/>
number_trees	300	1,000	100.0	<input checked="" type="checkbox"/>

Figura 2.14: Hiper-parâmetros do *Tuning*

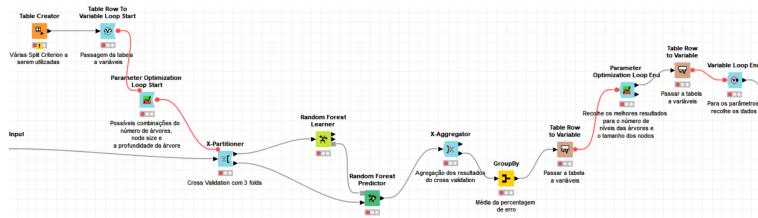


Figura 2.15: Nodos utilizados para o *tuning* do modelo de *Random Forest Tree*

Após a realização das várias iterações usadas neste *tuning*, obtiveram-se as 3 melhores combinações que os hiper-parâmetros geravam na *accuracy* dos modelos.

Row ID	number...	node_size	number...	D	Objective value	I	current...	S	Split
Row0	25	4	700	0.907563025210...	0	Gini			
Row2	20	3	800	0.906962785114...	2	Information...			
Row1	15	3	400	0.905162064825...	1	Information...			

Figura 2.16: Melhores resultados das combinações do *tuning*

De seguida, podemos escolher a melhor combinação e utilizá-la no modelo preditivo final. Assim, conseguimos fazer a otimização do modelo e prever a *feature accidents* com uma melhor *accuracy* na teoria.

2.6 Análise de resultados

Ao longo das várias submissões efetuadas, notamos nas implementações a nível de tratamento de dados, modelo e otimização que melhoraram o *score* significativamente.

Em primeiro lugar, a troca de *Decision Trees* para *Random Forest Trees* permitiu-nos obter uma melhoria significativa na *accuracy* final. Assim, o uso de grandes quantidades de árvores com um nível de diversidade entre elas, teve um peso importante nas diferenças das primeiras submissões.

Em segundo lugar, a remoção de *features* através da análise da matriz de correlação e de *feature selection* através de algoritmos genéticos, teve um peso considerável no aumento da precisão do modelo de *Machine Learning*. De realçar que com estas duas alterações no tratamento de dados e na modelação foi obtida uma *accuracy* “interna” (no *KNIME*) por volta dos 88%.

Por último, a otimização do modelo através do *tuning* permitiu-nos a obter uma *accuracy* ligeiramente superior, ficando situada entre os 89 e 90%. ²

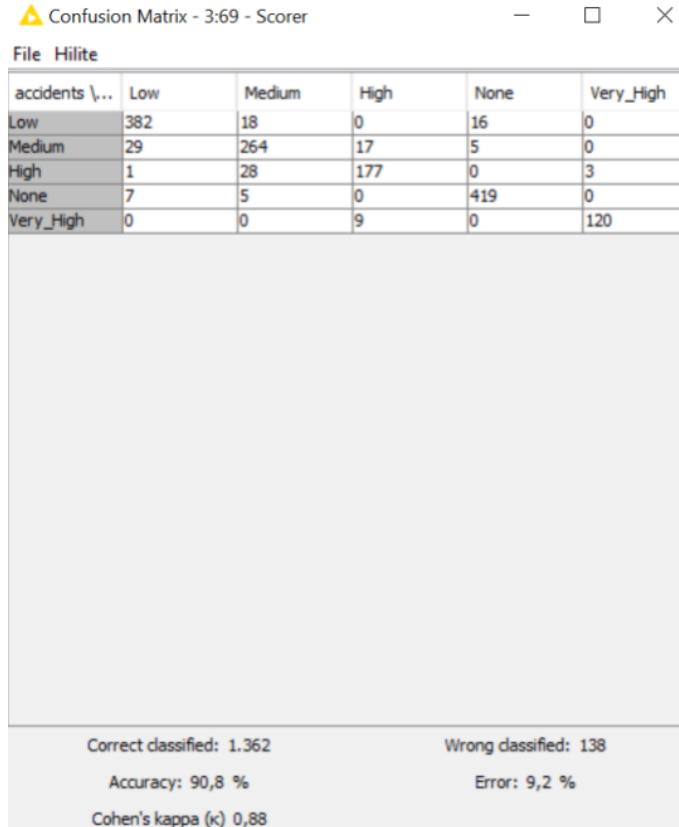


Figura 2.17: Exemplo de um dos resultados obtidos

Como é possível observar, o nosso modelo permite alcançar uma precisão acima dos 90% para certas partições do *dataset* de treino, pelo que poderíamos melhorar o tratamento de dados do modelo num trabalho futuro.

²De realçar que esta precisão é comprovada pelos valores obtidos no *tuning* (Figura 2.16)

3 | Students' Academic Performance Dataset

3.1 *Features* do conjunto de dados

As características que constituem o *dataset* [2] relativo à performance académica de um conjunto de alunos são as seguintes:

- **Gender** - o género do aluno;
- **Nationality** - a nacionalidade do aluno;
- **PlaceofBirth** - local de nascimento do aluno;
- **StageID** - ciclo de estudos em que o aluno se encontra;
- **GradeID** - ano em que o aluno se encontra matriculado;
- **SectionID** - turma a que o aluno pertence;
- **Topic** - tema da disciplina;
- **Semester** - semestre do ano escolar;
- **Relation** - encarregado de educação;
- **RaisedHands** - número de vezes que o aluno levantou a mão/participou nas aulas;
- **VisitedResources** - número de vezes que o aluno acedeu aos recursos do curso;
- **AnnouncementsView** - número de vezes que o aluno leu os anúncios do curso;
- **Discussion** - número de vezes que o aluno participou em grupos de discussão;
- **ParentAnsweringSurvey** - se os pais responderam a questionários enviados pela escola;
- **ParentschoolSatisfaction** - grau de satisfação dos pais com a escola;
- **StudentAbsenceDays** - número de dias que o aluno esteve ausente;
- **Class** - A nota dos estudantes é classificada em três intervalos numéricos baseado na sua nota final (Low-Level, Medium-Level, High-Level).

Feature	Tipo
Gender	Nominal
Nationality	Nominal
PlaceofBirth	Nominal
StageID	Nominal
GradeID	Nominal
SectionID	Nominal
Topic	Nominal
Semester	Nominal
Relation	Nominal
Raisedhands	Numeric
VisitedResources	Numeric
AnnouncementsView	Numeric
Discussion	Numeric
ParentAnsweringSurvey	Nominal
ParentSchoolSatisfaction	Nominal
StudentAbsenceDays	Nominal
Class	Nominal

Tabela 3.1: Tipos de dados do conjunto

3.2 Visualização e análise de dados

O estudo do conjunto de dados foi realizado com recurso a nodos de exploração visual que permitem a análise dos valores de cada *feature* e facilitam a identificação de padrões e situações anormais, tais como *outliers* e *missing values*, para que estes possam ser futuramente tratados e não tenham impacto na solução final.

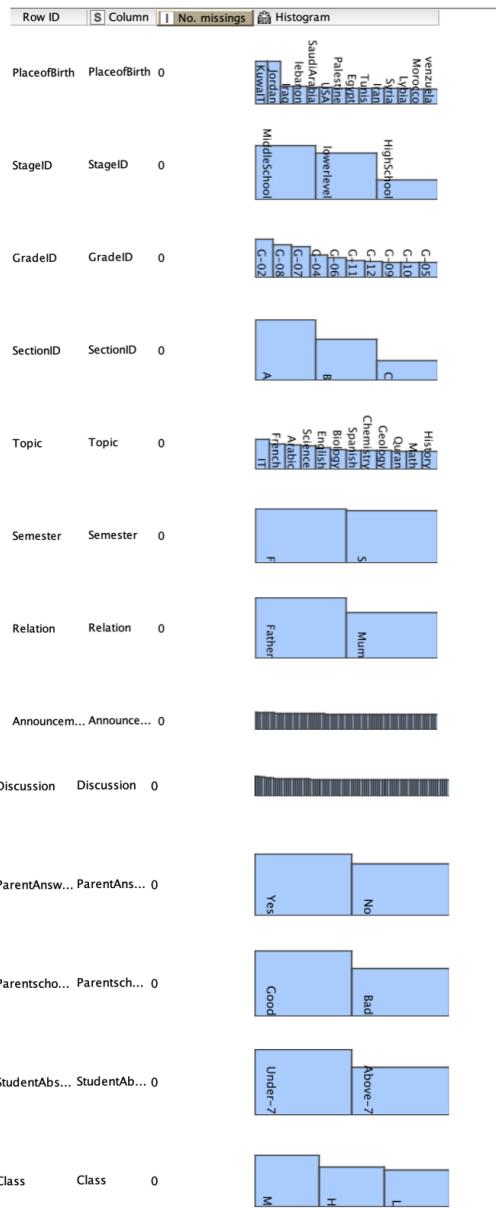


Figura 3.1: Distribuição das features do conjunto de dados

Como podemos observar pela Figura anterior não existem *missing values* no conjunto de dados em tratamento.

Através do nodo *Box Plot* é possível verificar rapidamente a existência de *outliers*. Na Figura 3.2 é possível constatar que os dados do tipo numérico do conjunto não apresentam grande dispersão.

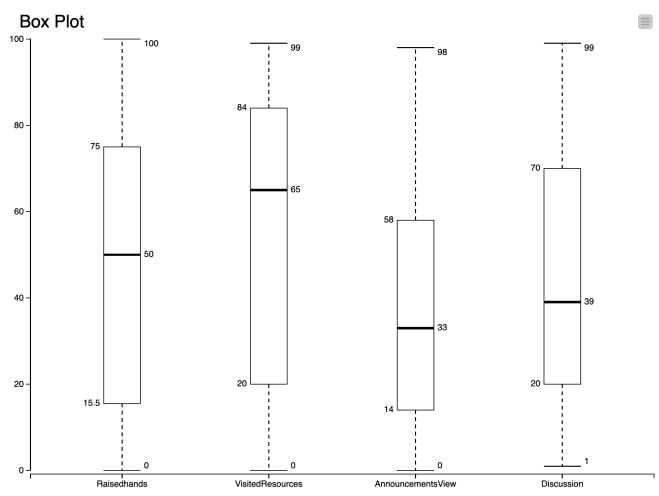


Figura 3.2: *Box Plot* dos dados numéricos do conjunto

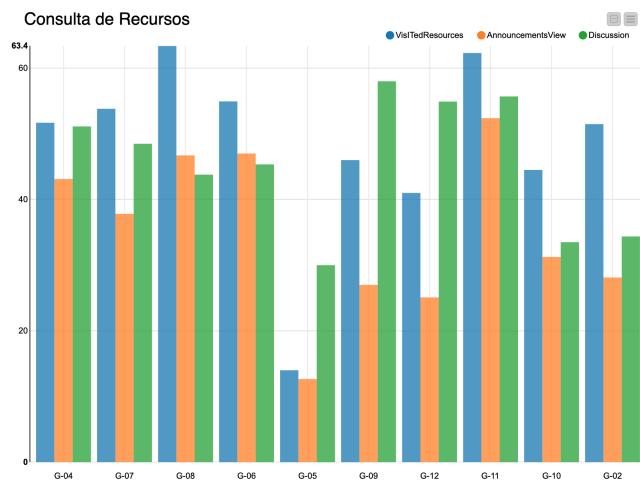


Figura 3.3: Consulta de recursos por turma

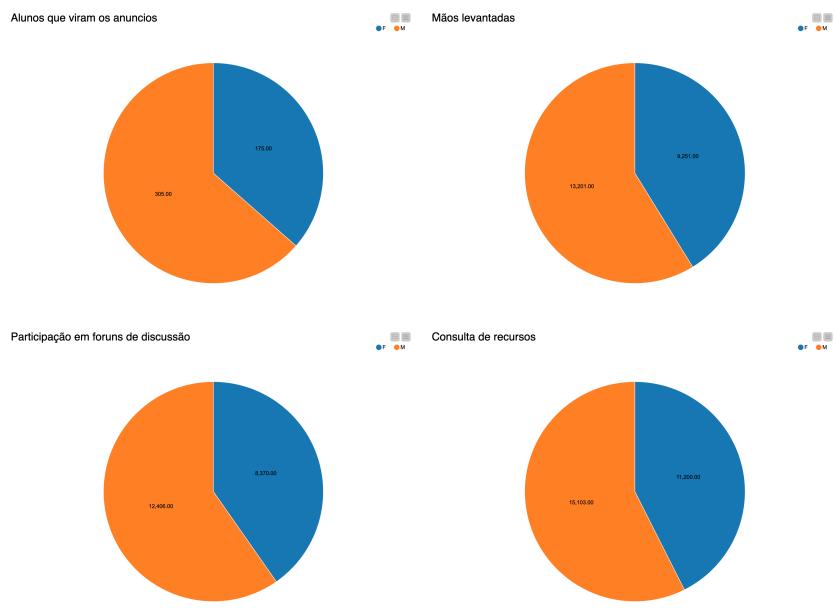


Figura 3.4: *Pie Charts* para diversas *features*, por género

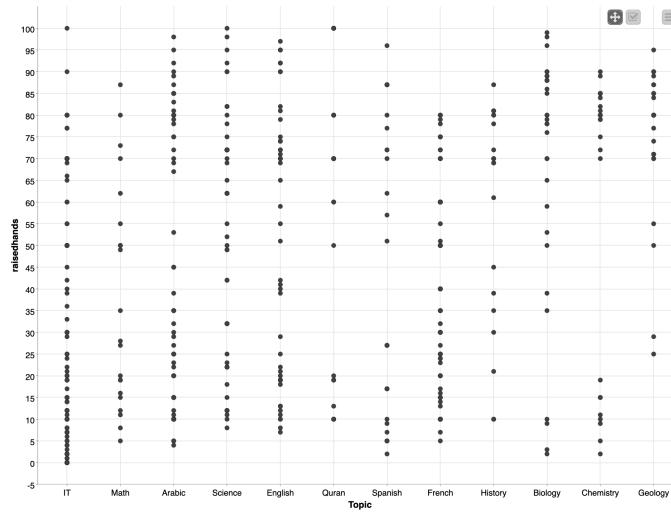


Figura 3.5: *Scatter Plot* com o registo das mãos levantadas por disciplina

3.3 Pré-processamento de dados

O primeiro passo no tratamento de dados foi uniformizar os nomes das colunas e alguns valores, por motivos estéticos e facilidade de leitura, pois este tratamento não se reflete nos resultados finais. Com o intuito de atingir a melhor precisão possível foram removidas as *features* *GradeID* e *SectionID* visto que o ano e a turma do aluno apresentam pouca influência no algoritmo de previsão do grau de satisfação dos pais com a escola.

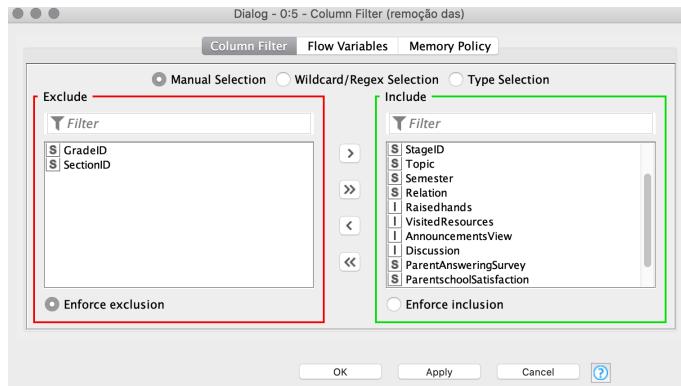


Figura 3.6: Definições aplicadas ao nodo *Column Filter*

De modo a evitar o enviesamento de algumas colunas o grupo decidiu normalizar as *features* numéricas do conjunto de dados entre os valores 0 e 1, como se pode verificar na Figura 3.7.

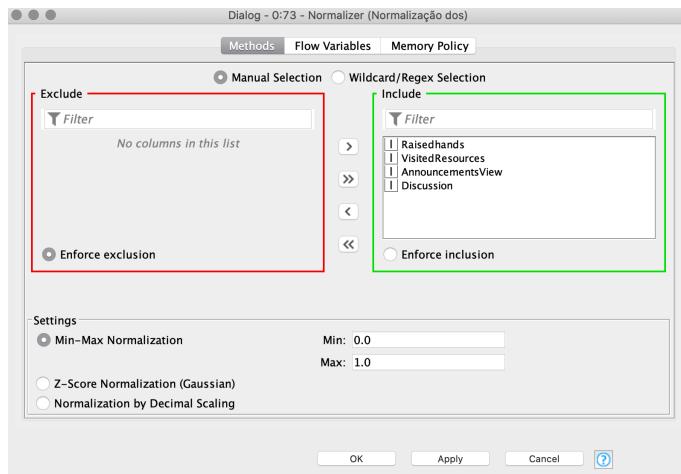


Figura 3.7: Normalização de dados numéricos

A coluna *Relation* indica-nos se o encarregado de educação do aluno é a mãe ou o pai. Esta *feature* que era inicialmente do tipo nominal foi mapeada para um inteiro através do nodo *Category To Number*.

O conjunto de valores possíveis para a *feature* *Class* corresponde a três letras

distintas. A partir dos nodos *Cell Replacer*, *Table Creator* e *String to Number* foram convertidos estes três valores nominais em três valores numéricos.

Row ID	column1	column2
Row0	L	0
Row1	M	1
Row2	H	2

Figura 3.8: Conversão da *feature Class*

Adicionalmente, separou-se o tratamento em dois metanodos: *Tratamento de Dados* que contempla a análise apresentada anteriormente e cujo objetivo é tratar os dados para prever a satisfação dos pais dos alunos.

O segundo metanodo, *Análise de Dados*, tem o objetivo de compreender e relacionar os restantes dados do conjunto e extrair conclusões adicionais à previsão da satisfação dos pais.

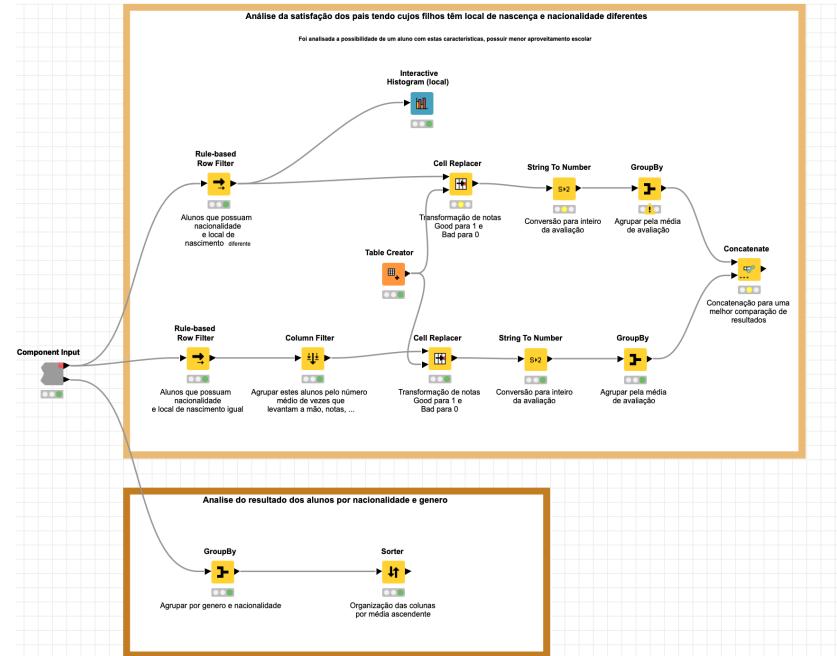


Figura 3.9: Metanodo *Análise de Dados*

Foi analisada a possibilidade de um aluno com local de nascença e nacionalidade diferentes possuir menor aproveitamento escolar e pode concluir-se pela Figura 3.10 que estas características não são significativas no aproveitamento escolar.

Row ID	D	Mean(ParentschoolSatisfaction)	I	Count*(ParentschoolSatisfaction)
Row0	0.673		55	
Row0_dup	0.6		425	

Figura 3.10: Análise da satisfação dos pais cujos filhos têm local de nascença e nacionalidade diferentes

Na tabela seguinte podemos observar a média dos resultados escolares dos alunos por nacionalidade e género e aferir aqueles que apresentam um melhor desempenho.

Row ID	S	Gender	S	Nationality	D	Mean(Class)	I	Count*(Class)
Row6	F		Lybia	0		2		
Row19	M		Lybia	0		4		
Row14	M		Iran	0.6		5		
Row23	M		Syria	0.6		5		
Row20	M		Morocco	0.667		3		
Row17	M		Kuwait	0.712		125		
Row13	M		Egypt	0.714		7		
Row18	M		Lebanon	0.833		6		
Row24	M		Tunis	0.909		11		
Row16	M		Jordan	0.914		93		
Row1	F		Iran	1		1		
Row11	F		Tunis	1		1		
Row25	M		USA	1		2		
Row4	F		Kuwait	1.074		54		
Row22	M		Saudi Arabia	1.286		7		
Row3	F		Jordan	1.304		79		
Row21	M		Palestine	1.4		20		
Row0	F		Egypt	1.5		2		
Row12	F		USA	1.5		4		
Row8	F		Palestine	1.5		8		
Row15	M		Iraq	1.625		16		
Row2	F		Iraq	1.667		6		
Row5	F		Lebanon	1.727		11		
Row9	F		Saudi Arabia	1.75		4		
Row7	F		Morocco	2		1		
Row10	F		Syria	2		2		
Row26	M		Venezuela	2		1		

Figura 3.11: Análise do resultado dos alunos por nacionalidade e género

No passo final do pré-processamento de dados aplicamos o nodo *Backward Feature Elimination* para perceber quais as *features* importantes para o modelo de previsão e a sua precisão.

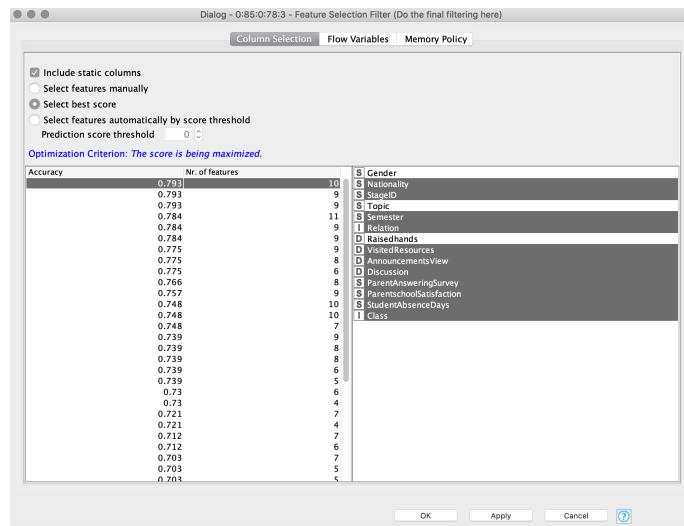


Figura 3.12: Resultado do nodo *Backward Feature Elimination*

3.4 Modelação

O modelo baseado em árvores de decisão escolhido para tratar os dados foi o *Random Forest*. Pode observar-se na Figura 3.13 o fluxo de modelação. O conjunto de dados foi partido de modo aleatório em duas partes, 70% para treino e 30% para teste.

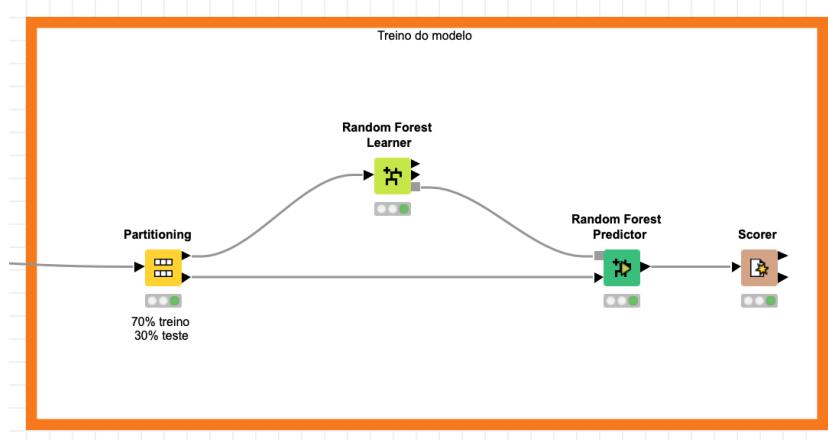


Figura 3.13: *Random Forest*

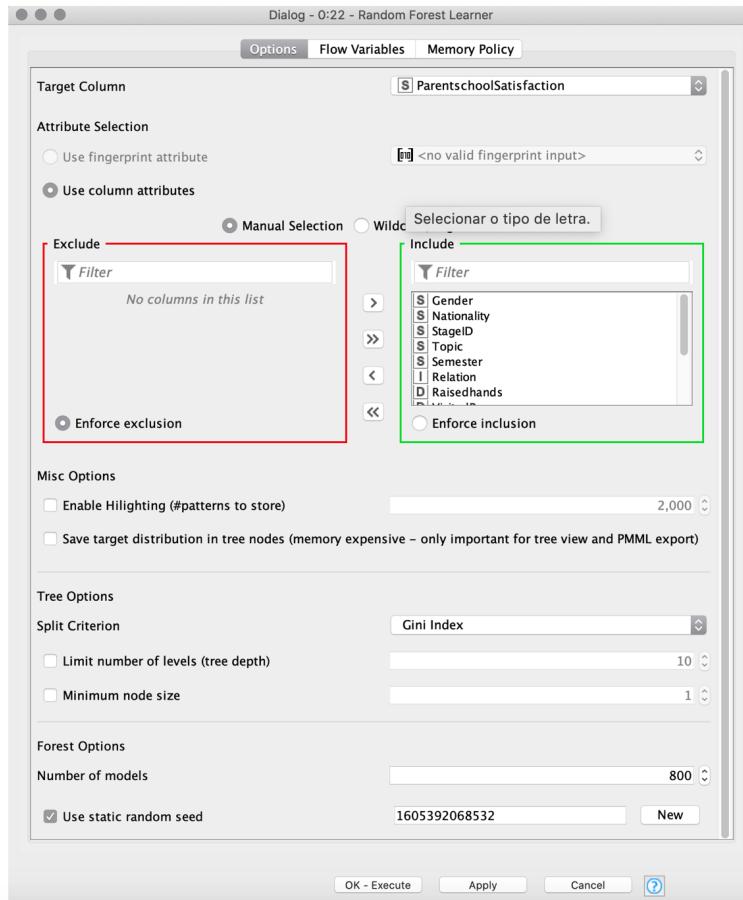


Figura 3.14: Definições aplicadas ao nodo *Random Forest*

Analizando a matriz de confusão apresentada na Figura 3.15 podemos constatar que o modelo apresenta uma precisão de 84.028%. De notar que foram classificados de forma errada 23 registos, sendo 17 destes falsos negativos, isto é, a satisfação dos pais é tida erradamente como boa. Com o intuito de reduzir este erro, será apresentada na secção seguinte a otimização do modelo.

Confusion Matrix - 0:82 - Scorer		
File	Hilite	
Parentscho...	Good	Bad
Good	80	6
Bad	17	41
Correct classified: 121		Wrong classified: 23
Accuracy: 84.028 %		Error: 15.972 %
Cohen's kappa (κ) 0.657		

Figura 3.15: Matriz de confusão

3.5 Otimização

Para maximizar a precisão apresentada anteriormente foi utilizado o seguinte conjunto de nodos (Figura 3.16) para descobrir quais os parâmetros da configuração ótima.

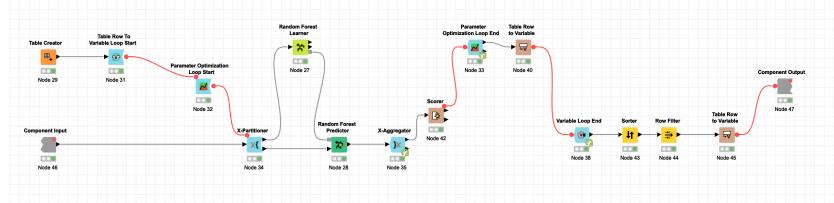


Figura 3.16: Metanodo tuning

Inicialmente foi construído um *Table Creator* com os três valores possíveis para a opção *Split Criterion* do nodo *Random Forest Learner*. Os valores possíveis são *Information Gain Ratio*, *Information Gain* e *Gini Index*. De seguida, ao nodo *Parameter Optimization Loop Start* aplicaram-se três parâmetros:

- *number_levels*: com valores entre 10 e 15;
- *node_size*: com valores entre 3 e 5;
- *number_trees*: com valores entre 300 e 700, num intervalo de 100 em 100

Em cada iteração, para cada árvore, o conjunto de dados é dividido em 10 partes através do método *K-Fold Cross Validation*, sendo k = 10. Dentro dessa iteração, para cada árvore são feitas 9 iterações com o conjunto de treino e 1 com o conjunto de teste, alternando essas 10 partes entre as iterações. Este método permite reduzir o erro e evitar o *overfitting* do modelo.

Row ID	numb...	node_...	numb...	Objective value	currentiteration	Split
Row0	14	3	300	0.831	0	Gini
Row1	13	3	300	0.829	1	InformationGain
Row2	14	3	700	0.835	2	InformationGainRatio

Figura 3.17: Configurações obtidas apóis otimização

Com base nos resultados obtidos o critério de partição *Information Gain Ratio* deveria ser a opção ideal para efetuar o treino do modelo pois apresenta a melhor precisão dos três técnicas testadas. A segunda melhor após a aplicação do *tuning* é o critério *Information Gain* e, por último, com menor precisão, *Gini Index*.

Finalmente, foi aplicado ao fluxo apresentado na Figura 3.18 as características que permitiam otimizar o modelo.

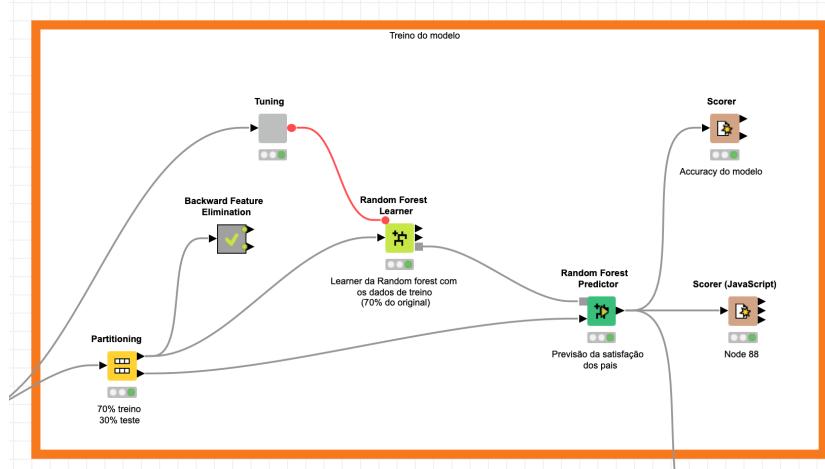


Figura 3.18: Modelação otimizada da *Random Forest*

Como se pode observar pela matriz de confusão apresentada na Figura 3.19 a precisão do modelo, ao contrário do que era esperado, diminuiu. Esta diminuição embora tenha sido inesperada não é um cenário completamente impossível uma vez que o método *K-Fold Cross Validation* particiona o conjunto de dados K vezes e nessas K vezes utiliza K-1 conjuntos distintos para treino.

Confusion Matrix - 0:25 - Scorer		
File Hilite		
Parentscho...	Good	Bad
Good	81	5
Bad	20	38

Correct classified: 119	Wrong classified: 25
Accuracy: 82.639 %	Error: 17.361 %
Cohen's kappa (κ) 0.623	

Figura 3.19: Matriz de confusão após otimização

3.6 Análise de resultados

Na Figura 3.20 apresenta-se a matriz de confusão do modelo de previsão final para a satisfação dos encarregados de educação. A precisão obtida final foi 82,64%. Num universo de 144 registo, 25 foram classificados erradamente. Foram classificados como falsos negativos 20 registos, este valor é o mais problemático uma vez que ter erradamente a satisfação dos encarregados de educação classificada como positiva revela-se pior tendo em conta os objetivos do modelo preditivo. Isto impede as escolas de melhorar e garantir que o rendimento dos seus alunos é o melhor possível.

Scorer View		Confusion Matrix		
		Bad (Predicted)	Good (Predicted)	
Bad (Actual)	Bad (Actual)	38	20	65.52%
	Good (Actual)	5	81	94.19%
		88.37%	80.20%	

Overall Statistics				
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
82.64%	17.36%	0.623	119	25

Figura 3.20: Scorer View

De seguida apresenta-se a *Receiver Operating Characteristic Curve (ROC curve)* que ilustra graficamente a precisão do sistema.

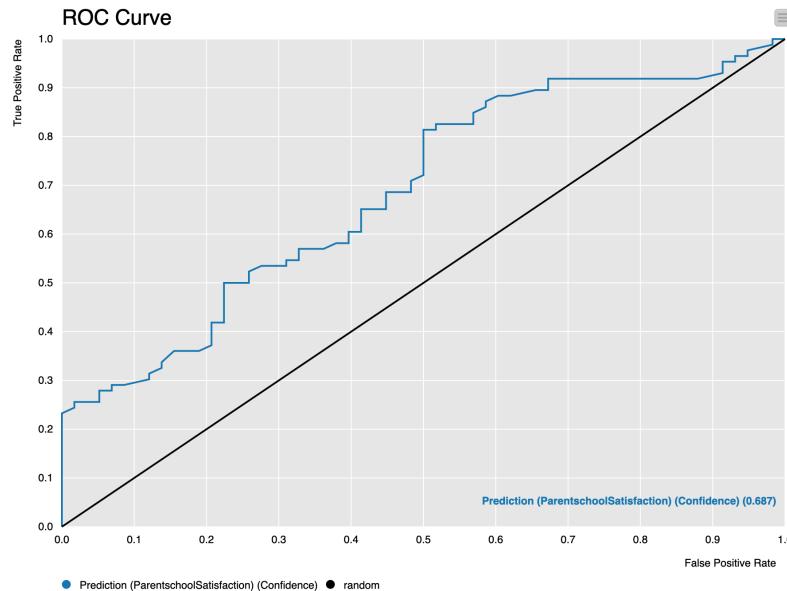


Figura 3.21: ROC Curve

4 | Conclusão

Em suma, na parte inicial deste relatório foi analisado o *dataset* relacionado com os incidentes rodoviários na cidade de Braga. Nesta secção foram tratados e analisados os dados de forma a otimizar o *dataset*.

Foram retiradas e alteradas *features*, aplicou-se um modelo preditivo, que na nossa opinião é o mais vantajoso, e avaliou-se os resultados obtidos. Durante a realização do trabalho, fomos também aprimorando o modelo preditivo através do acréscimo de hiper-parâmetros. Após a análise de dados, conseguimos observar que o nosso modelo tem uma precisão de cerca de 90%. Na segunda parte do relatório, foi realizado o tratamento de dados relativos ao *dataset* cujos dados caracterizam a performance académica de alunos. Nesta análise foram também tratados os dados de forma a otimizar ao máximo o modelo. O resultado obtido através do *tuning* não diferiu muito do resultado alcançado a partir da simples aplicação do modelo preditivo sem a optimização.

Neste trabalho, as dificuldades assentaram principalmente na escolha do tratamento de dados adequado. A escolha dos hiper-parâmetros constituiu parte das dificuldades encontradas. Finalmente, o grupo considera que conseguiu desenvolver modelos com uma previsão razoável. Apesar das dificuldades encontradas, conseguimos otimizar os dados de forma a obter a melhor *accuracy* possível. Nos dois *datasets* o modelo desenhado alcançou os objetivos desejados.

Bibliografia

- [1] Conjunto de dados de previsão do número de incidentes rodoviários
- [2] Students' Academic Performance Dataset