**Introdução**

Serve o presente relatório para explicitar a análise, preparação dos dados, extração de conhecimento, obtenção de resultados e análise crítica dos mesmos para o dataset da competição de Sistemas Baseados em Similaridade e o o dataset escolhido pelo grupo.

Os workflows criados visam a concessão e otimização dos modelos baseados em árvores usando um sistema de aprendizagem com supervisão de forma obter uma classificação final que no caso do dataset da competição diz respeito a uma previsão do número de incidentes rodoviários que irão acontecer numa cidade a uma dada hora e no caso do dataset escolhido pelo grupo diz respeito à previsão do vencedor de um jogo de League of Legends com dados apenas dos primeiros 10 minutos da partida.

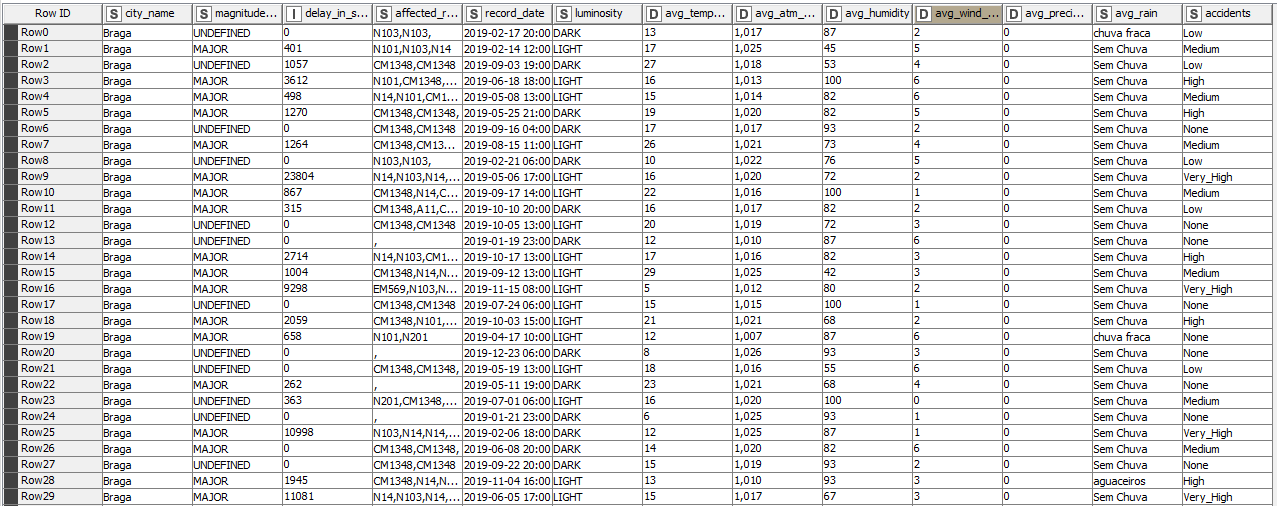
De forma a obter os melhores resultados nos modelos criados foi necessário por em prática as técnicas aprendidas na aula de preparação, extração de dados, seleção de features e hiper-parâmetros e validação de resultados.

**Dataset competição**

Para a competição foram-nos disponibilizados dois datasets. Um training\_data.csv e um test\_data.csv.

**Analise de dados**

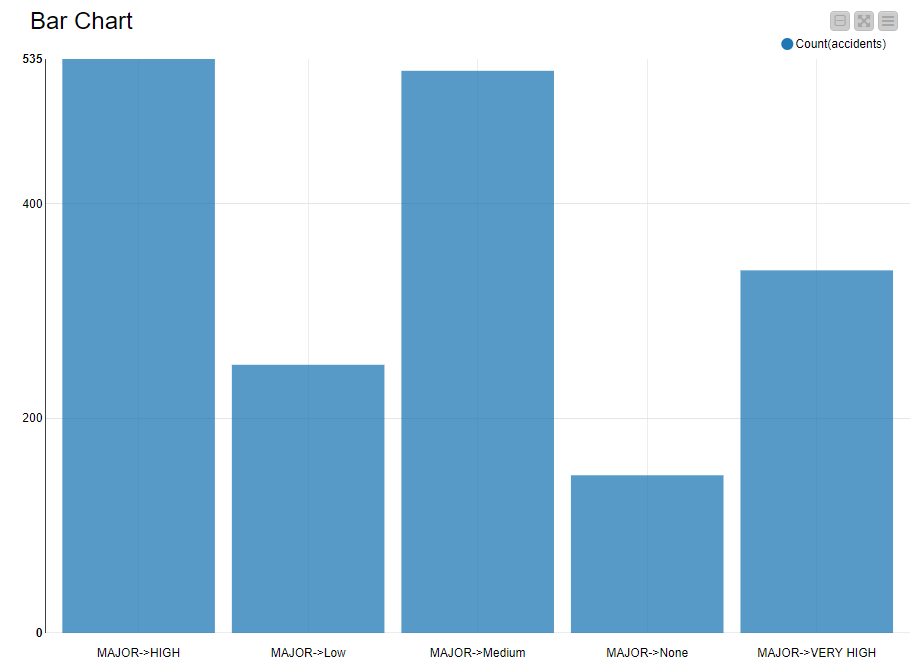
O taining set foi usado para treinar o modelo de Machine Learning criado. Neste encontra-se informação referente ao nível de incidentes rodoviários de cada registo para que treinar os modelos de aprendizagem.

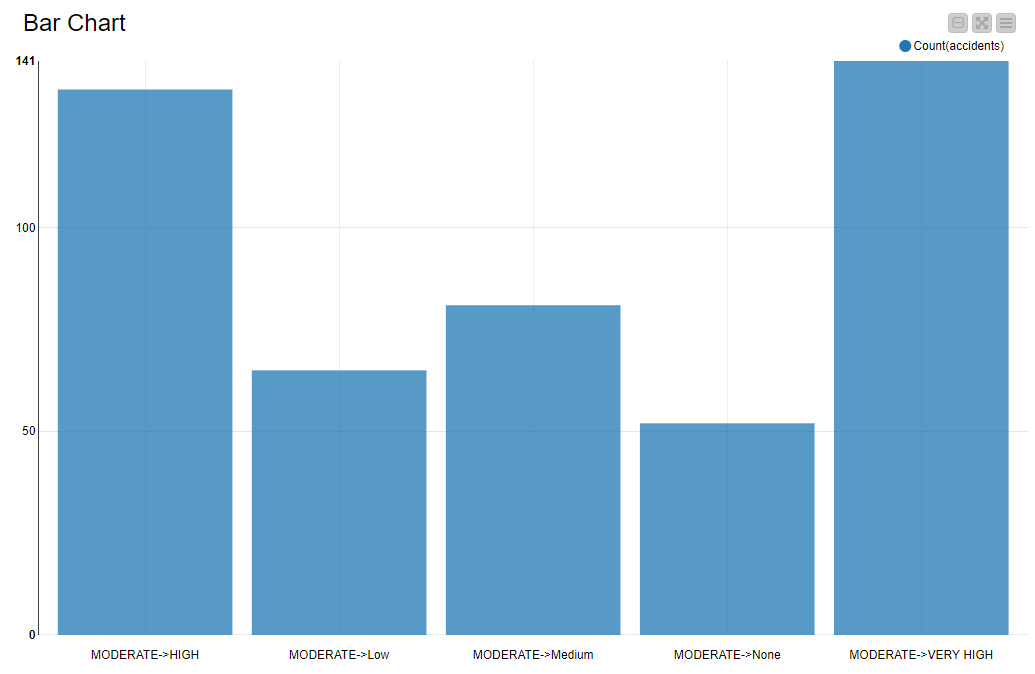


À primeira vista temos algumas features que intuitivamente parecem ter mais impacto na classificação final da intensidade dos acidentes do que outras. Começamos então isolar as essas feutures e ver graficamente o seu impacto.

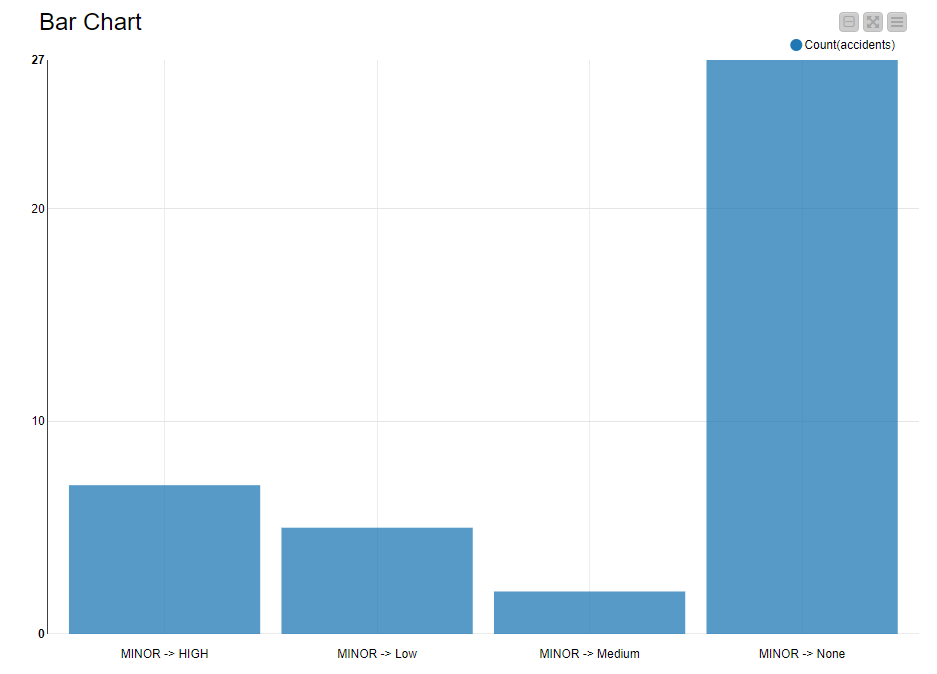
**Magnitude\_of\_delay**

Começando então pela magnitude do delay, podemos ver através dos seguintes gráficos de barras a distribuição do nível de acidentes rodoviários e a magnitude de atraso provocados pelos incidentes.

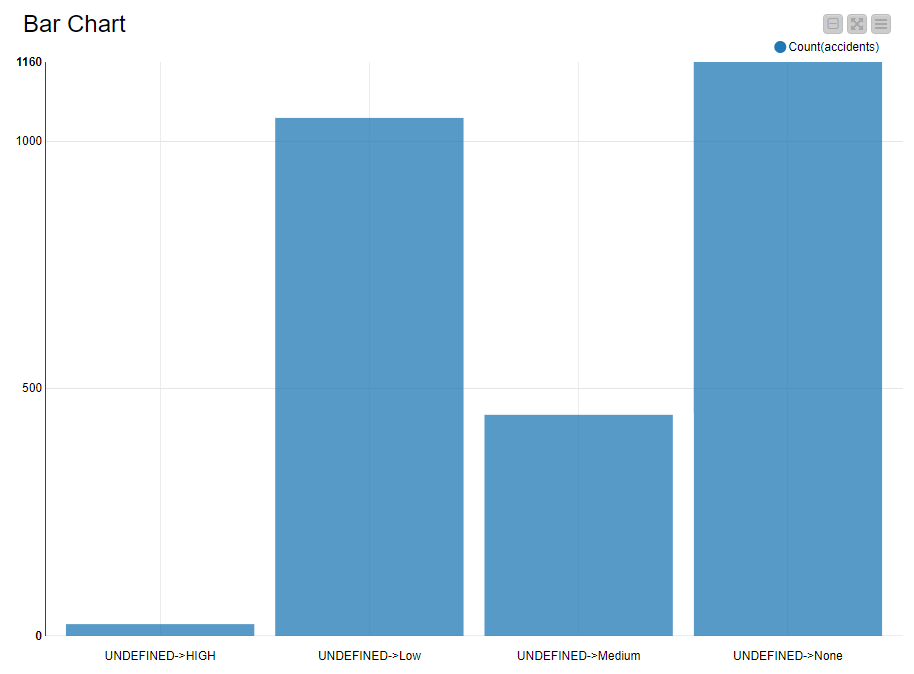


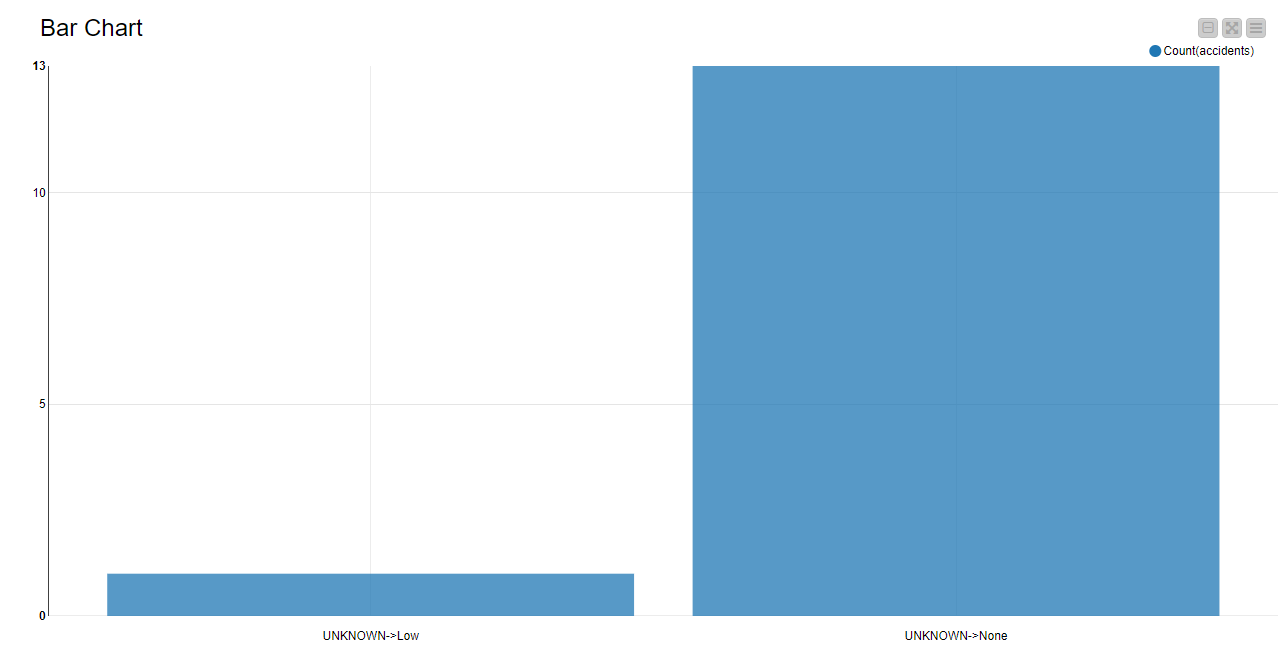


Como podemos ver quando existe uma magnitude alta ou moderada o delay o nível de acidentes tende a variar entre o “Very\_High”, “High” e “Medium”. Sendo que também existe uma percentagem considerável de “Low”.



Quando a magnitude é menor temos um maior nível de incidentes “None”.





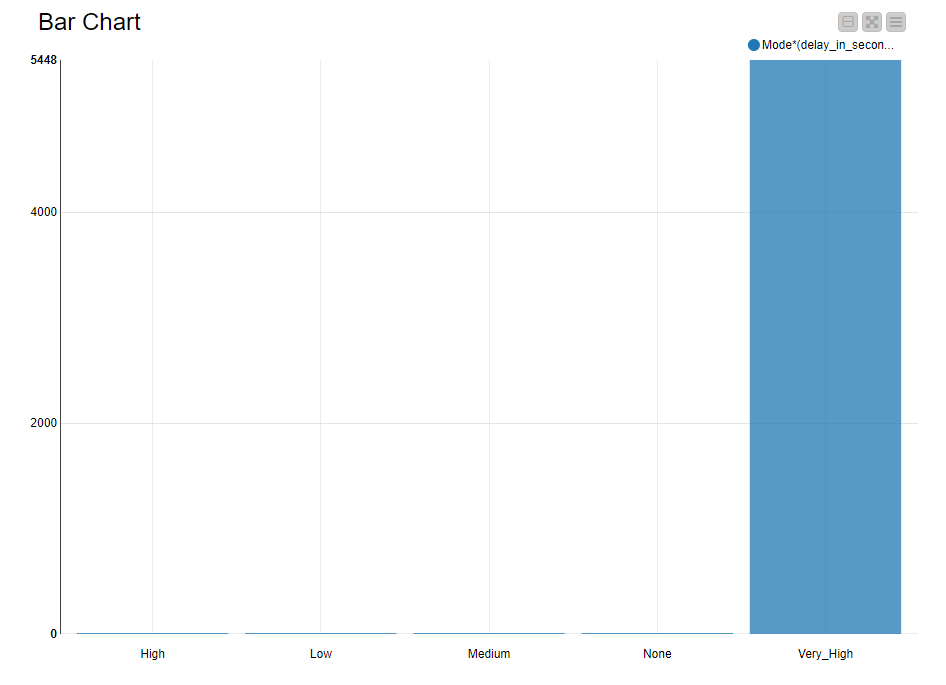
Quando a magnitude se apresenta indefinida ou desconhecida temos muito provavelmente um nível de incidentes baixo ou até inexistente.

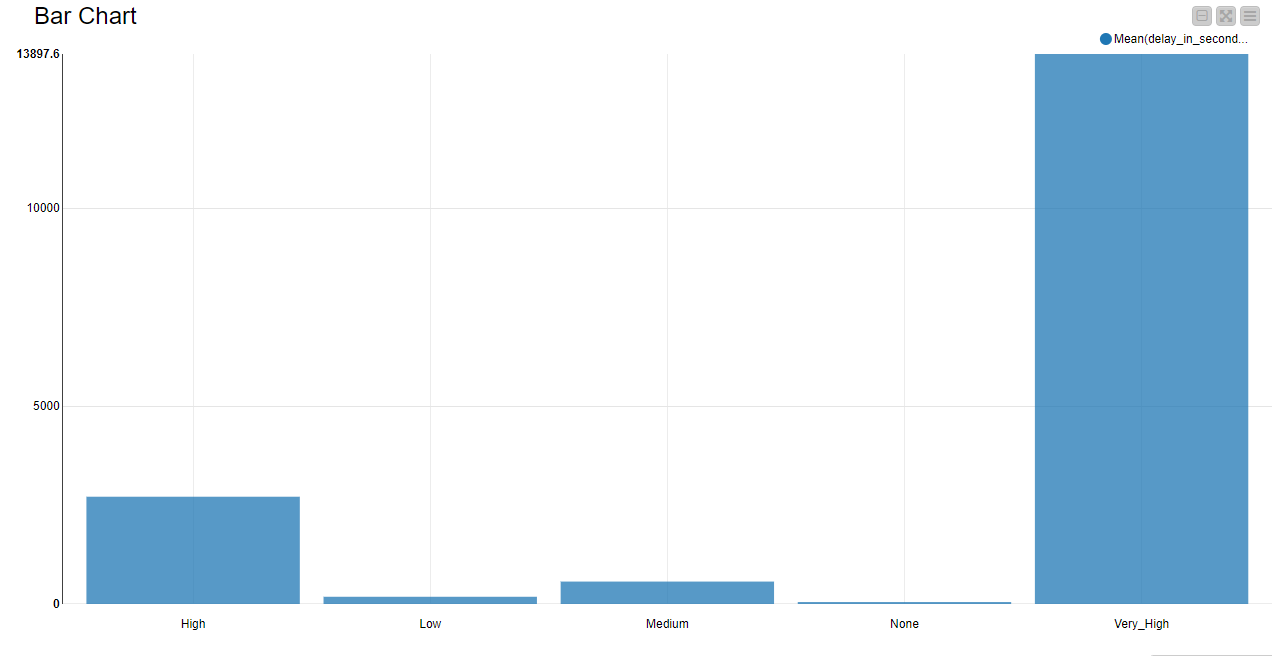
Como podemos ver a magnitude de atraso permite algumas previsões de forma intuitiva para prever o nível de incidentes. O que torna esta feature, à partida, uma forte candidata para a escolha do modelo final.

**Delay\_in\_secs**

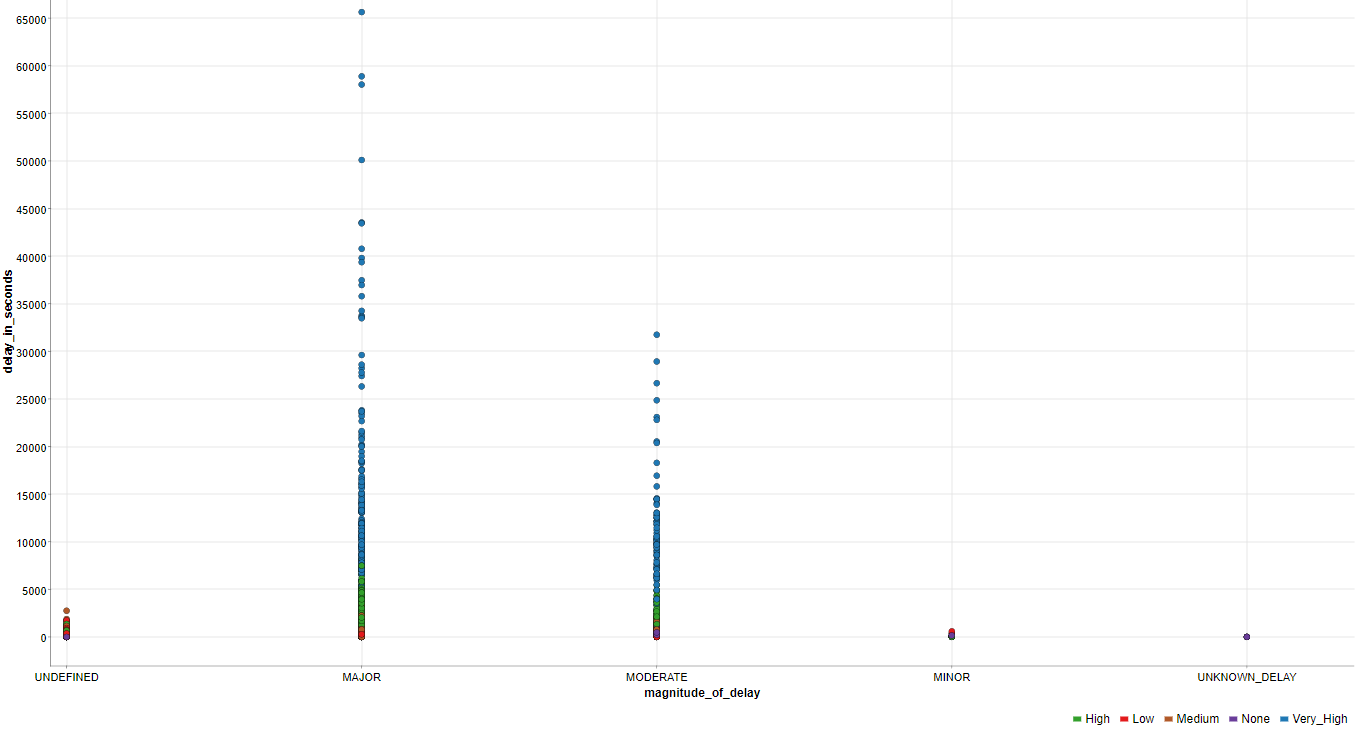
O dataset apresenta ainda o delay em segundos, provocado pelos incidentes que se verificam.

Começamos então por ver a média e moda de delay para cada nível de incidentes.





Destes valores podemos afirmar que o delay\_in\_secs pode ser usada como um forte indicativo para níveis de incidentes “Very\_High”.



Relacionando agora a magnitude de delay com o delay em segundos e o nível de incidentes podemos ver que há uma relação direta entre o tempo de delay e o nível de incidentes. Sendo que quanto maior for o delay maior a probabilidade de um nível de incidentes alto. Também podemos retirar que quanto maior for a magnitude do delay maior será a probabilidade de temos um grande delay em segundos.

**affected\_roads**

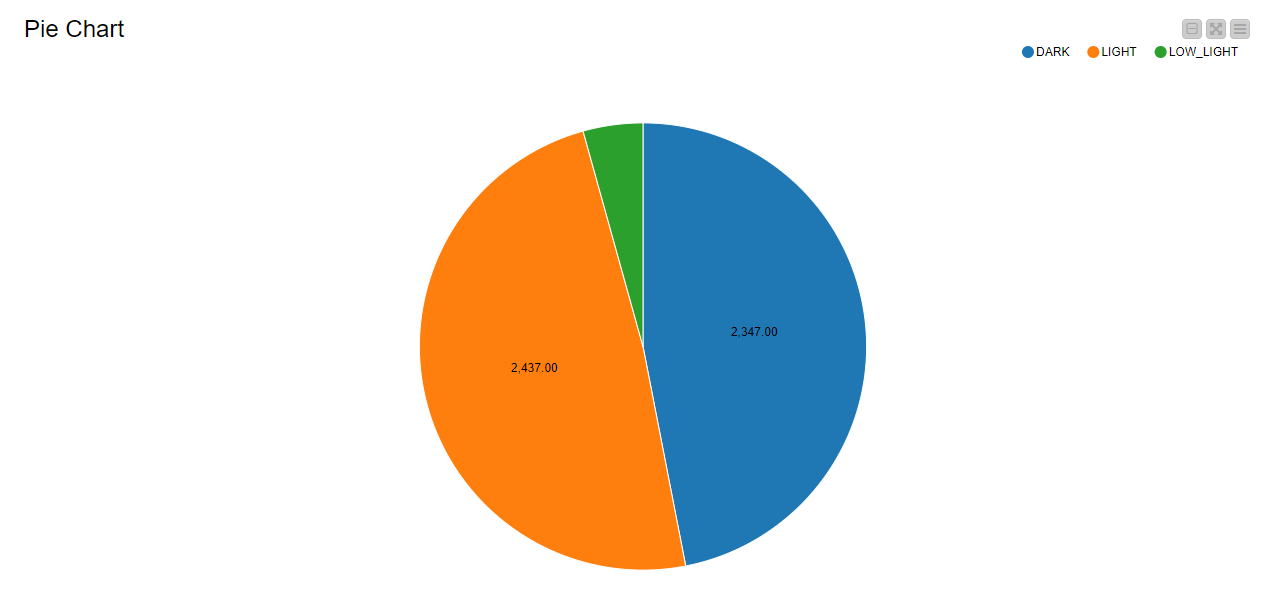
Uma das features disponíveis no dataset mais interessantes é a affected\_roads. No entanto da forma como os dados são fornecidos não é possível extrair informação relevante. Só após termos feito tratamento de dados, como mais à frente no relatório vai ser descrito, é que foi possível extrair informação relevante para o modelo.

**Record\_date**

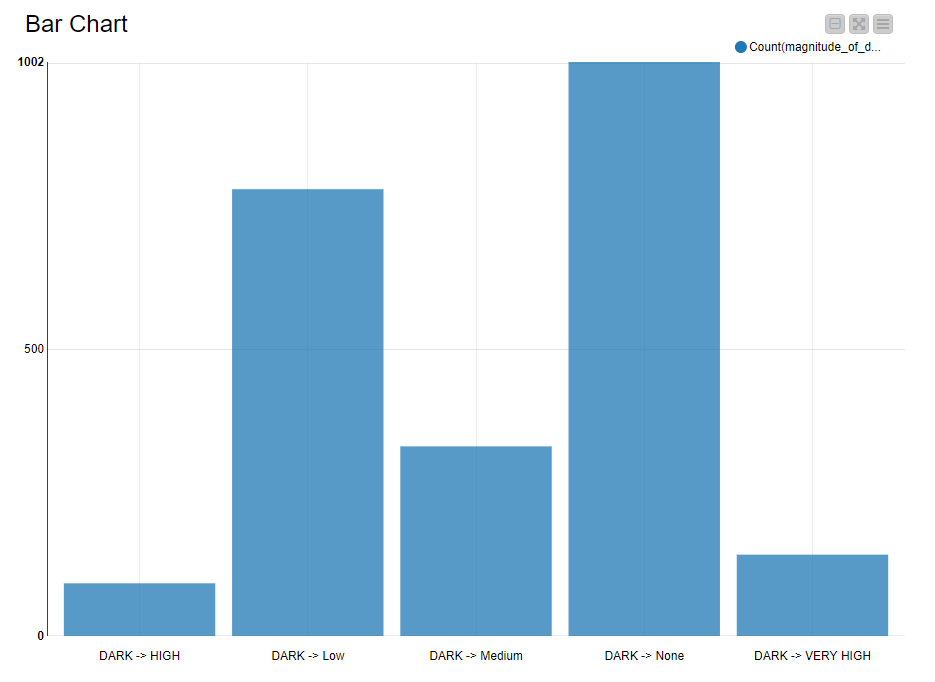
Tal como a affexted\_roads esta feature embora bastante relevante numa análise inicial não tem grande utilidade. Só após termos feito o tratamento de dados é que foi possível extrair informações relevantes para o modelo.

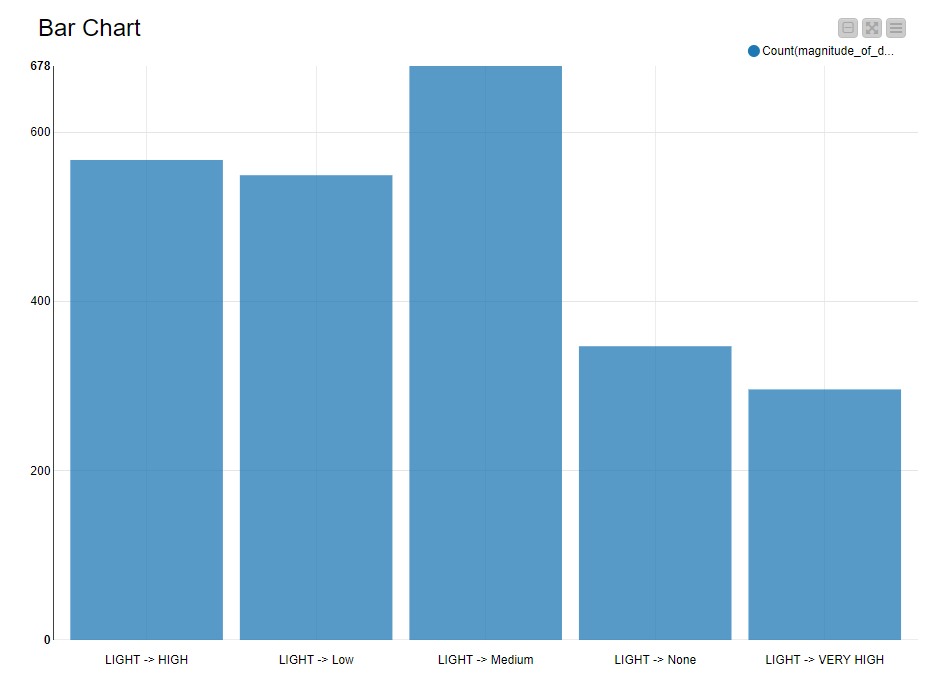
**Luminosity**

Uma das preocupações que tivemos ao analisar esta feature foi a distribuição de dados. Mas como podemos ver através do seguinte pie chart, temos uma percentagem de quase 50% de incidentes entre a luminosidade “Dark” e “Light”. Sendo que os únicos dados não muito relevantes são os da “Low\_Light” devido aos poucos registos.



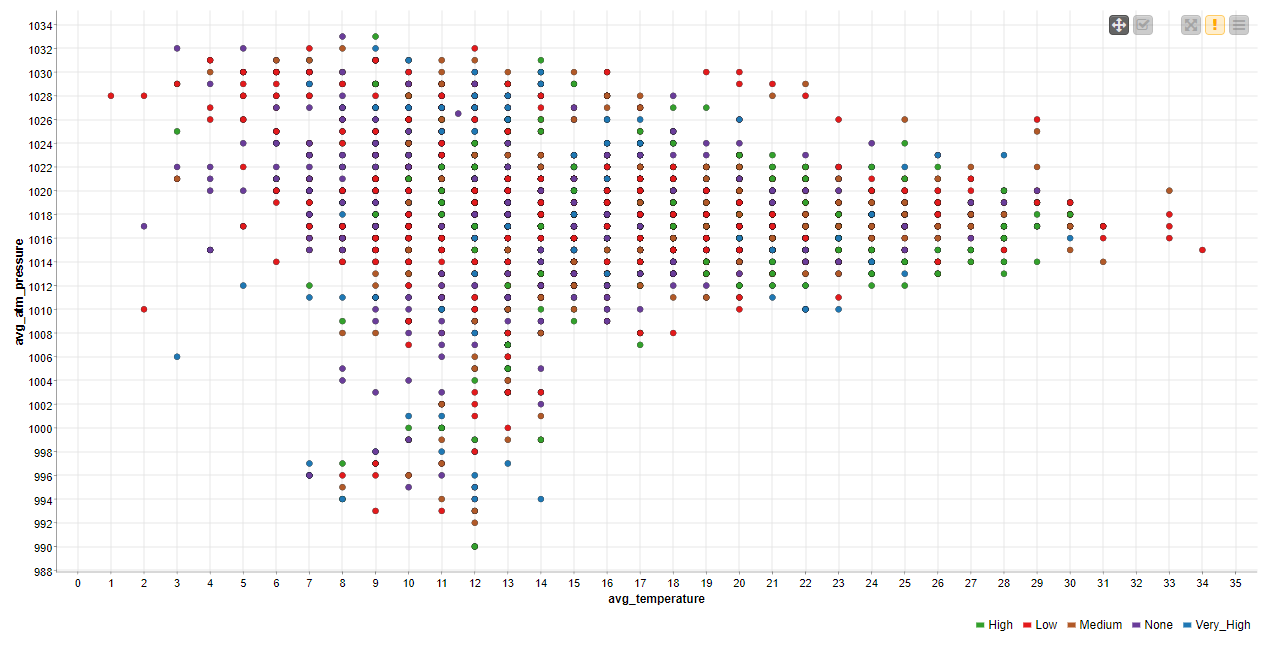
Após a análise do nível de incidentes nas diferentes condições de luminosidade é possível concluir que os registos com uma luminosidade “Light” se encontram bem distribuídos entre os diferentes níveis de incidentes, já para uma luminosidade “Dark” existe um maior número de registos com níveis de intensidade “Low” ou “None”.

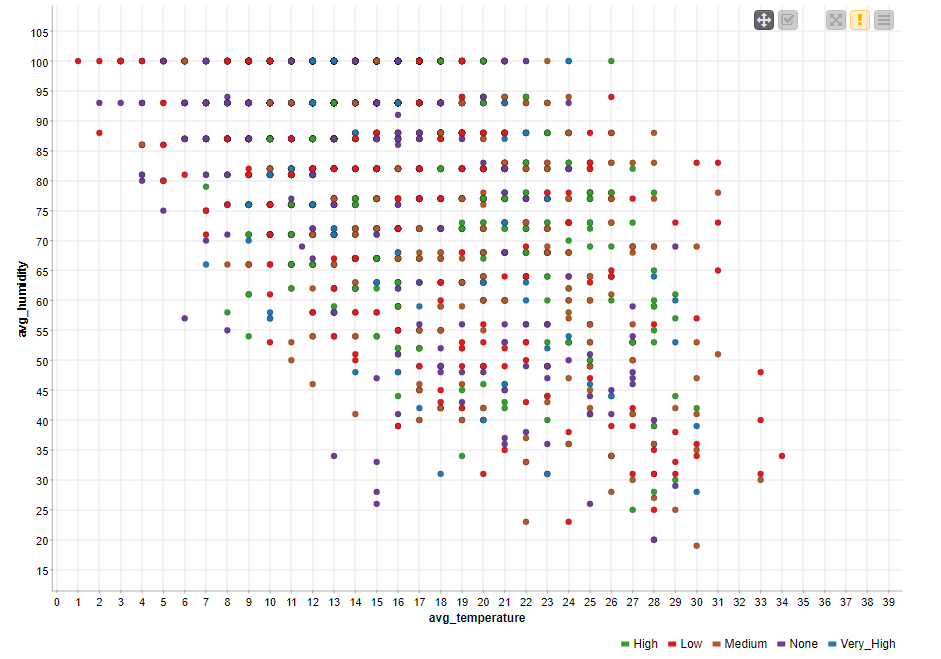


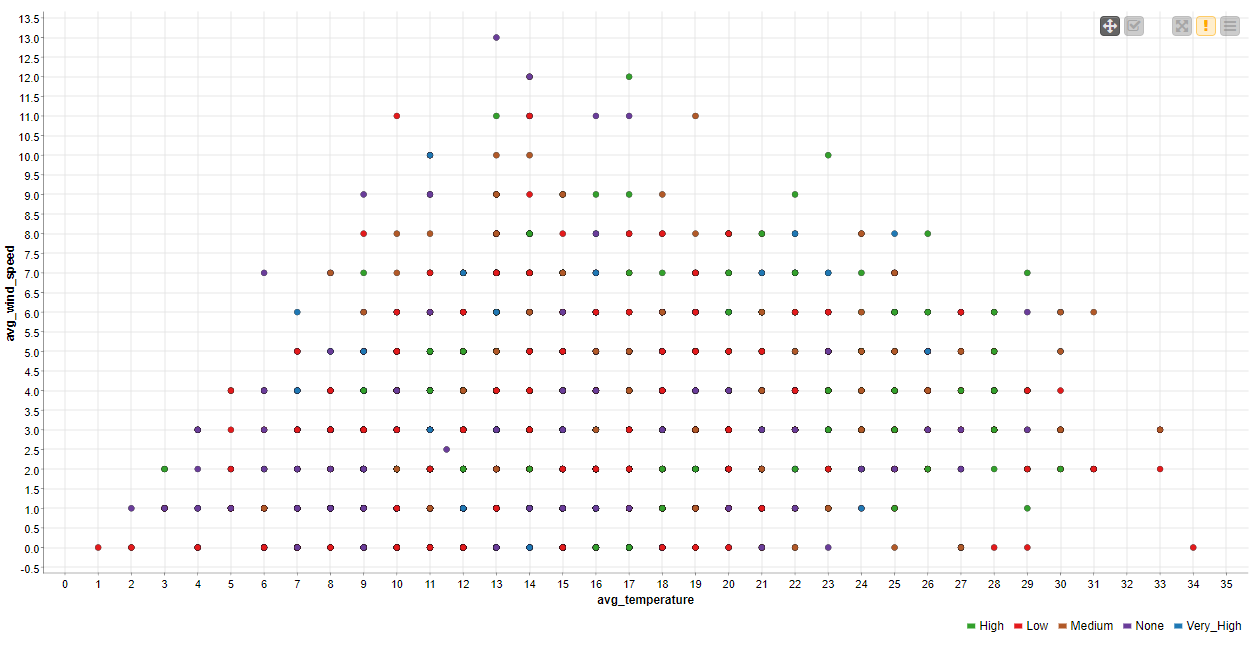


**avg\_temperature & avg\_atm\_pressure & avg\_humidity & avg\_wind\_speed**

**Olhando para a distribuição dos registos nos seguintes scatter plots à primeira vista parece não haver uma relação direta entre as caraterísticas metrológicas e o nível de incidentes. Podemos tirar esta conclusão devido à distribuição tendencialmente homogenia do nível de intensidade dos registos. Não podemos, no entanto, excluir estas features da seleção final para o treino do modelo, uma vez que os modelos de árvores podem encontrar uma combinação de features que englobem estas e que o ajudem no processo de classificação do nível de incidentes.**





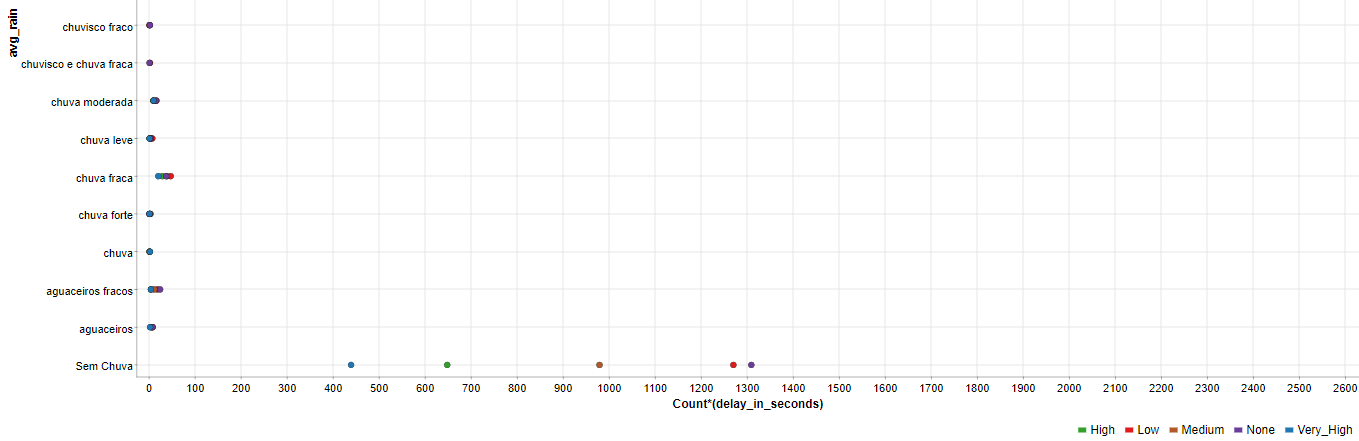


**avg\_precipitation**

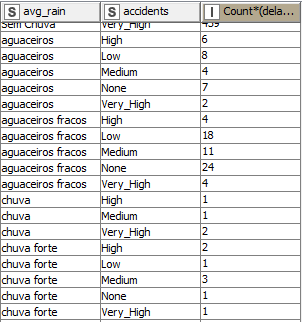
Ao contrário do resto das features relacionadas com a metrologia o uso da avg\_precipitation não vai ser importante para o treino do modelo e pode potencialmente ser projudicial. Isto devido ao valor dos dados nos registos do dataset treino. Se formos analisar a coluna referente a esta feature no dataset apenas três registos apresentam um valor diferente de 0. Podemos por isso considerar estes 3 registos como sendo outliers, ou seja, valores que podem comprometer de forma negativa os resultados de uma análise.

**avg\_rain**

De forma a analisar qual o impacto da feature avg\_rain foi feito o seguinte gráfico onde se apresenta o número de incidentes ocorridos para os diferentes níveis de incidentes perante as diferentes possibilidades de avg\_rain.



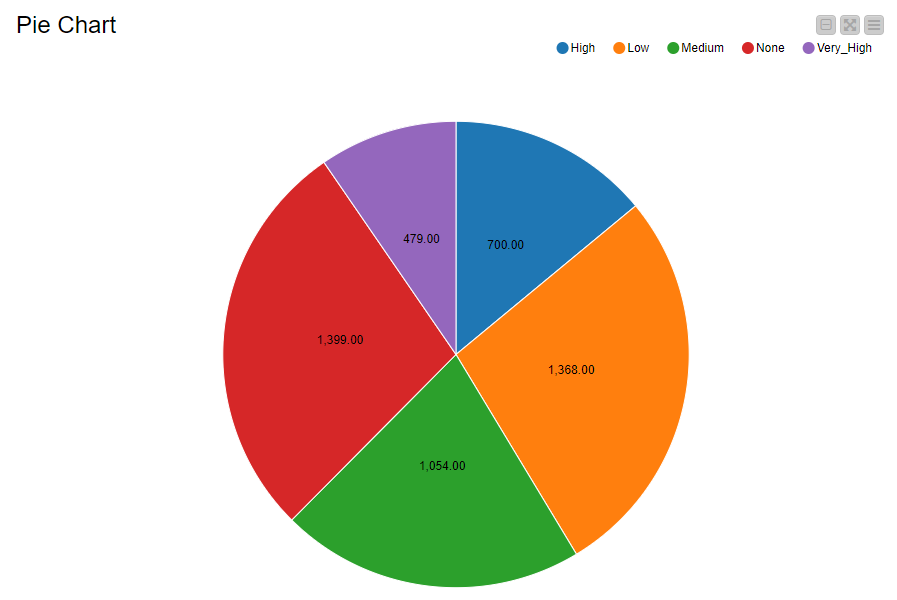
Olhando apenas para este gráfico parece que a avg\_rain pode ser um forte indicativo para um determinado nível de intensidade de incidentes, mas devido ao número de registos muito reduzido em algumas das combinações entre features, por exemplo, as que a seguir se apresentam podemos assumir estes valores como outliers o que significa que podem até ser prejudiciais para a classificação final do modelo treinado.



**accidents**

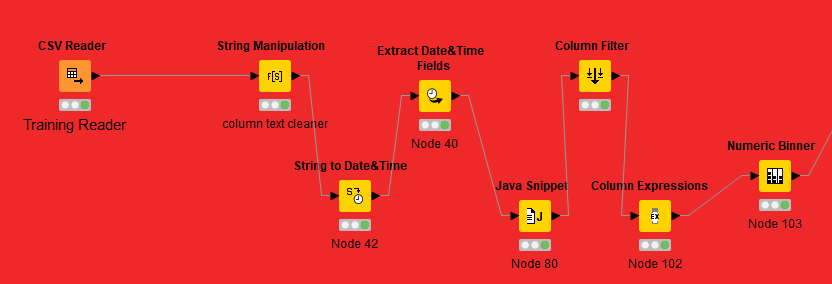
Finalmente temos a feature accidents que se refere ao nível de incidentes rodoviários que se verificam numa recor\_date. São os valores desta feature que queremos prever e é esta coluna que usamos para fazer uma aprendizagem supervisionada do modelo e é através desta que calculamos a precisão da solução encontrada.

Foi feita uma análise à distribuição dos dados e como se pode ver através do seguinte pie chart temos uma distribuição mais ou menos equilibrada entre os diferentes níveis de intensidade no dataset treino.



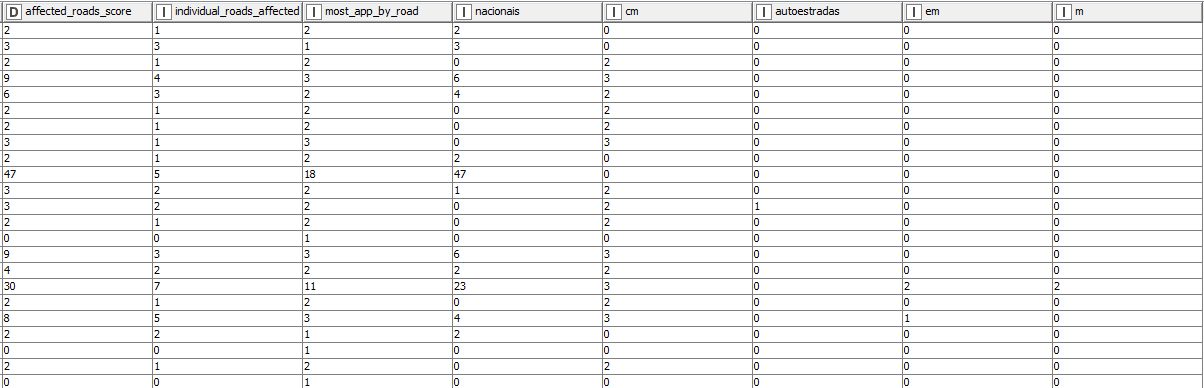
**Tratamento dos dados**

Embora uma grande parte dos dados do dataset treino se encontre pronto para treinar o modelo, foi, no entanto, necessário proceder ao tratamento e filtragem de algumas das features.



Começamos por retirar caracteres estranhos da coluna affected\_roads. Passamos depois à extração número do mês, semana, dia do ano, dia do mês, dia da semana e a hora da timestamp que se encontra na coluna record\_date.

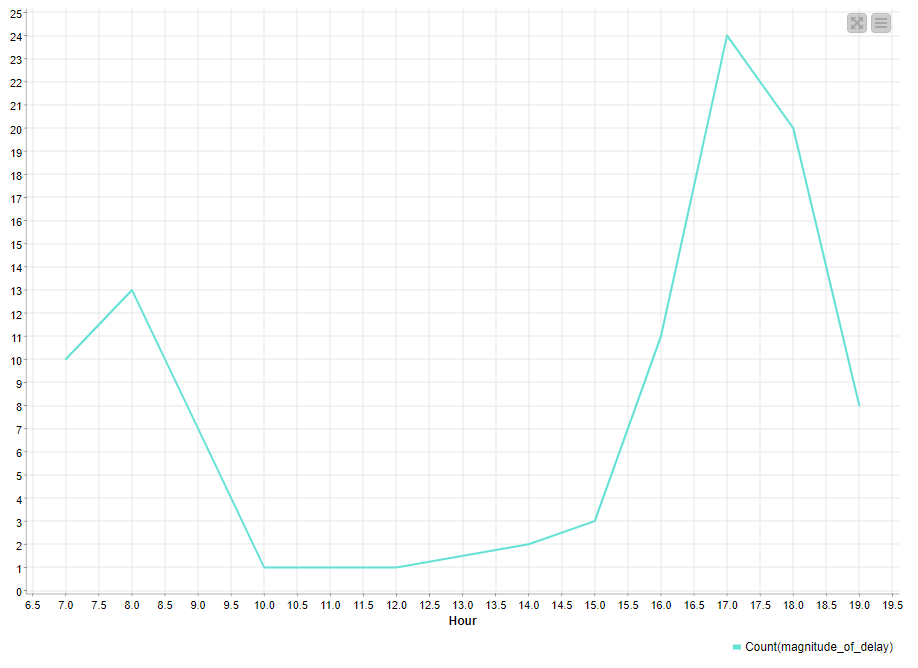
No nodo Java Snippet tentamos extrair o máximo de informação da coluna affected\_roads. Deste nodo surgem as seguintes colunas:



A coluna affected\_roads\_score indica o número com repetições de estradas afetadas, a coluna individual\_roads\_affected indica o número de estradas diferentes afetadas e a coluna mos\_app\_by\_road indica qual das estradas foi mais afetada.

O resto das colunas dizem-nos para cada registo qual a quantidade de estradas afetadas de cada tipo de estrada.

Foi ainda uma análise aos dias de semana e fins-de-semana e às horas de ponta e de que forma estas afetam a intensidade de acidentes.



(segunda very\_high)

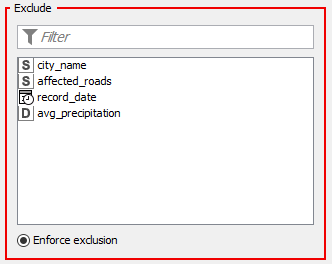


(dom very\_high)

Através da análise das horas de ponta para cada nível de incidente em cada dia da semana, foi-nos possível desenvolver um nodo que olhando para o dia da semana e hora calcula se estamos num dia da semana, ou fim de semana e se estamos ou não perante uma hora de ponta.

Agrupamos ainda os dados de delay\_in\_seconds em intervalos, para avaliar se o modelo se comportava melhor quando trabalhava sobre intervalos ou sobre todos os valores dos registos.

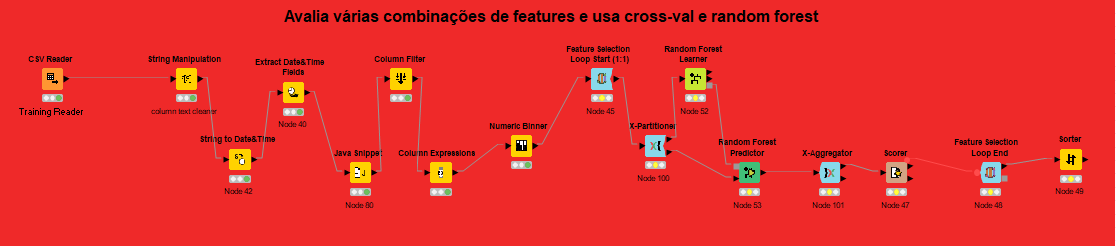
Finalmente, antes de treinar o modelo retiramos as features que não são relevantes para o modelo, tais como:

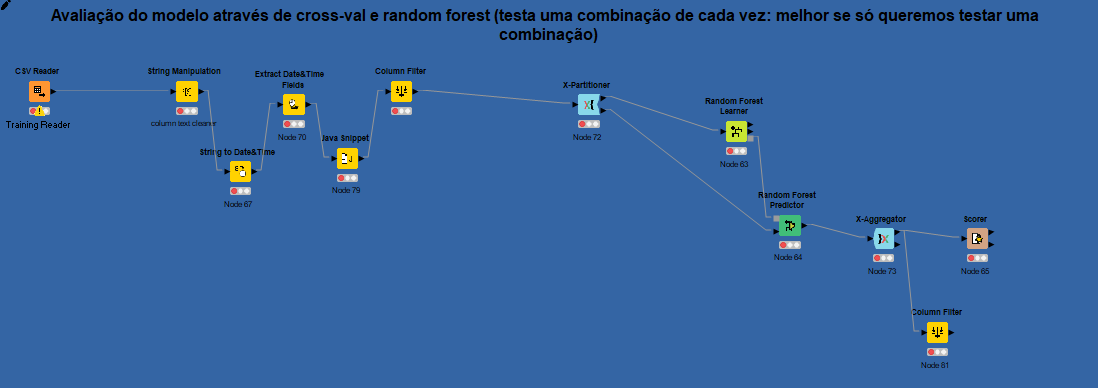


E para além destas ainda excluímos as colunas não selecionadas pelo feature selection, como vai ser referido mais à frente no relatório.

**Feature Selection**

De forma a selecionar a combinação de features que oferece uma melhor precisão ao modelo foram desenvolvidos dois worklows que nos permitiram avaliar as melhores combinações de features e o impacto que features individuais têm na precisão do modelo. Estes worflows são constituídos por técnicas aprendidas nas aulas de feature selection e cross validation para validação de resultados.





Após uma extensa análise, de várias tentativas e combinações chegamos à seleção de features que dá ao modelo a melhor precisão:

* magnitude\_of\_delay
* delay\_in\_seconds
* Day of month
* Day of year
* Hour
* Day of week (number)
* Nacionais
* Cm
* avg\_atm\_pressure
* Week

(Justificação para exclusão horas de pico e delay\_in\_seconds\_bin)

