Relatório TP1 SBS-MLFA

Luis Freitas - PG38347; Daniel Pereira

November 22, 2020

Abstract

1 Introdução

2 Metodologias

2.1 CRISP-DM

2.2 Modelo Baseados em Árvores

3 Arquiteturas e Ferramentas

4 DataSet da Competição

4.1 Contexto do Dataset

O dataset utilizado neste projeto é referente ao nível de incidentes rodoviários na zona de Braga e contém algumas features que representam a magnitude do atraso que se verifica a cada hora, a temperatura, a pressão atmosférica, a velocidade do vento e ainda outras features.

O principal objetivo é aprimorar um modelo de Machine Learning Baseado em Arvores mas também fazer uma analise exploratoria completa dos dados para se retirar algumas informações importantes do dataset. Outo ponto passa por conseguir preprarar o dataset com técnicas de engenharia de dados para se obter os melhores resultados possiveis. Este dataset é referente a uma competição da plataforma Kaggle

Em relação as features do dataset estas são apresentadas da seguinte forma:

1. **city_name** - nome da cidade em causa;
2. **record_date** - o timestamp associado ao registo
3. **magnitude_of_delay** - magnitude do atraso provocado pelos incidentes que se verificam no record_date correspondente;
4. **delay_in_seconds** - atraso, em segundos, provocado pelos incidentes que se verificam no record_date correspondente;
5. **affected_roads** - estradas afectadas pelos incidentes que se verificam no record_date correspondente;
6. **luminosity** - o nível de luminosidade que se verificava na cidade de Braga;
7. **avg_temperature** - valor médio da temperatura para o record_date na cidade de Braga;
8. **avg_atm_pressure** - valor médio da pressão atmosférica para o record_date na cidade de Braga;
9. **avg_humidity** - valor médio da humidade para o record_date na cidade de Braga;
10. **avg_wind_speed** - valor médio da velocidade do vento para o record_date na cidade de Braga;
11. **avg_precipitation** - valor médio de precipitação para o record_date na cidade de Braga;
12. **avg_rain** - avaliação qualitativa do nível de precipitação para o record_date na cidade de Braga;
13. **accidents** - indicação acerca do nível de incidentes rodoviários que se verificam no record_date correspondente na cidade de Braga;

4.2 Análise e Transformação dos Dados

O DataSet utilizado apresenta cinco mil linhas e foi analisado de forma a serem retiradas informações que pudessem ser uteis no desenvolvimento do modelo.

(DANIEL ESCREVE AQUI A ANÁLISE DE DADOS QUE JUSTIFIQUE TODAS AS TRANSFORMAÇÕES FEITAS)

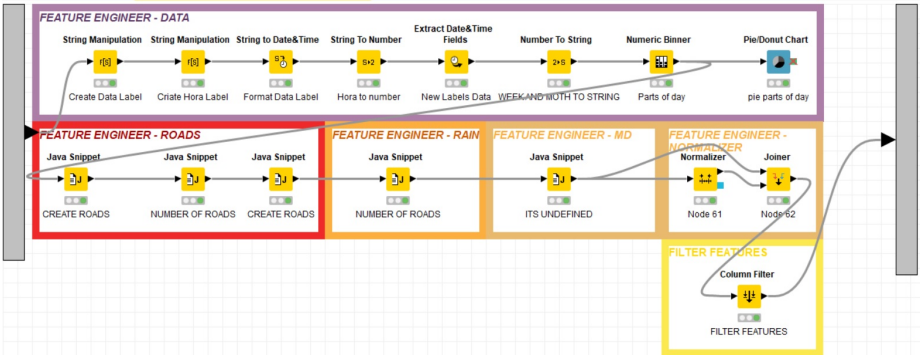

Figure 1: Workflow Data Preparation

As transformações de dados realizadas foram uteis para os resultados do modelo e foram suportadas pela analise de dados efetuada e explicada no capitulo anterior.

Como se pode ver na figura em baixo, pode-se retirar várias informações de algumas features, como por exemplo da record_date e da affected_roads. Nesta forma inicial estas duas features não contém muita informação pertinente para o modelo.

Row ID	city_name	magnit...	delay_i...	affecte...	record_date	luminosity	avg_te...	avg_at...	avg_hu...	avg_wi...	avg_pr...	avg_rain	accidents
Row0	Braga	UNDEFINED	0	N103,N103,	2019-02-17 20:00	DARK	13	1,017	87	2	0	chuva fraca	Low
Row1	Braga	MAJOR	401	N101,N103,...	2019-02-14 12:00	LIGHT	17	1,025	45	5	0	Sem Chuva	Medium
Row2	Braga	UNDEFINED	1057	CM1348,CM...	2019-09-03 19:00	DARK	27	1,018	53	4	0	Sem Chuva	Low
Row3	Braga	MAJOR	3612	N101,CM13...	2019-06-18 18:00	LIGHT	16	1,013	100	6	0	Sem Chuva	High
Row4	Braga	MAJOR	498	N14,N101,C...	2019-05-08 13:00	LIGHT	15	1,014	82	6	0	Sem Chuva	Medium
Row5	Braga	MAJOR	1270	CM1348,CM...	2019-05-25 21:00	DARK	19	1,020	82	5	0	Sem Chuva	High
Row6	Braga	UNDEFINED	0	CM1348,CM...	2019-09-16 04:00	DARK	17	1,017	93	2	0	Sem Chuva	None
Row7	Braga	MAJOR	1264	CM1348,CM...	2019-08-15 11:00	LIGHT	26	1,021	73	4	0	Sem Chuva	Medium
Row8	Braga	UNDEFINED	0	N103,N103,	2019-02-21 06:00	DARK	10	1,022	76	5	0	Sem Chuva	Low
Row9	Braga	MAJOR	23804	N14,N103,N...	2019-05-06 17:00	LIGHT	16	1,020	72	2	0	Sem Chuva	Very_High
Row10	Braga	MAJOR	867	CM1348,N1...	2019-09-17 14:00	LIGHT	22	1,016	100	1	0	Sem Chuva	Medium
Row11	Braga	MAJOR	315	CM1348,A1...	2019-10-10 20:00	DARK	16	1,017	82	2	0	Sem Chuva	Low
Row12	Braga	UNDEFINED	0	CM1348,CM...	2019-10-05 13:00	LIGHT	20	1,019	72	3	0	Sem Chuva	None
Row13	Braga	UNDEFINED	0	,	2019-01-19 23:00	DARK	12	1,010	87	6	0	Sem Chuva	None
Row14	Braga	MAJOR	2714	N14,N103,C...	2019-10-17 13:00	LIGHT	17	1,016	82	3	0	Sem Chuva	High
Row15	Braga	MAJOR	1004	CM1348,N1...	2019-09-12 13:00	LIGHT	29	1,025	42	3	0	Sem Chuva	Medium
Row16	Braga	MAJOR	9298	EM569,N10...	2019-11-15 08:00	LIGHT	5	1,012	80	2	0	Sem Chuva	Very_High
Row17	Braga	UNDEFINED	0	CM1348,CM...	2019-07-24 06:00	LIGHT	15	1,015	100	1	0	Sem Chuva	None
Row18	Braga	MAJOR	2059	CM1348,N1...	2019-10-03 15:00	LIGHT	21	1,021	68	2	0	Sem Chuva	High
Row19	Braga	MAJOR	658	N101,N201	2019-04-17 10:00	LIGHT	12	1,007	87	6	0	chuva fraca	None
Row20	Braga	UNDEFINED	0	,	2019-12-23 06:00	DARK	8	1,026	93	3	0	Sem Chuva	None
Row21	Braga	UNDEFINED	0	CM1348,CM...	2019-05-19 13:00	LIGHT	18	1,016	55	6	0	Sem Chuva	Low
Row22	Braga	MAJOR	262	,	2019-05-11 19:00	DARK	23	1,021	68	4	0	Sem Chuva	None
Row23	Braga	UNDEFINED	363	N201,CM13...	2019-07-01 06:00	LIGHT	16	1,020	100	0	0	Sem Chuva	Medium
Row24	Braga	UNDEFINED	0	,	2019-01-21 23:00	DARK	6	1,025	93	1	0	Sem Chuva	None
Row25	Braga	MAJOR	10998	N103,N14,N...	2019-02-06 18:00	DARK	12	1,025	87	1	0	Sem Chuva	Very_High
Row26	Braga	MAJOR	0	CM1348,CM...	2019-06-08 20:00	DARK	14	1,020	82	6	0	Sem Chuva	Medium
Row27	Braga	UNDEFINED	0	CM1348,CM...	2019-09-22 20:00	DARK	15	1,019	93	2	0	Sem Chuva	None
Row28	Braga	MAJOR	1945	CM1348,N1...	2019-11-04 16:00	LIGHT	13	1,010	93	3	0	aguaceiros	High
Row29	Braga	MAJOR	11081	N14,N103,N...	2019-06-05 17:00	LIGHT	15	1,017	67	3	0	Sem Chuva	Very_High
Row30	Braga	UNDEFINED	0	CM1348,CM...	2019-10-22 03:00	DARK	8	1,016	87	3	0	Sem Chuva	None
Row31	Braga	MAJOR	639	N14,CM134...	2019-09-28 18:00	DARK	17	1,019	82	1	0	Sem Chuva	Medium
Row32	Braga	UNDEFINED	123	CM1348,N1...	2019-07-05 10:00	LIGHT	20	1,017	88	4	0	Sem Chuva	Medium
Row33	Braga	UNDEFINED	232	CM1348,CM...	2019-10-19 01:00	DARK	13	1,011	100	6	0	Sem Chuva	Low
Row34	Braga	MODERATE	1018	N14,	2019-12-13 21:00	DARK	12	1,021	82	1	0	Sem Chuva	Medium
Row35	Braga	MAJOR	277	CM1348,CM...	2019-08-10 14:00	LIGHT	22	1,021	73	4	0	Sem Chuva	Low
Row36	Braga	UNDEFINED	0	CM1348,CM...	2019-10-31 23:00	DARK	17	1,024	100	3	0	Sem Chuva	None
Row37	Braga	MAJOR	9454	N103,N14,N...	2019-07-16 17:00	LIGHT	25	1,013	78	2	0	Sem Chuva	Very_High
Row38	Braga	UNDEFINED	0	CM1348,CM...	2019-10-04 03:00	DARK	16	1,019	93	0	0	Sem Chuva	None
Row39	Braga	UNDEFINED	565	CM1348,CM...	2019-05-10 23:00	DARK	13	1,021	100	1	0	Sem Chuva	Medium

Figure 2: Data Set

A partir de técnicas de tratamento de dados foi possivel criar novas features

geradas a partir de informação disponivel na feature record_date, foi separada a data da hora em duas novas features, foi criada a nova coluna do trimestre e das semans, assim como o dia da semana e a parte do dia.

Row ID	Data	Hora	Quarter	Month (number)	Week	Day of week (number)	Parte do Dia
Row0	2019-02-17	20	1	2	7	7	Noite
Row1	2019-02-14	12	1	2	7	4	Tarde
Row2	2019-09-03	19	3	9	36	2	Noite
Row3	2019-06-18	18	2	6	25	2	Noite
Row4	2019-05-08	13	2	5	19	3	Tarde
Row5	2019-05-25	21	2	5	21	6	Noite
Row6	2019-09-16	4	3	9	38	1	Madrugada
Row7	2019-08-15	11	3	8	33	4	Manhã
Row8	2019-02-21	6	1	2	8	4	Manhã
Row9	2019-05-06	17	2	5	19	1	Tarde
Row10	2019-09-17	14	3	9	38	2	Tarde
Row11	2019-10-10	20	4	10	41	4	Noite
Row12	2019-10-05	13	4	10	40	6	Tarde
Row13	2019-01-19	23	1	1	3	6	Noite
Row14	2019-10-17	13	4	10	42	4	Tarde
Row15	2019-09-12	13	3	9	37	4	Tarde
Row16	2019-11-15	8	4	11	46	5	Manhã
Row17	2019-07-24	6	3	7	30	3	Manhã
Row18	2019-10-03	15	4	10	40	4	Tarde
Row19	2019-04-17	10	2	4	16	3	Manhã
Row20	2019-12-23	6	4	12	52	1	Manhã
Row21	2019-05-19	13	2	5	20	7	Tarde
Row22	2019-05-11	19	2	5	19	6	Noite
Row23	2019-07-01	6	3	7	27	1	Manhã
Row24	2019-01-21	23	1	1	4	1	Noite
Row25	2019-02-06	18	1	2	6	3	Noite
Row26	2019-06-08	20	2	6	23	6	Noite
Row27	2019-09-22	20	3	9	38	7	Noite
Row28	2019-11-04	16	4	11	45	1	Tarde
Row29	2019-06-05	17	2	6	23	3	Tarde
Row30	2019-10-22	3	4	10	43	2	Madrugada
Row31	2019-09-28	18	3	9	39	6	Noite
Row32	2019-07-05	10	3	7	27	5	Manhã
Row33	2019-10-19	1	4	10	42	6	Madrugada
Row34	2019-12-13	21	4	12	50	5	Noite
Row35	2019-08-10	14	3	8	32	6	Tarde
Row36	2019-10-31	23	4	10	44	4	Noite

Figure 3: Novas Features geradas da **record_date**

A feature **Parte do Dia** foi concebida a partir da **Hora** e foram selecionados intervalos de forma a que cada parte do dia ficasse com frequencias identicas.

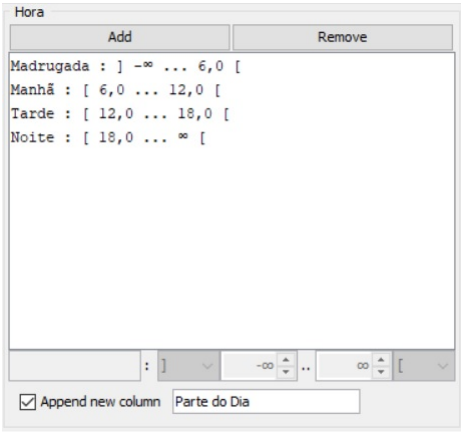

Figure 4: Configuração da Feature **Parte do Dia**

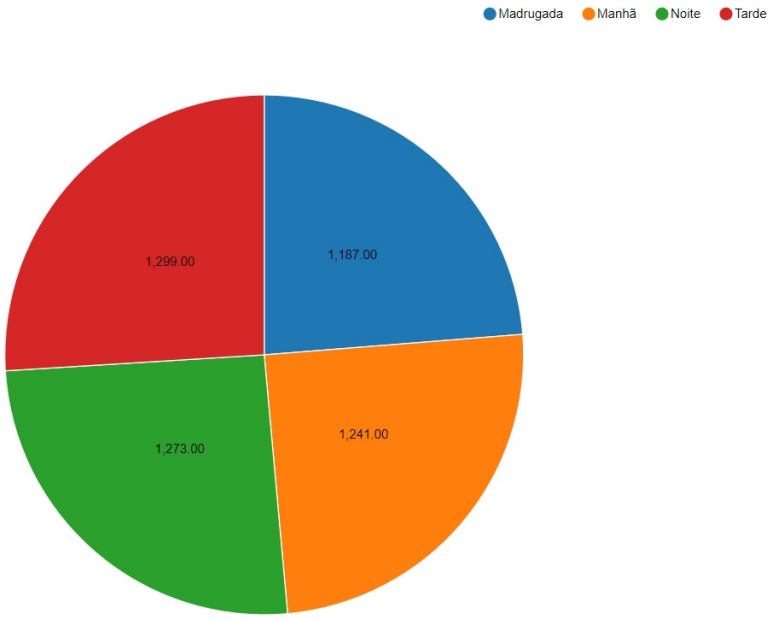

Figure 5: Distribuição da Feature **Parte do Dia**

Outra das transformações mais importantes foi sobre a feature **affected_roads**, em baixo podemos ver as novas colunas geradas pelas transformações de dados nessa feature.

Como se pode ver, as features ROAD CM, ROAD M, ROAD EM, ROAD IP, ROAD N e ROAD A representam o tipo de estrada. Se a estrada daquele tipo estiver afetada no acidente, o valor dessa feature vai ser "1", caso não esteja, vai ser "0", por exemplo, na "Row 2" a coluna "ROAD N" tem valor "1",

Row ID	S ROAD CM	S ROAD M	S ROAD EM	S ROAD IP	S ROAD N	S ROAD A	I NRROADS	S roads_affect	S repeat_road
Row0	0	0	0	0	1	0	2	A	ROADS DUPLICATED
Row1	0	0	0	0	1	0	3	B	CORRECT
Row2	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row3	1	0	0	0	1	0	9	C	ROADS DUPLICATED
Row4	1	0	0	0	1	0	6	C	CORRECT
Row5	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row6	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row7	1	0	0	0	0	0	3	B	ROADS DUPLICATED
Row8	0	0	0	0	1	0	2	A	ROADS DUPLICATED
Row9	0	0	0	0	1	0	47	C	CORRECT
Row10	1	0	0	0	1	0	3	B	CORRECT
Row11	1	0	0	0	0	1	3	B	CORRECT
Row12	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row13	0	0	0	0	0	0	0	A	NOT ROADS
Row14	1	0	0	0	1	0	9	C	CORRECT
Row15	1	0	0	0	1	0	4	B	CORRECT
Row16	1	1	1	0	1	0	30	C	CORRECT
Row17	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row18	1	0	1	0	1	0	8	C	CORRECT
Row19	0	0	0	0	1	0	2	A	CORRECT
Row20	0	0	0	0	0	0	0	A	NOT ROADS
Row21	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row22	0	0	0	0	0	0	0	A	NOT ROADS
Row23	1	0	0	0	1	0	3	B	ROADS DUPLICATED
Row24	0	0	0	0	0	0	0	A	NOT ROADS
Row25	0	0	0	0	1	0	21	C	ROADS DUPLICATED
Row26	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row27	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row28	1	0	0	0	1	0	9	C	CORRECT
Row29	1	0	0	0	1	1	36	C	CORRECT
Row30	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row31	1	0	0	0	1	0	3	B	ROADS DUPLICATED
Row32	1	0	0	0	1	0	3	B	CORRECT
Row33	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row34	0	0	0	0	1	0	1	A	CORRECT
Row35	1	0	0	0	0	0	2	A	ROADS DUPLICATED
Row36	1	0	0	0	0	0	2	A	ROADS DUPLICATED

Figure 6: Novas Features geradas a partir da

logo, estradas do tipo "N" estão afetadas pelo acidente. A feature "NROADS" representa o número de estradas afetadas pelos acidentes, está variavel tem muita importancia no modelo gerado porque se houver um grande número de estradas é quase certo que o acidente é grave, este valor é justificado com a seguinte caixa de bigodes.

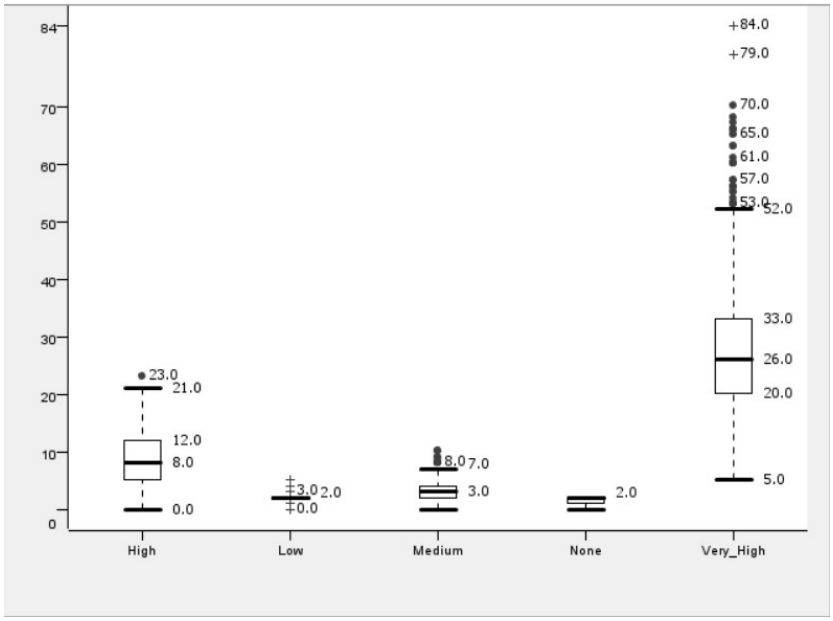

Figure 7: Box Plot NRROADS vs Count(accuracy)

A feature "roads_affect" é usada para categorizar a feature "NRROADS". É claro que se ambas estas features entrarem no modelo este terá problemas de multicolineariadade, mas foram as duas geradas para se perceber qual delas funcionará melhior no modelo. A feature conta com a categoria "A","B" e "C" e é dividida da seguinte forma:

```
String[] parts = c_affected_roads.split(",");
out_NRROADS = parts.length;

if (out_NRROADS >= 3 && out_NRROADS < 6 ){
    out_roads_affect = "B";
}else if (out_NRROADS >= 6){
    out_roads_affect = "C";
}else {
    out_roads_affect = "A";
}
```

Finalmente, a última feature a ser gerada ("repeat_road") revela se na feature das "roads_affect" existem estradas repetidas, não esxistem estradas, ou simplesmente existem estradas mas não repetidas.

Foram efetuadas mais algumas transformações, das quais categorizar, de uma forma diferente, a feature "avg_rain". Tendo o valor "0" caso a "avg_rain" tenha o valor "Sem Chuva", e o valor "1" caso contrário. Foi também criada uma feature para categorizar de forma diferente a feature **magnitude_of_delay**, caso seja "UNDEFINED", ficará com o valor "1", e toma o valor "0" caso contrário.

Foram acrescentadas todas as features numéricas de forma normalizada ao dataset para depois se comparar de que forma é que o modelo se comporta com algumas normalizações.

4.3 Modelação

4.4 Análise de Resultados

5 DataSet 1

5.1 Contexto do Dataset

5.2 Análise e Compreensão dos Dados

5.3 Modelação

5.4 Análise de Resultados

6 Conclusão