

Universidade do Minho
Departamento de Informática

Mestrado Integrado em Engenharia Informática

Classificadores e Sistemas Conexionistas



Redes Neurais Recorrentes para previsão do fluxo de tráfego rodoviário

Grupo 7
A81765 Joana Matos
A78156 Nuno Silva
A80624 Sofia Teixeira

Braga
Maio, 2020

Conteúdo

1	Objetivos	2
2	Metodologia	2
3	Análise do <i>Dataset</i>	2
3.1	Traffic.Flow_Braga	3
3.2	Traffic.Incidents_Braga	4
3.3	Weather_Braga_Descriptions	6
3.4	Weather_Braga	8
4	Decisões de Conceção	9
4.1	Traffic.Flow_Braga	9
4.2	Traffic.Incidents_Braga	10
4.3	Weather_Braga_Descriptions	10
5	Junção dos Datasets	11
6	Tratamento de Dados	12
6.1	Correlação dos Dados	13
6.2	<i>Label Encoding</i>	15
6.3	Coerência dos Dados	17
6.4	<i>Missing Values</i>	18
6.5	<i>Binning</i>	18
6.6	Preenchimento de <i>timesteps</i>	19
7	Modelo Desenvolvido	20
7.1	Características do treino	20
7.2	<i>Tuning</i> e <i>Cross Validation</i>	21
8	Discussão dos Resultados	21
9	Conclusão	23

1 Objetivos

Uma rede neuronal recorrente (RNN) é uma classe das redes neurais artificiais onde as conexões entre nodos dão à rede memória e a noção de ordem e tempo. Estas redes podem usar o seu estado interno para processar sequências de *inputs*. Uma das variantes desta classe e a que é usada neste trabalho são as redes LSTM (*long short-term memory*) que é capaz de aprender dependências a longo período. O objetivo destas é preservar a estimativa de erro que é propagada ao longo do tempo para, assim, treinarem o modelo de maneira mais aprofundada e durante mais iterações.

O principal objetivo deste trabalho é, usando RNNs, prever o fluxo de tráfego a curto e longo prazo, mais concretamente o **speed.diff**, da cidade de **Braga** de acordo com 4 ficheiros acerca de informações que poderão estar direta ou indiretamente envolvidas no tráfego da cidade.

2 Metodologia

É possível referir alguns aspetos a tratar antes de realizar o trabalho e observando apenas os dados que se têm. Daqui podem-se planear algumas tarefas:

1. decidir que *features* são realmente importantes. Em caso negativo, eliminar as colunas correspondentes;
2. tratar de *missing values*;
3. fazer *one hot encoding* ou *label encoding* a dados que se achem pertinentes;
4. mudar os tipos de *features*. Por exemplo, é do interesse do grupo mudar datas de *strings* para o tipo *Date&Time*;
5. juntar os ficheiros iniciais;
6. definir os *timesteps* adequados;
7. definir a rede neuronal;
8. fazer *tuning* aos parâmetros da rede;
9. comparar resultados.

De notar que esta ordem pode mudar ligeiramente na execução do trabalho.

3 Análise do *Dataset*

Tal como referido anteriormente, este *dataset* contém 4 ficheiros: um com os aspetos referentes ao tráfego e suas ruas; um que refere os incidentes na cidade bem como onde aconteceram; um com informações categóricas da meteorologia e, por último, com a informação numérica da meteorologia na cidade. De seguida, estes ficheiros vão ser descritos com a máxima destreza para ser possível chegar à conclusão de quais terão mais peso na modelação do modelo final.

3.1 Traffic_Flow_Braga

Row ID	S city_name	I road_num	S road_name	S functional_road_class_desc	I current_speed	I free_flow_speed	I speed_diff	I current_travel_time	I free_flow_travel_time	I time_diff	S creation_date
Row0	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 19:05:02.000000
Row1	Braga	2	Avenida Central	Local High Importance Road	4	12	8	458	152	306	2019-01-15 19:05:02.000000
Row2	Braga	3	Rua de Caires	Local Connecting Road	26	36	10	51	37	14	2019-01-15 19:05:02.000000
Row3	Braga	4	N14 Bosch	Other Major Road	19	46	27	118	48	70	2019-01-15 19:05:02.000000
Row4	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 19:25:00.000000
Row5	Braga	2	Avenida Central	Local High Importance Road	4	12	8	458	152	306	2019-01-15 19:25:00.000000
Row6	Braga	3	Rua de Caires	Local Connecting Road	23	35	12	58	38	20	2019-01-15 19:25:00.000000
Row7	Braga	4	N14 Bosch	Other Major Road	24	46	22	93	48	45	2019-01-15 19:25:00.000000
Row8	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 19:45:02.000000
Row9	Braga	2	Avenida Central	Local High Importance Road	3	14	11	610	130	480	2019-01-15 19:45:02.000000
Row10	Braga	3	Rua de Caires	Local Connecting Road	35	35	0	38	38	0	2019-01-15 19:45:02.000000
Row11	Braga	4	N14 Bosch	Other Major Road	34	47	13	66	47	19	2019-01-15 19:45:02.000000
Row12	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 20:05:02.000000
Row13	Braga	2	Avenida Central	Local High Importance Road	8	12	4	229	152	77	2019-01-15 20:05:02.000000
Row14	Braga	3	Rua de Caires	Local Connecting Road	35	35	0	38	38	0	2019-01-15 20:05:02.000000
Row15	Braga	4	N14 Bosch	Other Major Road	36	47	11	62	47	15	2019-01-15 20:05:02.000000
Row16	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 20:25:00.000000
Row17	Braga	2	Avenida Central	Local High Importance Road	12	12	0	152	152	0	2019-01-15 20:25:00.000000
Row18	Braga	3	Rua de Caires	Local Connecting Road	35	35	0	38	38	0	2019-01-15 20:25:00.000000
Row19	Braga	4	N14 Bosch	Other Major Road	47	47	0	47	47	0	2019-01-15 20:25:00.000000
Row20	Braga	1	Avenida da Liberdade	Local High Importance Road	20	20	0	14	14	0	2019-01-15 20:45:01.000000
Row21	Braga	2	Avenida Central	Local High Importance Road	9	12	3	203	152	51	2019-01-15 20:45:01.000000

Figura 1: Visualização do *dataset* Traffic_Flow_Braga.

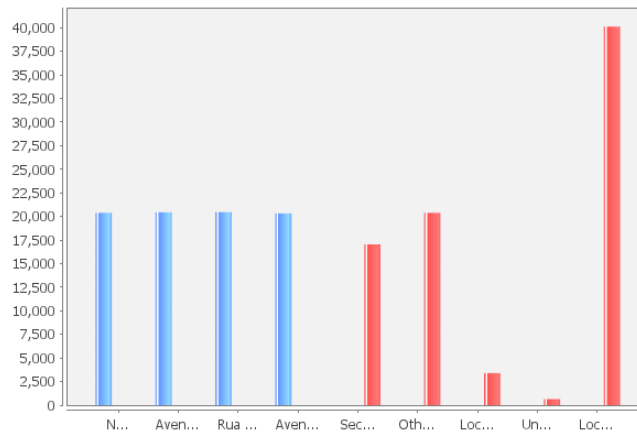
Este ficheiro contém 81600 ocorrências das quais existem as seguintes colunas:

- **city_name**: nome da cidade.
- **road_num**: número atribuído a cada rua. Este valor varia de 1 a 4.
- **road_name**: nome da rua de acordo com o número. Pode ter os nomes "Avenida da Liberdade", "Avenida Central", "Rua de Caires" ou "N14 Bosch", respetivamente.
- **functional_road_class_desc**: classe atribuída a cada rua. Pode ser "Local High Importance Road", "Other Major Road", "Secondary Road", "Local Connecting Road" ou "Unknown Road".
- **current_speed**: velocidade atual do veículo. Pode variar entre 0 Km/h a 48 Km/h.
- **free_flow_speed**: velocidade quando não há tráfego.
- **speed_diff**: diferença entre a velocidade quando não há tráfego e a velocidade atual.
- **current_travel_time**: tempo de viagem atual.
- **free_flow_travel_time**: tempo de viagem quando não há tráfego e tempo de viagem atual.
- **time_diff**: diferença entre tempo de viagem atual.
- **creation_date**: data de criação no formato MM/DD/AAAA HH:MM:SS.

Na figura 2 é possível observar algumas estatísticas dos dados numéricos deste ficheiro. Já na figura 4 observa-se a ocorrência dos atributos "road_name" e "functional_road_class_desc" onde no primeiro pode-se concluir que as quatro ruas, N14 Bosch, Avenida Central, Rua de Caires e Avenida da Liberdade, têm exatamente o mesmo número de ocorrências. Já na função das ruas, tem-se que a maioria especifica o tipo de estradas locais de alta importância. Com menos ocorrências as do tipo outras estradas principais, com ainda menos as estradas secundárias, depois as estradas de conexão local e em minoria estradas desconhecidas.

Nenhuma das colunas, tanto numéricas como nominais, tinham *missing values*.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance
road_num	<input type="checkbox"/>	1	4	2.501	1.117	1.248
current_speed	<input type="checkbox"/>	0	48	27.072	12.215	149.204
free_flow_speed	<input type="checkbox"/>	0	49	30.982	11.950	142.791
speed_diff	<input type="checkbox"/>	0	43	3.910	7.125	50.764
current_travel_time	<input type="checkbox"/>	0	1832	127.439	120.598	14543.762
free_flow_travel_time	<input type="checkbox"/>	0	381	99.157	79.158	6266.034
time_diff	<input type="checkbox"/>	0	1666	28.282	68.286	4663.026

Figura 2: Estatística dos atributos numéricos *dataset* Traffic_Flow_Braga.Figura 3: Estatística dos atributos nominais *dataset* Traffic_Flow_Braga.

3.2 Traffic Incidents Braga

Row ID	S d...	S descript...	S caus...	S from_road	S to_road	S affect...	S incident...	S magnit...	I length...	I delay...	S incident_date	D latitude	D longitude
Row0	Braga	queuing traffic	?	Quinteiro (N103)	Braga-Circular (Ferre...	N103	Jam	Moderate	615	111	2019-01-15 19:05:...	41.537	-8.452
Row1	Braga	queuing traffic	?	Rua Manuel Al...	Rua Da Quinta De Sa...	N14	Jam	Moderate	1008	210	2019-01-15 19:05:...	41.532	-8.44
Row2	Braga	queuing traffic	?	Braga-Circular ...	Avenida General Nort...	N101	Jam	Moderate	615	184	2019-01-15 19:05:...	41.56	-8.419
Row3	Braga	stationary tr...	?	Avenida Gener...	Braga-Circular (N101)	N101	Jam	Major	322	232	2019-01-15 19:05:...	41.558	-8.418
Row4	Braga	stationary tr...	?	Cm1327 (N201)	N201 (Rua De Cima) (...)	N201	Jam	Major	208	235	2019-01-15 19:05:...	41.559	-8.446
Row5	Braga	slow traffic	?	Junction (Brag...	Rua De SãEo Martinh...	?	Jam	Minor	381	28	2019-01-15 19:05:...	41.55	-8.433
Row6	Braga	stationary tr...	?	Avenida Centr...	Avenida General Nort...	N101	Jam	Major	324	275	2019-01-15 19:05:...	41.552	-8.423
Row7	Braga	stationary tr...	?	Rua De Camoe...	Avenida Central (Rua...	?	Jam	Major	308	178	2019-01-15 19:05:...	41.555	-8.417
Row8	Braga	stationary tr...	?	Rua De Sao Do...	N101 (Rua Do Consel...	?	Jam	Major	147	265	2019-01-15 19:05:...	41.558	-8.416
Row9	Braga	stationary tr...	?	SãEo Vicente	Largo das A...	?	Jam	Major	452	183	2019-01-15 19:05:...	41.562	-8.415
Row10	Braga	slow traffic	?	N103 (SãEo V...	Avenida General Nort...	?	Jam	Minor	1744	156	2019-01-15 19:05:...	41.557	-8.407
Row11	Braga	stationary tr...	?	SãEo Vã-tor	Avenida Padre Jã%lo ...	?	Jam	Major	440	281	2019-01-15 19:05:...	41.564	-8.406
Row12	Braga	closed	?	Avenida Douto...	Avenida Robert Smith	?	Road Closed	Undefined	88	0	2019-01-15 19:05:...	41.542	-8.404
Row13	Braga	stationary tr...	?	N103 (Rua Dos...	Braga-Circular (Rua D...	?	Jam	Major	322	223	2019-01-15 19:24:...	41.547	-8.434
Row14	Braga	stationary tr...	?	Avenida Centr...	Avenida General Nort...	N101	Jam	Major	324	277	2019-01-15 19:24:...	41.552	-8.423
Row15	Braga	stationary tr...	?	Avenida Gener...	Braga-Circular (N101)	N101	Jam	Major	322	268	2019-01-15 19:24:...	41.558	-8.418
Row16	Braga	slow traffic	?	Rua Dom Antã...	N101 (Braga Norte) (...)	?	Jam	Minor	732	39	2019-01-15 19:24:...	41.56	-8.407
Row17	Braga	slow traffic	?	Avenida Douto...	Rua De SãEo Josã@/...	?	Jam	Minor	792	69	2019-01-15 19:24:...	41.553	-8.407
Row18	Braga	closed	?	Avenida Douto...	Avenida Robert Smith	?	Road Closed	Undefined	88	0	2019-01-15 19:24:...	41.542	-8.404
Row19	Braga	stationary tr...	?	Avenida Centr...	Avenida General Nort...	N101	Jam	Major	198	448	2019-01-15 19:44:...	41.552	-8.422
Row20	Braga	closed	?	Avenida Douto...	Avenida Robert Smith	?	Road Closed	Undefined	88	0	2019-01-15 19:44:...	41.542	-8.404
Row21	Braga	closed	?	Avenida Douto...	Avenida Robert Smith	?	Road Closed	Undefined	88	0	2019-01-15 20:04:...	41.542	-8.404

Figura 4: Visualização do *dataset* Traffic_Incidents_Braga.

Este ficheiro contém 83347 ocorrências das quais existem as seguintes colunas:

- **city_name**: nome da cidade.
- **description**: descrição do incidente. Pode ser "stationary traffic", "closed", "queuing traffic", "slow traffic", "roadworks", "bridge closed", "tunnel closed" ou "incident".
- **cause_of_incident**: causa do incidente. Pode tomar os valores "new roadworks layout", "roadworks" ou "incident".
- **from_road**: rua inicial do incidente.
- **to_road**: rua final do incidente.
- **affected_roads**: ruas afetadas pelo incidente. Pode tomar vários valores dos 31 possíveis.
- **incident_category_desc**: categoria do incidente. Pode ser devido a "Jam", "Road Closed", "Road Works" ou "Unknown Incident".
- **magnitude_of_delay_desc**: magnitude do atraso. Pode ser "Major", "Undefined", "Moderate", "Minor" ou "Unknown Delay".
- **length_in_meters**: comprimento da rua afetado pelo incidente, em metros.
- **delay_in_seconds**: atraso que o incidente está a provocar, em segundos.
- **incident_date**: data do incidente no formato MM/DD/AAAA HH:MM:SS.
- **latitude**: latitude do incidente. Este valor pode variar de 41.483 a 42.373.
- **longitude**: longitude do incidente. Este valor varia de -87.905 a -8.369.

A partir da figura 5 é possível observar alguns valores referentes às estatísticas dos atributos numéricos. Para estes valores não existem *missing values*.

Já na figura 6 tem-se o gráfico de barras das colunas "incident_category_desc" e "magnitude_of_delay_desc", azul e vermelho, por esta ordem. Na categoria do incidente em maioria é o tráfego. De seguida, é o facto de existirem ruas fechadas, depois são obras na estrada e em minoria por razões desconhecidas. Já na magnitude do incidente temos o número de ocorrências decrescentemente como maior, indefinido, moderado, menor e em minoria por motivo desconhecido. Para os dados nominais, as colunas "cause_of_incident" e "affected_roads" têm 82921 e 42366 *missing values*, respetivamente.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance
+ length_in_meters	<input type="checkbox"/>	14	19670230	2439.608	192698.097	37132556779.180
+ delay_in_seconds	<input type="checkbox"/>	0	5176	143.732	168.624	28433.943
+ latitude	<input type="checkbox"/>	41.483	42.373	41.545	0.018	0.000
+ longitude	<input type="checkbox"/>	-87.905	-8.369	-8.426	1.030	1.062

Figura 5: Estatística dos atributos numéricos do *dataset* Weather_Braga_Descriptions.

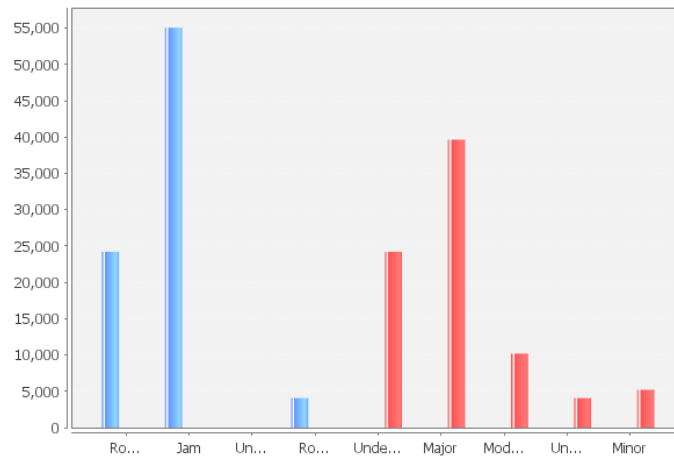


Figura 6: Estatística dos atributos nominais do *dataset* Weather_Braga_Descriptions.

3.3 Weather_Braga_Descriptions

Row ID	S city_name	S cloudiness	S atmosp...	S snow	S thunde...	S rain	S sunrise	S sunset	S creation_date
Row0	Braga	algumas nuvens	N/A	N/A	N/A	N/A	2019-01-15 ...	2019-01-15...	2019-01-15 19:05:00.000000
Row1	Braga	algumas nuvens	N/A	N/A	N/A	N/A	2019-01-15 ...	2019-01-15...	2019-01-15 20:04:59.000000
Row2	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-15 ...	2019-01-15...	2019-01-15 21:04:58.000000
Row3	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-15 ...	2019-01-15...	2019-01-15 22:04:58.000000
Row4	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-15 ...	2019-01-15...	2019-01-15 23:04:58.000000
Row5	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 00:04:59.000000
Row6	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 01:04:58.000000
Row7	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 02:04:58.000000
Row8	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 03:04:58.000000
Row9	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 04:04:58.000000
Row10	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 06:04:59.000000
Row11	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 12:04:58.000000
Row12	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 13:05:00.000000
Row13	Braga	N/A	nÃ@voa	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 14:04:59.000000
Row14	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 15:05:00.000000
Row15	Braga	N/A	nÃ@voa	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 16:04:59.000000
Row16	Braga	N/A	N/A	N/A	N/A	chuva leve	2019-01-16 ...	2019-01-16...	2019-01-16 17:04:59.000000
Row17	Braga	N/A	N/A	N/A	N/A	chuva leve	2019-01-16 ...	2019-01-16...	2019-01-16 18:04:59.000000
Row18	Braga	nuvens quebr...	N/A	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 19:04:59.000000
Row19	Braga	N/A	nÃ@voa	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 20:04:58.000000
Row20	Braga	N/A	neblina	N/A	N/A	N/A	2019-01-16 ...	2019-01-16...	2019-01-16 21:04:59.000000

Figura 7: Visualização do *dataset* Weather_Braga_Descriptions.

Este ficheiro contém 6820 ocorrências das quais existem as seguintes colunas:

- **city_name**: nome da cidade.
- **cloudiness**: característica das nuvens. Pode tomar qualquer um dos 10 nomes disponíveis.
- **atmosphere**: característica da atmosfera naquele dia. Pode tomar um de 7 nomes diferentes.
- **snow**: característica que dita se houve neve ou não. Apenas toma o valor de "N/A".
- **thunderstorm**: característica de tempestades em caso afirmativo. Pode tomar um dos valores "N/A", "trovoada", "trovoada com chuva fraca" ou "trovoada com chuva forte".
- **rain**: característica que refere que tipo de chuva existiu. Pode ser um de 10 nomes possíveis.
- **sunrise**: hora em que se deu o nascer do sol no formato MM/DD/AAAA HH:MM:SS.
- **sunset**: hora em que se deu o pôr do sol no formato MM/DD/AAAA HH:MM:SS.
- **creation_date**: data de criação no formato MM/DD/AAAA HH:MM:SS.

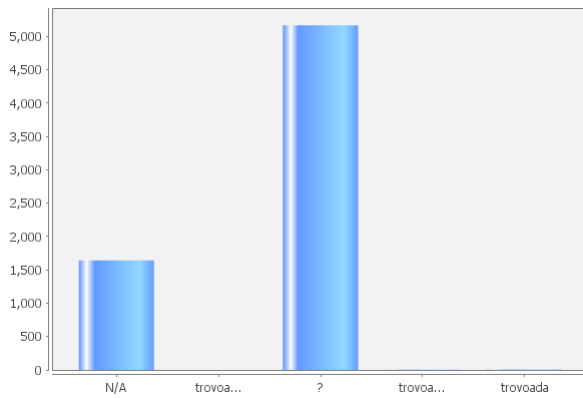


Figura 8: Distribuição para o atributo "thunderstorm".

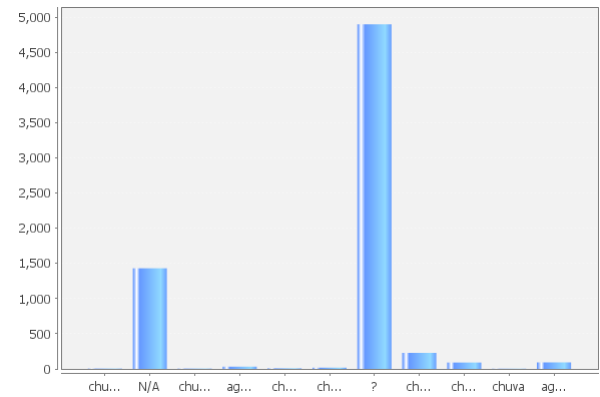


Figura 9: Distribuição para o atributo "rain".

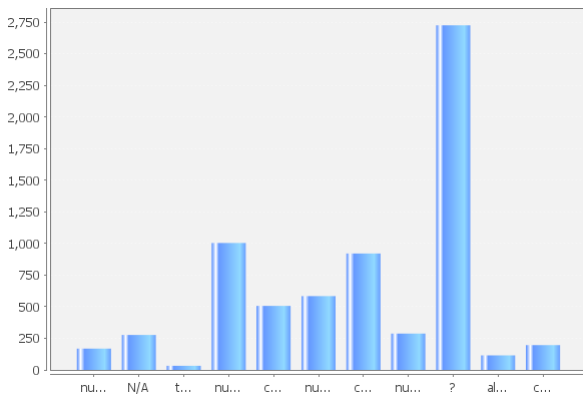


Figura 10: Distribuição para o atributo "cloudiness".

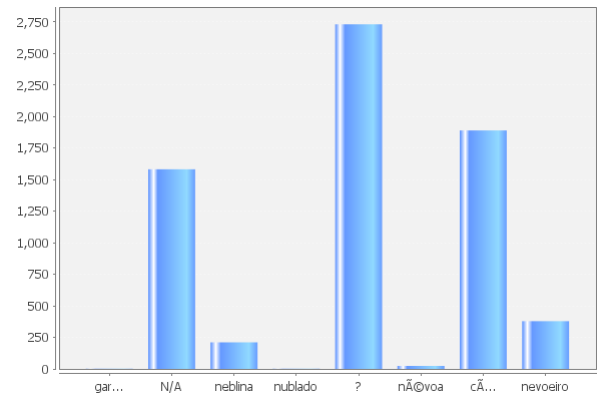


Figura 11: Distribuição para o atributo "atmosphere".

Este ficheiro não contém dados numéricos e os dados nominais existentes, exceto datas ou o nome da cidade, contém todos *missing values*. Para a "cloudiness" existem 2726, na "atmosphere" existem 2731, para "snow" existem 5177, para "thunderstorm" 5166 e, finalmente, para "rain" existem 4898 *missing values*.

Analisando a figura 8, a maioria dos valores são *missing values*. De seguida existem as linhas N/A. Com pouca frequência tem-se trovoada, trovoada com chuva fraca e trovoada com chuva forte.

No atributo da figura 9, também tem a maioria com *missing values* seguido de N/A. De seguida, tem-se chuva fraca, aguaceiros fracos, chuva moderada, aguaceiros, chuva leve, chuva forte, chuveiro fraco, chuveiro e chuva fraca e chuva, por esta ordem de frequência.

Para a figura 10, mais uma vez a maioria são *missing values*. De seguida, de maior frequência para o menor, tem-se nuvens quebradas, céu pouco nublado, nuvens dispersas, céu claro, nublado, N/A, céu limpo, nuvens quebrados, algumas nuvens e, finalmente, tempo nublado.

Para finalizar, na figura 11, a maioria da frequência encontra-se nos *missing values*. Posteriormente, o céu limpo, N/A, nevoeiro, neblina, névoa, garoa fraca e, por fim, nublado por esta ordem de maior para menos frequência.

3.4 Weather_Braga

Row ID	S cit...	I tem...	I atm...	I hum...	I wind...	I clouds	I precipit...	S current...	S sunrise	S sunset	S creation_date
Row0	Braga	8	1019	93	1	12	0	DARK	2019-01-15...	2019-01-15 ...	2019-01-15 19:05:00.000000
Row1	Braga	9	1020	93	1	20	0	DARK	2019-01-15...	2019-01-15 ...	2019-01-15 20:04:59.000000
Row2	Braga	10	1020	87	2	75	0	DARK	2019-01-15...	2019-01-15 ...	2019-01-15 21:04:58.000000
Row3	Braga	9	1020	93	3	75	0	DARK	2019-01-15...	2019-01-15 ...	2019-01-15 22:04:58.000000
Row4	Braga	10	1020	93	4	75	0	DARK	2019-01-15...	2019-01-15 ...	2019-01-15 23:04:58.000000
Row5	Braga	10	1020	87	4	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 00:04:59.000000
Row6	Braga	10	1020	87	3	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 01:04:58.000000
Row7	Braga	10	1019	87	3	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 02:04:58.000000
Row8	Braga	9	1019	93	4	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 03:04:58.000000
Row9	Braga	9	1019	87	3	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 04:04:58.000000
Row10	Braga	9	1018	87	3	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 06:04:59.000000
Row11	Braga	10	1019	93	2	75	0	LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 12:04:58.000000
Row12	Braga	11	1018	87	2	75	0	LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 13:05:00.000000
Row13	Braga	11	1018	87	2	75	0	LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 14:04:59.000000
Row14	Braga	11	1018	93	3	75	0	LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 15:05:00.000000
Row15	Braga	11	1018	93	2	75	0	LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 16:04:59.000000
Row16	Braga	11	1018	93	1	75	0	LOW_LIGHT	2019-01-16...	2019-01-16 ...	2019-01-16 17:04:59.000000
Row17	Braga	11	1018	100	1	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 18:04:59.000000
Row18	Braga	10	1019	100	2	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 19:04:59.000000
Row19	Braga	10	1019	100	2	75	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 20:04:58.000000
Row20	Braga	9	1019	100	0	40	0	DARK	2019-01-16...	2019-01-16 ...	2019-01-16 21:04:59.000000

Figura 12: Visualização do *dataset* Weather_Braga.

Este ficheiro contém 6821 ocorrências das quais tem as seguintes colunas:

- **city_name**: nome da cidade.
- **temperature**: temperatura naquele momento. Varia entre 1 a 34°C.
- **atmospheric_pressure**: pressão atmosférica. Varia de 990 a 1033 hPa.
- **humidity**: humidade presente no ar. Varia de 19 a 100%.
- **wind_speed**: velocidade do vento. Pode variar de 0 a 14 Km/h.
- **clouds**: nuvens no céu. Pode variar de 0 a 100%.
- **precipitation**: precipitação. Varia de 0 a 5.
- **current_luminosity**: luminosidade atual. Pode ser "DARK", "LIGHT" ou "LOW_LIGHT".
- **sunrise**: hora em que se deu o nascer do sol no formato MM/DD/AAAA HH:MM:SS.
- **sunset**: hora em se deu o pôr do sol no formato MM/DD/AAAA HH:MM:SS.
- **creation_date**: data de criação no formato MM/DD/AAAA HH:MM:SS.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance
temperature	<input type="checkbox"/>	1	34	15.119	5.506	30.312
atmospheric_pressure	<input type="checkbox"/>	990	1033	1017.864	5.956	35.469
humidity	<input type="checkbox"/>	19	100	81.831	17.190	295.491
wind_speed	<input type="checkbox"/>	0	14	3.051	2.180	4.752
clouds	<input type="checkbox"/>	0	100	35.266	34.921	1219.485
precipitation	<input type="checkbox"/>	0	5	0.002	0.106	0.011

Figura 13: Detalhe dos dados numéricos do *dataset* Weather_Braga.

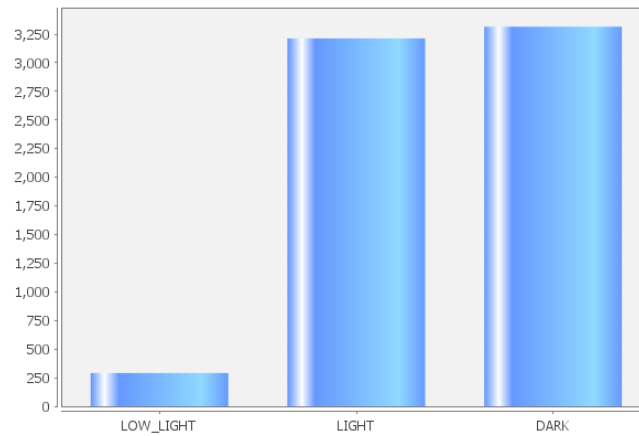


Figura 14: Detalhe da coluna **current_luminosity** do *dataset* Weather_Braga.

É possível observar pela figura 14 o mínimo, máximo, média, desvio padrão e variância para cada *feature* numérica deste ficheiro. Nenhuma destas colunas tem *missing values* incluindo os dados nominais.

A *feature* "current_luminosity" é o único dado nominal que não seja acerca de datas ou o nome da cidade neste ficheiro. Sendo assim, é possível observar que, das 6821 ocorrências existentes, a maioria corresponde a uma clara luminosidade e a uma luminosidade escura, enquanto que a minoria encontra-se em luminosidade pouco clara. Desta maneira, é possível concluir que, sabendo que o dia ocupa mais horas que a noite, existem menos registos do dia do que da noite, ou seja, este ficheiro contém mais informação acerca da meteorologia durante a noite.

4 Decisões de Conceção

Nesta fase do projeto foi realizada um estudo ponderado acerca das várias componentes dos *datasets*. Assim, tendo também em conta a análise anteriormente feita, foram removidos certos elementos e decidido o valor a calcular.

4.1 Traffic_Flow_Braga

Os parâmetros **city_name** e **Year** apresentam valores constantes e, portanto, desnecessários.

Os elementos **road_name** e **road_num** são ambos identificadores das 4 ruas existentes. Visto isto, apenas um deles é necessário, logo foi escolhido o elemento **road_num** para identificar as ruas uma vez que este é um valor numérico. Para além destes dois, existe também o elemento **functional_road_class_desc** que foi considerado não dar informação relevante visto ser praticamente igual à distinção já existente entre ruas.

Por fim, é necessário considerar as colunas acerca do tráfego - **current_speed**, **free_flow_speed**, **speed_diff**, **current_travel_time**, **free_flow_travel_time**, **time_diff** - e decidir quais vão permanecer no *dataset*.

Os elementos **free_flow_speed** e **free_flow_travel_time** são valores que são fixos (ou quase) relativos a cada rua. Sendo assim, só se encontram presentes com o objetivo de influenciar o **speed_diff** e o **time_diff**, uma vez que as colunas de velocidade e as colunas de tempo estão respetivamente relacionadas da seguinte maneira:

$$speed_diff = free_flow_speed - current_speed \quad (1)$$

$$time_diff = free_flow_time_travel - current_time_travel \quad (2)$$

Para além disto, também existe uma relação entre velocidade e tempo. Logicamente, quanto menor a velocidade, maior o tempo de viagem, significando, assim, que estes são inversamente proporcionais. Com efeito, é possível observar na figura 15 que os parâmetros **current_speed** e **current_travel_time** possuem uma grande correlação e confirma-se que são, de facto, inversamente proporcionais. Observa-se também que os parâmetros **time_diff** e **speed_diff** também estão bastante correlacionados.

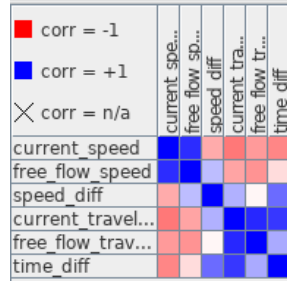


Figura 15: Correlação entre *speed* e *time*

Efetivamente, é possível verificar uma maior relação entre os valores de velocidade e os valores de tempo uma vez que foi possível verificar a seguinte expressão:

$$\frac{current_speed}{free_flow_speed} \approx \frac{free_flow_travel_time}{current_travel_time} \quad (3)$$

Tendo isto em conta, chegou-se à conclusão que o ideal seria calcular um novo valor que consiga representar a informação retirada destas 6 colunas:

$$speed_diff = \frac{current_speed}{free_flow_speed} \quad (4)$$

Este novo **speed_diff** representa então uma percentagem. Se este valor for, por exemplo, 0.3 significa que a velocidade atual está a 30% do que seria quando o fluxo está livre. Assim, sabendo os valores *free flow* da velocidade e do tempo, seria possível obter os valores atuais através das seguintes expressões:

$$current_speed = speed_diff * free_flow_speed \quad (5)$$

$$current_travel_time = speed_diff * free_flow_travel_time \quad (6)$$

4.2 Traffic Incidents Braga

Neste *dataset* vão ser necessários os elementos **from_road to_road** e **incidente_date** para fazer a conexão dos *datasets*.

Ponderando nos restantes dados, foi concluído que a informação relevante a retirar deste *dataset* é o nível de impacto que um incidente tem no tráfego. Assim, é possível observar que o elemento que fornece esta informação é o **magnitude_of_delay_desc**, sendo que este adota os valores **Minor**, **Moderate**, **Major** e **Undefined**.

4.3 Weather Braga Descriptions

Tendo em conta a anterior análise efetuada a este *dataset*, foi possível verificar que o parâmetro **snow** apenas possui valores **N/A** e *missing values*, o que o torna um parâmetro dispensável, visto não fornecer qualquer informação.

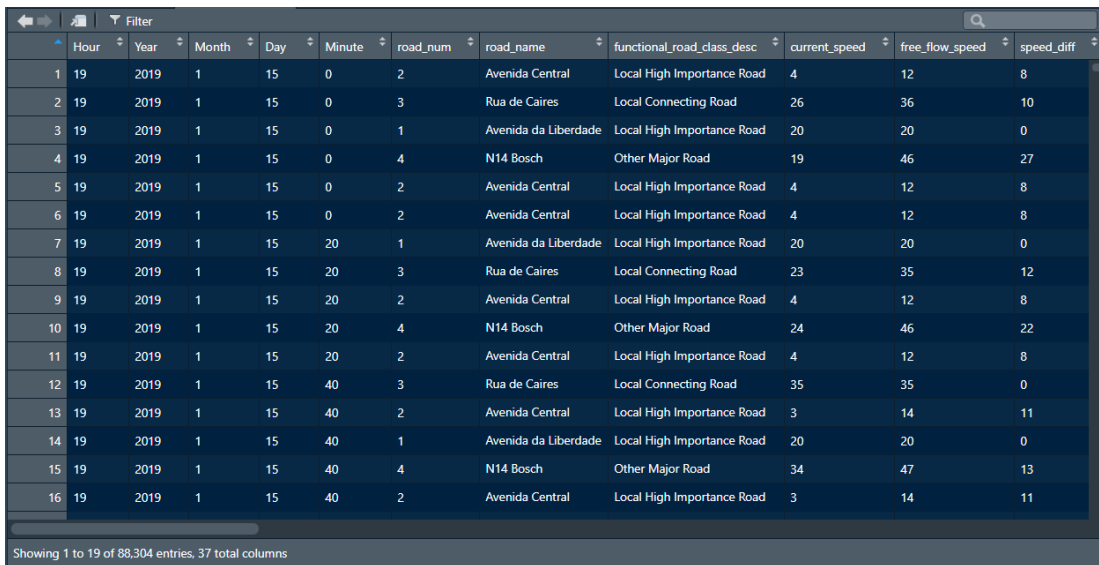
Este *dataset* também possui novamente o componente **city_name** que é constante e portanto irrelevante.

Por fim, ao observar o conteúdo do elemento **atmosphere**, reparou-se que este pode ser considerado como uma descrição das nuvens e que registos existentes ocorrem maioritariamente quando não existe registo no parâmetro **cloudiness**. Visto isto, foi retirado o elemento **atmosphere**, sendo os seus valores passados para o elemento **cloudiness**.

5 Junção dos Datasets

De forma a facilitar todo o processo de junção dos dados e de visualização gráfica desse mesmo, foram utilizadas a linguagem de programação **R** do IDE, **RStudio**, e da plataforma **Knime**. Estes, graças à sua enorme capacidade no tratamento de dados, interactividade, sintaxe de fácil utilização e interpretação, tomaram-se um ponto substancial na tomada desta decisão.

Na figura 16 é demonstrada a visualização de um *dataframe* que posteriormente será exportado como *csv*. Esta capacidade de visualização, oferecida pelo RStudio, permite uma grande interatividade, tal como indicado anteriormente, devido à possibilidade de utilizar os *sliders* para percorrer as colunas ou as linhas, ou então a utilização de filtros.



	Hour	Year	Month	Day	Minute	road_num	road_name	functional_road_class_desc	current_speed	free_flow_speed	speed_diff
1	19	2019	1	15	0	2	Avenida Central	Local High Importance Road	4	12	8
2	19	2019	1	15	0	3	Rua de Caires	Local Connecting Road	26	36	10
3	19	2019	1	15	0	1	Avenida da Liberdade	Local High Importance Road	20	20	0
4	19	2019	1	15	0	4	N14 Bosch	Other Major Road	19	46	27
5	19	2019	1	15	0	2	Avenida Central	Local High Importance Road	4	12	8
6	19	2019	1	15	0	2	Avenida Central	Local High Importance Road	4	12	8
7	19	2019	1	15	20	1	Avenida da Liberdade	Local High Importance Road	20	20	0
8	19	2019	1	15	20	3	Rua de Caires	Local Connecting Road	23	35	12
9	19	2019	1	15	20	2	Avenida Central	Local High Importance Road	4	12	8
10	19	2019	1	15	20	4	N14 Bosch	Other Major Road	24	46	22
11	19	2019	1	15	20	2	Avenida Central	Local High Importance Road	4	12	8
12	19	2019	1	15	40	3	Rua de Caires	Local Connecting Road	35	35	0
13	19	2019	1	15	40	2	Avenida Central	Local High Importance Road	3	14	11
14	19	2019	1	15	40	1	Avenida da Liberdade	Local High Importance Road	20	20	0
15	19	2019	1	15	40	4	N14 Bosch	Other Major Road	34	47	13
16	19	2019	1	15	40	2	Avenida Central	Local High Importance Road	3	14	11

Showing 1 to 19 of 88,304 entries. 37 total columns

Figura 16: Exemplo das funcionalidades oferecidas pelo RStudio

Olhando mais concretamente para as informações que são disponibilizadas, tem-se em conta que os *datasets* são relativos à mesma cidade, assim será de esperar que a meteorologia seja igual durante certos períodos de tempo para toda a cidade. Assim, irá proceder-se à junção dos *datasets* por via dos registos temporais. À primeira vista, verificou-se que existe essa possibilidade, porém os minutos podem diferir, tal como os segundos, impossibilitando uma junção "direta".

Com esta adversidade, separaram-se os registos temporais em novas colunas destinadas a cada parte do registo, ou seja, uma coluna para "Horas", "Minutos", "Ano", "Dia" e "Mês", desprezando assim os registos para os segundos. Assim, tornou-se possível a junção dos *datasets* através desses 5 elementos.

Posteriormente ao *merge* realizado, dividiu-se este novo *dataframe* em dois, um com registos dos acidentes (**accidents**) e outro com registos sem acidentes (**no_accidents**). Estes dois novos *dataframes* foram "filtrados" a partir das colunas associadas ao percurso/localização. Ou seja, a partir da coluna **road_name**, verificou-se que se trata de uma *substring* da coluna **from_road** ou da **to_road**. Caso esta condição se verifique, é concluído que está presente uma situação de acidente e, assim, este registo, fará parte do *dataframe accidents*. Caso contrário, o registo fará parte do *dataframe no_accidents*. Posteriormente à separação dos dados, foram substituídas todas as colunas referentes ao *dataset Traf-*

fic_Incidents_Braga, por **N/A** no *dataframe* **no_accidents**. Assim sendo, e por via de outros tratamentos de dados, é possível dar o conhecimento ao modelo das diferenças entre o tráfego normal ou por via de acidente.

Por fim, juntaram-se ambos os *dataframes*, e aplicou-se a função *unique*, onde esta se encarrega de remover os duplicados, deixando assim o *dataset* com cerca de 88304 linhas. Para além da *unique*, aplicou-se a função *complete.cases* sobre as *features* "Hour", "Minute", "Day", "Month" e "Year", em que esta se encarrega por indicar as linhas em que não existe valores **N/A**, nestas mesmas colunas.

6 Tratamento de Dados

Uma vez que nesta fase está efetuada a junção dos *datasets*, avança-se para o tratamento dos dados do mesmo através da plataforma **Knime**, sendo possível observar o trabalho realizado na mesma nas figuras 17, 18, 19 e 20.

Nesta secção irá analisar-se a correlação dos vários parâmetros de maneira a ser realizada uma melhor escolha do que deverá permanecer no *dataset*.

Para além disto, é implementado *label encoding* e são tratados os *missing values* de modo a que os valores do *dataset* fiquem prontos para a fase seguinte.

Outro aspeto relevante considerado é a coerência dos dados, uma vez que a informação dos vários elementos não pode ser contraditória.

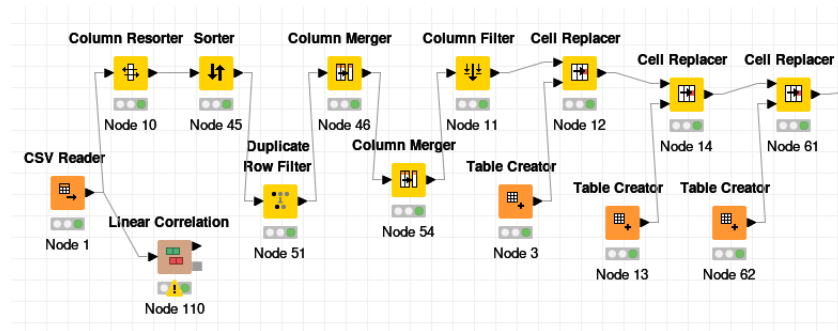


Figura 17: Knime *workflow*: tratamento de dados - parte 1.

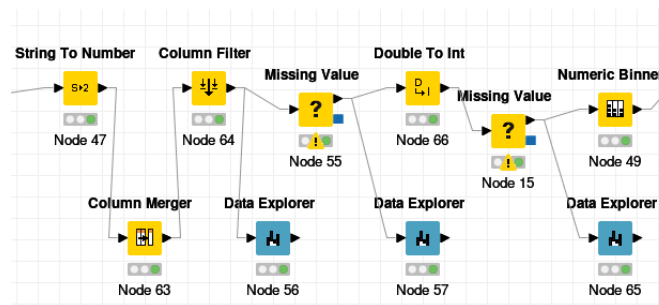
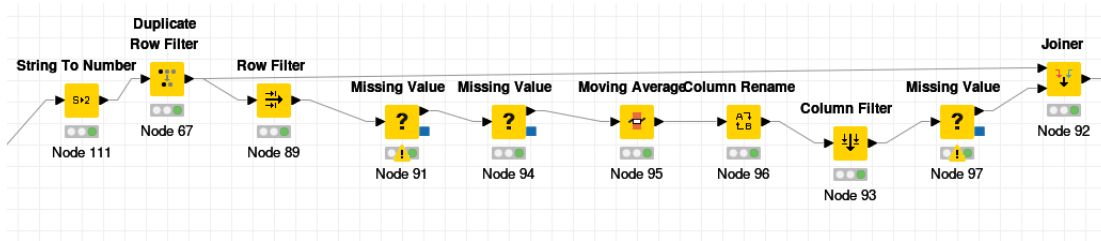
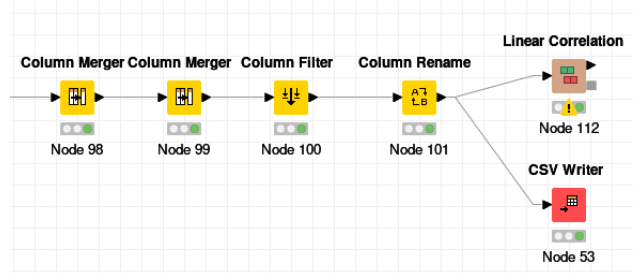


Figura 18: Knime *workflow*: tratamento de dados - parte 2.

Figura 19: Knime *workflow*: tratamento de dados - parte 3.Figura 20: Knime *workflow*: tratamento de dados - parte 4.

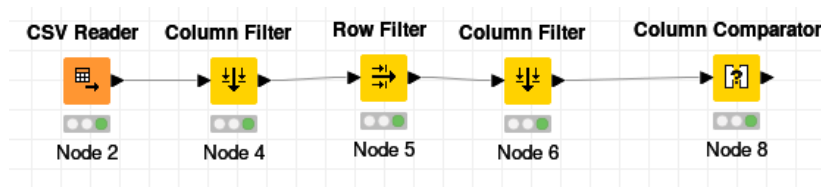
6.1 Correlação dos Dados

Primeiramente, ao analisar os vários parâmetros foram encontrados alguns que se referem à mesma informação. Com efeito, é de notar que os elementos **rain** e **precipitation** se referem à mesma informação. Da mesma maneira, salientam-se os elementos **cloudiness** e **clouds**. Posto isto, a informação destes parâmetros foi unida num só, ou seja, através do nodo **Column Merger** do Knime a formação presente no parâmetro **rain** será transportada para o parâmetro **precipitation** e a informação presente no parâmetro **clouds** será transportada para o parâmetro **cloudiness**.

De seguida, foi dada atenção aos parâmetros **sunrise** e **sunset**. Estas informações não são relevantes para o tráfego neste estado. No entanto, são importantes para se saber se é dia ou noite, amanhecer ou anoitecer, ou seja, essencialmente para saber a luminosidade.

Assim, as datas de ambos os parâmetros foram separadas e foram criadas mais duas colunas, **SunsetHourDiff** e **SunriseHourDiff**. Estes dois novos elementos comparam a hora, que foi previamente separada dos respetivos parâmetros, com a hora atual e calculam a sua diferença.

Com estes valores é possível saber quando é noite (a luminosidade é **DARK**) e quando é dia (a luminosidade é **LIGHT**). No entanto, existe a questão de quando ocorre o **LOW_LIGHT**. Para determinar isto, analisou-se o *dataset* através do *workflow* representado na figura 21.

Figura 21: Knime *workflow*: análise de luminosidade-sunrise-sunset

Concluiu-se então que o **LOW_LIGHT** ocorre até duas horas depois da hora do **sunrise** e desde duas horas antes do **sunset**. Deste modo, foram preenchidos os *missing values* do parâmetro **current_luminosity** e descartadas as informações relativas ao amanhecer e anoitecer.

Por fim, observou-se a correlação entre todas as *features* existentes. A partir da figura 22, os

quadrados que quanto mais azul estão, maior relação de proporcionalidade têm. De maneira oposta, quanto mais vermelhos estão, maior relação de proporcionalidade inversa têm. Sendo assim, é possível concluir que as *features* que afetam o **speed_diff** são:

- Month
- road_num
- free_flow_travel_time
- humidity
- HourSunrise
- SunriseHourDiff
- weekday
- Hour
- current_speed
- temperature
- wind_speed
- HourSunset
- SunsetHourDiff

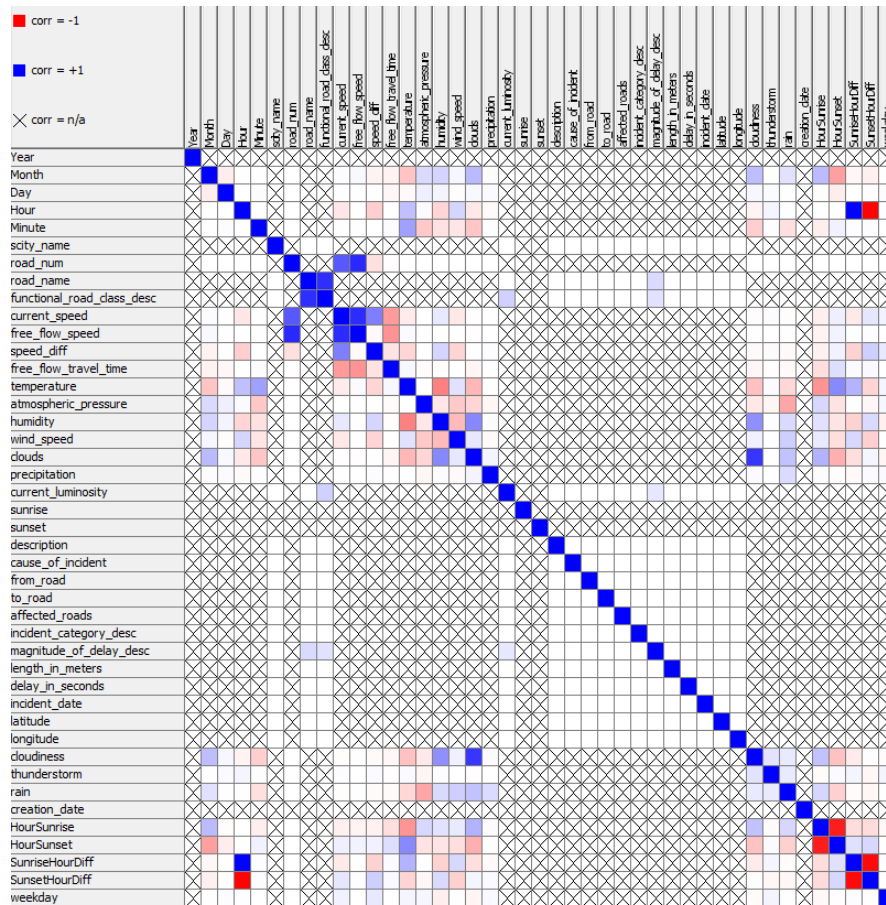


Figura 22: Correlação inicial entre as *features*.

No entanto, tendo em conta os pontos apresentados nesta secção e na secção 4, a lista de *features* que foram decididas manter não se trata da acima apresentada. Desta maneira, foram retirados os elementos **current_speed**, **free_flow_travel_time**, **HourSunrise**, **HourSunset**, **SunriseHourDiff** e **SunsetHourDiff** e foi adicionado o elemento **current_luminosity**.

Para além disso, são necessários os elementos **Day** e **Minute** uma vez que fazem parte da data. Também considerou-se que os elementos **cloudiness**, **precipitation** e **thunderstorm**, tendo em conta o senso comum, seriam componentes relevantes para o tráfego.

Quanto ao parâmetro **humidity**, embora tenha influência no **speed_diff**, apresenta *missing values*

difíceis de tratar devido à escassez de registos e, tendo em conta a sua correlação com os parâmetros **temperatura**, **cloudiness** e **wind_speed**, é possível considerar que a informação que nos fornece já está em parte a ser adquirida através desses outros parâmetros.

Assim, o *dataset* fica com as seguintes *features*:

- Month
- Day
- Hour
- Minute
- weekday
- road_num
- current_luminosity
- temperature
- precipitation
- cloudiness
- wind_speed
- thunderstorm
- speed_diff

Posteriormente ao resto de tratamento de dados que ainda será abordado nas seguintes secções, foi analisado novamente a correlação dos dados, obtendo-se a figura 23.

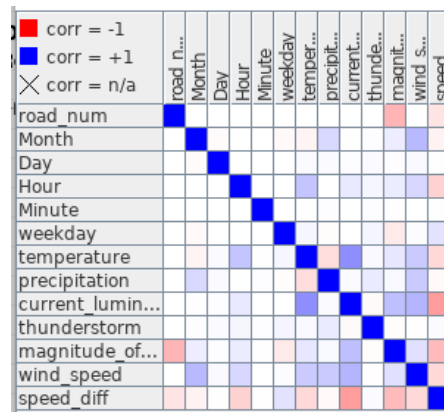


Figura 23: Correlação final entre as *features*.

6.2 Label Encoding

Para melhor desempenho, é necessário passar os parâmetros nominais para numéricos. Foi então decidido recorrer ao *label encoding* para fazer esta passagem. Para realizar este processo foram utilizados no Knime os nodos **Table Creator** e **Cell Replacer**.

Ao parâmetro **rain**, antes deste ser unido com o **precipitation**, foram atribuídos valores inteiros no intervalo [1,4], como é possível observar na figura 24.

Row ID	S column1	S column2
Row0	chuva leve	1
Row1	chuva fraca	2
Row2	chuva mo...	3
Row3	chuva	3
Row4	aguaceiros	2
Row5	aguaceiro...	1
Row6	chuva forte	4
Row7	chuvisco ...	1
Row8	chuvisco ...	2

Figura 24: *Label Encoding*: **rain**.

Ao parâmetro **cloudiness** foram atribuídos valores inteiros no intervalo $[0,100]$, como é possível observar na figura 25, uma vez que este ia posteriormente ser unido com o **clouds**, que continha valores semelhantes.

Row ID	S column1	S column2
Row0	céu limpo	0
Row1	céu claro	0
Row2	nuvens di...	20
Row3	algumas ...	40
Row4	céu pouc...	50
Row5	nuvens q...	60
Row6	tempo nu...	70
Row7	nublado	70

Figura 25: *Label Encoding*: **cloudiness**.

Ao parâmetro **atmosphere**, antes deste ser unido com o **cloudiness**, foram atribuídos valores inteiros no intervalo $[0,100]$, como é possível observar na figura 26.

Row ID	S column1	S column2
Row0	céu limpo	0
Row1	neblina	30
Row2	névoa	40
Row3	garoa fraca	20
Row4	nublado	70
Row5	nevoeiro	80

Figura 26: *Label Encoding*: **atmosphere**.

Ao parâmetro **magnitude_of_delay_desc** foram atribuídos valores inteiros no intervalo $[1,4]$, como é possível observar na figura 27.

Row ID	S column1	S column2
Row0	Minor	1
Row1	Moderate	2
Row2	Major	3
Row3	Undefined	4

Figura 27: *Label Encoding*: **magnitude_of_delay_desc**.

Ao parâmetro **thunderstorm** foi atribuído o valor inteiro 1, como é possível observar na figura 27, uma vez que foi decidido que esta coluna serviria apenas para informar se ocorreu uma tempestade ou não.

Row ID	S column1	S column2
Row0	trovoada	1
Row1	trovoada com chuva fraca	1
Row2	trovoada com chuva forte	1

Figura 28: *Label Encoding*: **thunderstorm**.

Ao parâmetro **current_luminosity** foram atribuídos valores inteiros no intervalo $[0,2]$, como é possível observar na figura 29,

Row ID	S column1	S column2
Row0	DARK	0
Row1	LOW_LIGHT	1
Row2	LIGHT	2

Figura 29: *Label Encoding*: **current_luminosity**.

6.3 Coerência dos Dados

Um aspeto importante no *dataset* é que a informação esteja corretamente dividida e coerente entre si. Por esta razão, foram feitos vários tratamentos dos dados de maneira a garantir estes dois aspetos.

Quanto ao primeiro tópico, foi observado que no parâmetro **thunderstorm** existem registos que referem chuva. Por conseguinte, efetuou-se a separação da informação relativa à chuva, tal como é possível observar na figura 30 sendo esta posteriormente adicionada ao parâmetro **precipitação**.

Row ID	S column1	S column2
Row0	trovoada	0
Row1	trovoada com chuva fraca	2
Row2	trovoada com chuva forte	4

Figura 30: Coerência dos dados: trovoada - chuva.

Quanto ao segundo, uma vez que vários parâmetros estão interligados, é necessário confirmar que a informação que estes fornecem não é contraditória. Assim, foram feitas várias correspondências entre parâmetros através dos nodos **Table Creator** e **Cell Replacer** do Knime.

Primeiramente, caso o parâmetro **thunderstorm** indique que houve trovoada, o parâmetro **cloudiness** tem de indicar que existiam nuvens uma vez que o contrário seria impossível. Deste modo, foi criada uma correspondência entre a trovoada e as nuvens, tal como é possível observar na figura 31.

Row ID	S column1	S column2
Row0	trovoada	100
Row1	trovoada ...	100
Row2	trovoada ...	100

Figura 31: Coerência dos dados: trovoada - nuvens.

Da mesma maneira, caso o parâmetro **precipitation** indique que choveu, o parâmetro **cloudiness** tem de indicar que existiam nuvens uma vez que o contrário seria impossível. Deste modo, foi criada uma correspondência entre a chuva e as nuvens, tal como é possível observar na figura 32.

Row ID	S column1	S column2
Row0	1	20
Row1	2	30
Row2	3	50
Row3	4	60

Figura 32: Coerência dos dados: chuva - nuvens.

Por outro lado, também o inverso tem de ser efetuado, ou seja, se o parâmetro **cloudiness** indicar céu limpo, o parâmetro **precipitation** tem de indicar que não choveu uma vez que o contrário seria impossível. Deste modo, foi criada outra correspondência entre a chuva e as nuvens, tal como é possível observar na figura 33.

Row ID	S column1	S column2
Row0	0	0
Row1	20	?
Row2	40	?
Row3	50	?
Row4	60	?
Row5	70	?
Row6	100	?

Figura 33: Coerência dos dados: céu limpo - sem chuva.

Por fim, no parâmetro **magnitude_of_delay_desc**, tinha-se reparado, numa prévia análise ao *dataset* **Traffic Incidents Braga**, que os registos correspondentes aos valores 4 (previamente **Undefined**) são caracterizados como *road closed*. Tendo em conta esta informação, foram analisados os registos do **speed_diff** correspondentes a estes valores, sendo constatado que o valor era sempre 1.

Em suma, quando a rua se encontra fechada, o *dataset* declara a velocidade como sendo a velocidade de quando o fluxo de tráfego está livre. Após ponderação, concluiu-se que faria mais sentido ser considerado que a velocidade era nula uma vez que não se pode passar numa rua fechada. Assim, foi criada uma correspondência entre ambos os parâmetros, como é possível observar na figura 34.

Row ID	S column1	S column2
Row0	4	0
Row1	3	?
Row2	2	?
Row3	1	?
Row4	0	?

Figura 34: Coerência dos dados: rua fechada - velocidade zero.

6.4 Missing Values

Na secção de análise dos dados foi constatado que o *dataset* continha vários *missing values*, logo estes foram resolvidos com o nodo **Missing Values** do Knime. Vários parâmetros foram sujeitos ao mesmo tratamento destes *missing values*, uma vez que o procedimento trata-se de em tornar os dados coerentes, ou seja, dar ligeira continuidade aos acontecimentos registados.

Assim, para os parâmetros **magnitude_of_delay**, **thunderstorm**, **precipitation**, **wind_speed** e **cloudiness**, esta continuidade foi adquirida através de **Moving Average**. De seguida, estes parâmetros foram completados com um **Fix Value** com o valor de 0, visto que, para além de dar continuidade suficiente para que as alterações se assemelhem à realidade, só se pode assumir que no resto dos *missing values* não ocorreram os eventos.

Quanto ao parâmetro **current_luminosity**, como estes valores já tinham sido preenchidos numa versão anterior, sabe-se que as alterações entre valores já estão registadas. Por esta razão, estes *missing values* foram resolvidos com **Next Value**.

Em relação ao parâmetro **temperature**, a solução não pode ser como as anteriores, uma vez que os registos em falta não podem ser simplesmente colocados a zero e esta se comporta de maneira diferente dos parâmetros anteriores. Assim, foi reduzido o *dataset* a apenas uma rua e os *missing values* foram tratados com **Linear Interpolation**, uma vez que os valores aumentam e decrescem continuamente de forma linear.

6.5 Binning

Uma parte importante do *dataset* é que este esteja dividido em períodos de tempo iguais. Sendo assim, após análise, este foi dividido em intervalos de 20 minutos. No parâmetro **Minute** todos os

valores no intervalo $[0,19]$ foram colocados a 0, todos os valores no intervalo $[20,39]$ foram colocados a 20 e todos os valores no intervalo $[40,59]$ foram colocados a 40, tal como demonstrado na figura 35.

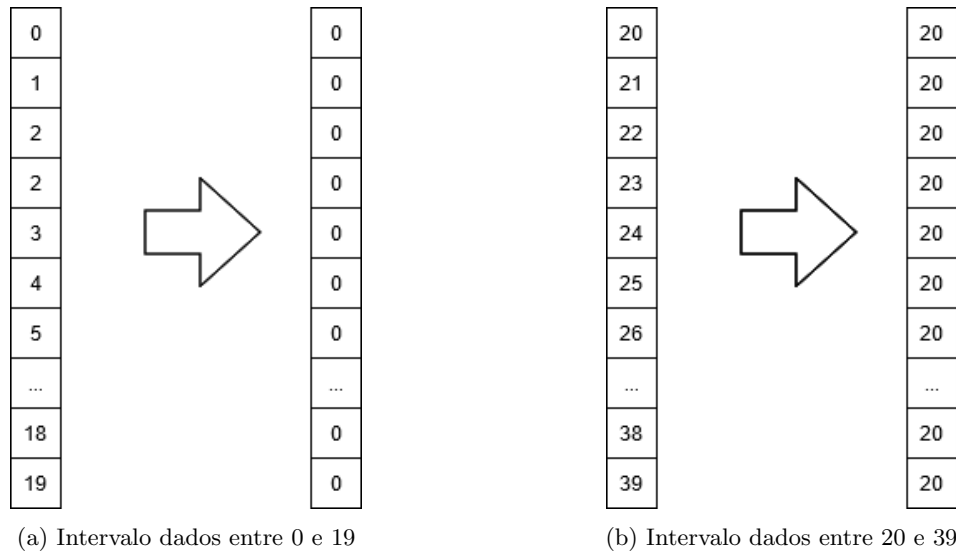


Figura 35: "Round Down" aplicado aos minutos

6.6 Preenchimento de *timesteps*

Com o intuito de ajudar o modelo na compreensão sobre as situações de acidente ou não, é necessário criar os registos em falta, por exemplo os registos do mês de janeiro só começam a partir do dia 15. Assim sendo, e tendo em conta que existem meses que diferem na quantidade de dias, foram criados três *dataframes* sobre o qual cada um incidia nessa quantidade. Este processo de criação tem de gerar todas as combinações possíveis de registos diários e das ruas existentes. Com isto em consideração, existirão quatro registos para cada instante diferindo na rua, como está ilustrado na figura 39.

Day	Month	Year	Hour	Minute	Road_Num
1	1	2019	19	0	1
1	1	2019	19	0	2
1	1	2019	19	0	3
1	1	2019	19	0	4
1	1	2019	19	20	1
...

Figura 36: Exemplo criação dos *timesteps*.

Após a criação e junção destes *dataframes*, procedeu-se à remoção dos registos sobre o qual já se possuíam dados, para isso utilizou-se um *subset* dos dados que foram fornecidos, e aplicou-se a diferença entre os dois *frames*. Assim conseguiram-se obter os registos inexistentes nos dados fornecidos, onde estes, posteriormente, serão concatenados com o resultado desta diferença, em que as colunas em falta serão preenchidas com **N/As**.

Tendo em conta que se está a utilizar uma RNN e, portanto, prever futuros acontecimentos, a ordem dos dados é imperativa. Como tal, ao *dataframe* anteriormente descrito, aplicou-se uma ordenação, para assim satisfazer este requisito.

Nesta fase existia uma coisa a tratar: as linhas que foram acrescentadas sem valores. Nesta fase, utilizou-se o Knime com o *workflow* demonstrado na figura 37.

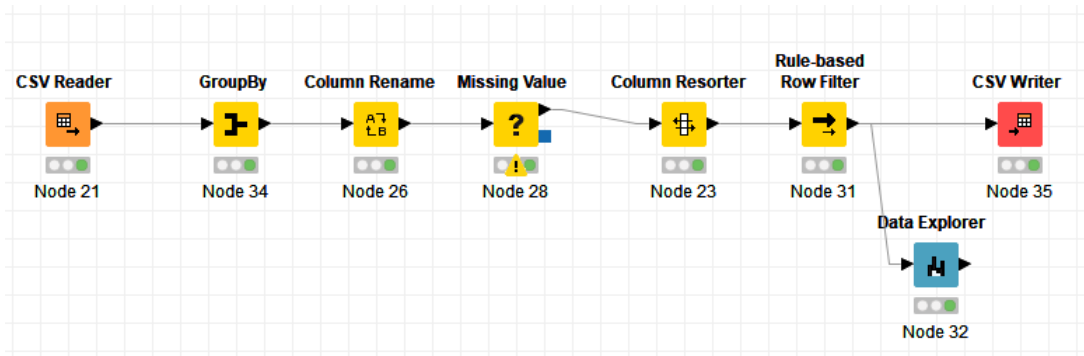


Figura 37: Workflow do Knime.

No **speed_diff** foi necessário fazer uma análise mais profunda. Como o *dataset* começa apenas a 15 de janeiro e como o mês de março se encontrava muito incompleto, removeram-se estes dois meses por completo do *dataset*. O mês de fevereiro, abril, julho, agosto, novembro e dezembro tinham alguns dias em falta e decidiu-se removê-los também. Esta tarefa foi feita com o nodo **Rule-based Row Filter**.

Por fim, foram colocados os *missing values* do **speed_diff** com o valor fixo de -99, o que terá utilidade no próximo passo do projeto.

7 Modelo Desenvolvido

7.1 Características do treino

Inicialmente, todas as *features* foram normalizadas com valores entre 0 e 1. Posto isto, o *dataset* é transformado num problema de *supervised learning*.

Posteriormente, o *dataset* é separado em dados de treino e de teste e depois separados em variáveis de *input* e *output*. Finalmente, o *input* X é tornado no formato 3D esperado pelas LSTMs: (*samples*, *timesteps*, *features*).

De seguida, é possível fazer *fit* e definir o modelo LSTM. Antes de definir a primeira camada LSTM, procedeu-se à inserção de uma camada de *masking* que tem o objetivo de ignorar na aprendizagem os campos que sejam iguais a -99.

A primeira camada LSTM contém 128 neurónios. O *input_shape* é de 1 *timestep* com 13 *features*. São usadas também camadas de Dropout, Dense e mais duas LSTM. No fim, o modelo é compilado usando a função de perda MAE com o otimizador Nadam, que foi o otimizador que deu *loss* mais baixa, comparadamente aos outros possíveis.

```

1 def build_model(train_X=train_X, dropout_rate=0.6):
2     model = Sequential()
3     model.add(Masking(mask_value=-99., input_shape=(train_X.shape[1], train_X.shape[2])))
4     model.add(LSTM(128, input_shape=(train_X.shape[1], train_X.shape[2]),
5         return_sequences=True))
6     model.add(Dropout(dropout_rate))
7     model.add(LSTM(256, return_sequences=True, dropout=dropout_rate))
8     model.add(LSTM(512, dropout=dropout_rate))
9     model.add(Dense(50, activation='tanh'))
10    model.add(Dropout(dropout_rate))
11    model.add(Dense(1, activation='linear'))
12
13    opt = keras.optimizers.Nadam(learning_rate=0.001)
14    model.compile(loss='mae', optimizer=opt)

```

```
14 return model
```

Listing 1: Camadas do modelo.

7.2 Tuning e Cross Validation

Os hiperparâmetros são variáveis de configuração que são externas ao modelo, ou seja, são especificados antes do treino. Estes definem os conceitos de maior nível em relação ao modelo. Com os valores certos, os hiperparâmetros eliminam as hipóteses de haver *overfitting* e *underfitting*. O objetivo do *tuning* é descobrir os melhores valores para estes hiperparâmetros.

A *cross validation* é uma técnica usada que avalia a capacidade de generalizar um modelo, a partir dos dados. Tem o objetivo de estimar quão preciso o modelo em causa é. Consiste em particionar dados em conjuntos mutuamente exclusivos e o uso destes para estimar os parâmetros do modelo para o treino.

Na biblioteca **scikit-learn** existe o **Grid-search** que devolve os hiperparâmetros que obtenham os melhores resultados a partir de *cross validation*.

```
1 # define the grid search parameters
2 grid_param_LSTM = {
3     'epochs': [50,60,70,80],
4     'learning_rate':[0.001,0.0001]
5 }
6
7 model_LSTM=KerasRegressor(build_fn=build_model, batch_size=72)
8
9 grid = GridSearchCV(estimator=model_LSTM,
10                    param_grid=grid_param_LSTM,
11                    scoring={'neg_mean_absolute_error'},
12                    refit='neg_mean_absolute_error',
13                    cv=2)
14
15 grid_result = grid.fit(X=test_X, y=test_y, verbose=0)
16
17 # summarize results
18 print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))
19 means = grid_result.cv_results_['mean_test_neg_mean_absolute_error']
20 stds = grid_result.cv_results_['std_test_neg_mean_absolute_error']
21 params = grid_result.cv_results_['params']
22 for mean, stdev, param in zip(means, stds, params):
23     print("%f (%f) with: %r" % (mean, stdev, param))
```

Listing 2: Uso do GridSearch.

É testado o mesmo modelo com 50, 60, 70 e 80 épocas e com taxas de aprendizagem de 0.001 e 0.0001 usando a *loss* MAE (*mean absolute error*). Desta forma, os gráficos são desenhados na mesma proporção que o *dataset* sem ser necessário fazer desnormalização.

8 Discussão dos Resultados

Posteriormente ao uso do Grid-search, alguns parâmetros foram alterados:

- épocas: 70
- *learning_rate*: 0.001

- *batch_size*: 75
- *dropout_rate*: 0.6

Best: -0.008161 using {'epochs': 70, 'learning_rate': 0.001}

Figura 38: Resultado do Grid-search para as *epochs* e *learning_rate*.

Best: -0.015880 using {'batch_size': 75, 'dropout_rate': 0.6}

Figura 39: Resultado do Grid-search para o *batch_size* e *dropout_rate*.

Depois de retificar os hiperparâmetros, obteve-se o gráfico da figura 40 com umas *loss* de, aproximadamente, 0.66%.

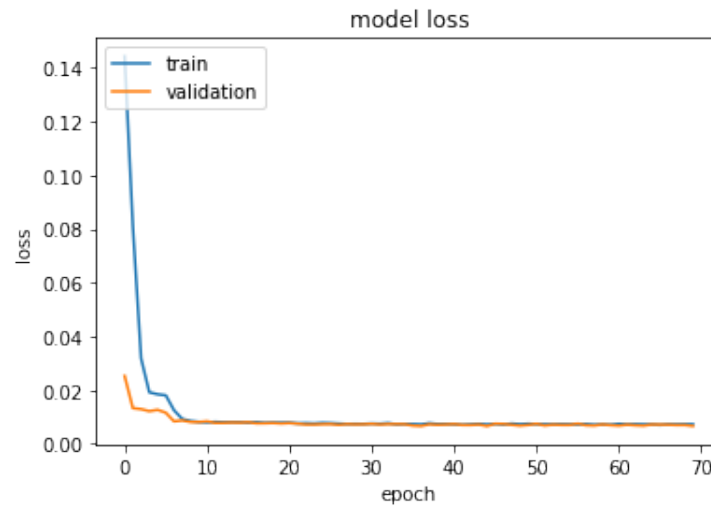


Figura 40: Gráfico de *loss*.

Este valor de *loss* é bastante bom talvez pelo facto de o **speed_diff** ser uma percentagem e, consequentemente, ver mais facilmente o que poderia afetar o tráfego, uma vez que se excluíram as *features* que poderiam complicar a previsão. Ou seja, sendo a velocidade máxima de uma rua diferente das outras, se existe a mesma quantidade de tráfego em cada uma, a percentagem será a mesma para todas, o que contribuiu para um modelo mais preciso.

9 Conclusão

Neste trabalho, o grupo deparou-se com algumas dificuldades no tratamento de dados para decidir o que fazer com algumas *features* e foi aqui que se ocupou grande parte da realização do projeto. Nesta fase ocorreu uma falha, uma vez que estava previsto manter o parâmetro **cloudiness**, mas este foi removido. Após esta falha ser descoberta, o *dataset* foi corrigido e testado. Contudo, visto o resultado obtido ter piorado, foi decidido manter o *dataset* original.

Já a realização do modelo foi mais fluída pelo que a experiência anterior em redes RNN ajudou a fácil realização das camadas necessárias conseguindo assim uma *loss* de 0.66%.

Este projeto poderia ter obtido consideravelmente melhores resultados ou até outras configurações implementadas se a gestão do tempo tivesse sido melhor. Apesar disso, os elementos do grupo estão relativamente satisfeitos com os resultados finais.