[MEST] Previsão de novos comentários

Grupo 4E:
- Benedita Bacelar, up201909937
- Hugo Fernandes, up201909576
- Ricardo Pinto, up201806849

Problema

"O meu *post* vai ter comentários nas próximas 24h?"

- Problema estatístico de classificação:
  - Análise de um conjunto de dados classificado.
  - Determinação de tendências e correlações.
  - Classificação de novos dados.

Dados

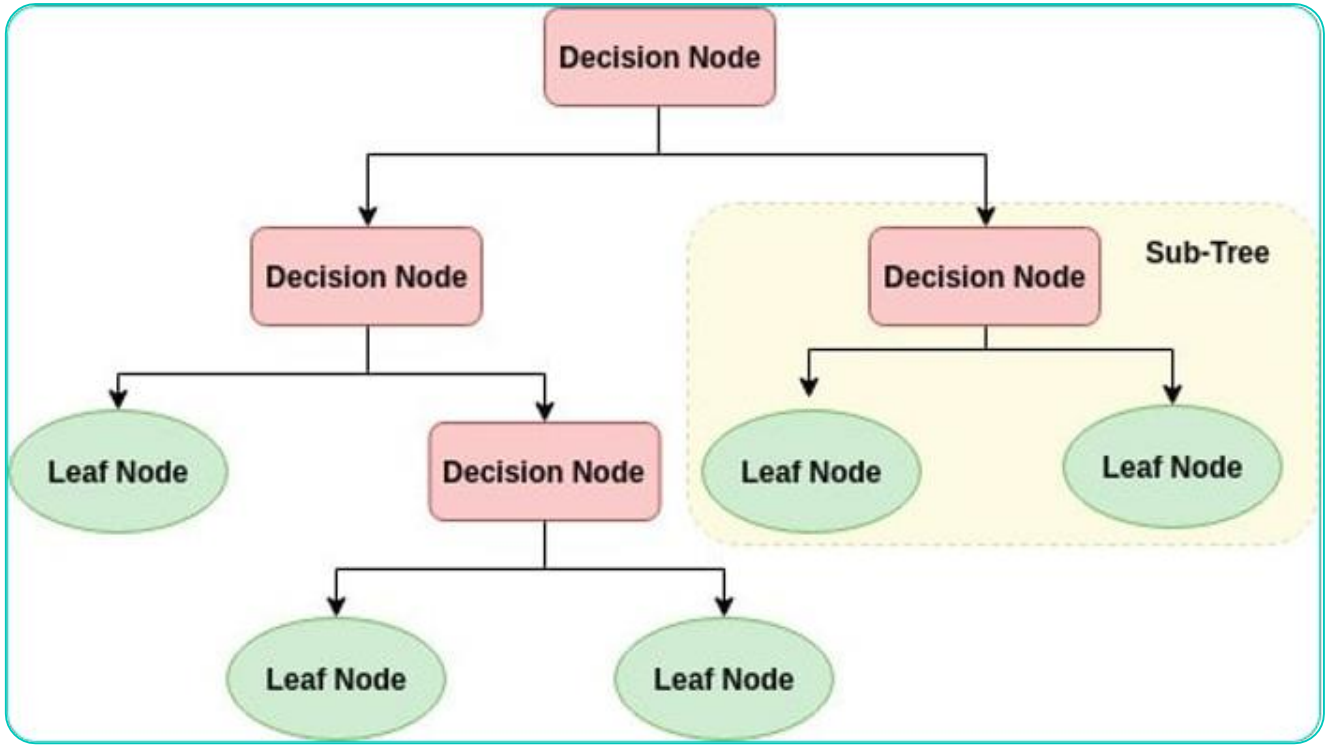
- Compostos por 14 valores.
  - 1 valor de identificação (ID).
  - 1 valor de classificação (has_new_comments).
  - 10 valores de variáveis que caracterizam cada entrada na tabela.
  - 2 variáveis de texto que caracterizam dias da semana.
- Dois ficheiros de dados:
  - dev.csv ficheiro com 10,000 entradas previamente classificadas.
  - new.csv ficheiro com 1000 entradas que devem ser classificadas pelo grupo de trabalho e submetido no Kaggle.
  - Um outro ficheiro foi disponibilizado de forma a demostrar em que forma deviam ser submetidos ficheiros no Kaggle.

Preparação dos dados

- Foram testados diferentes modelos de previsão que manipulavam as variáveis como por exemplo modelos que utilizam apenas algumas das variáveis com elevada correlação com a variável de classificação.
- Modelos foram também optimizados com uso de operadores do rapidminer em termos de critério principal de avaliação e outras componentes configuráveis do algoritmo.
- Tentou-se otimizar o conjunto de dados do ficheiro dev.csv de forma a remover outliers contudo a forma como os tentamos identificar não funcionou, devido a quantidade de entradas o rapidminer tinha problemas de memória e não conseguiu apresentar resultados.

Decision Tree

- Algoritmo de aprendizagem supervisionada.
- Pode ser utilizado para problemas de classificação.
- Permite a criação de um modelo que pode ser usado para prever o valor da variável alvo, através de simples decisões que ele toma.



Random Forest

- Consiste num numero elevado de Decision Trees que trabalham em conjunto.
- Cada uma das Decision Trees chega a uma conclusão sobre a variável alvo.
- Todas as conclusões são consideradas e a decisão que tiver mais votos das arvores é a que o modelo acredita que será o resultado mais provável.

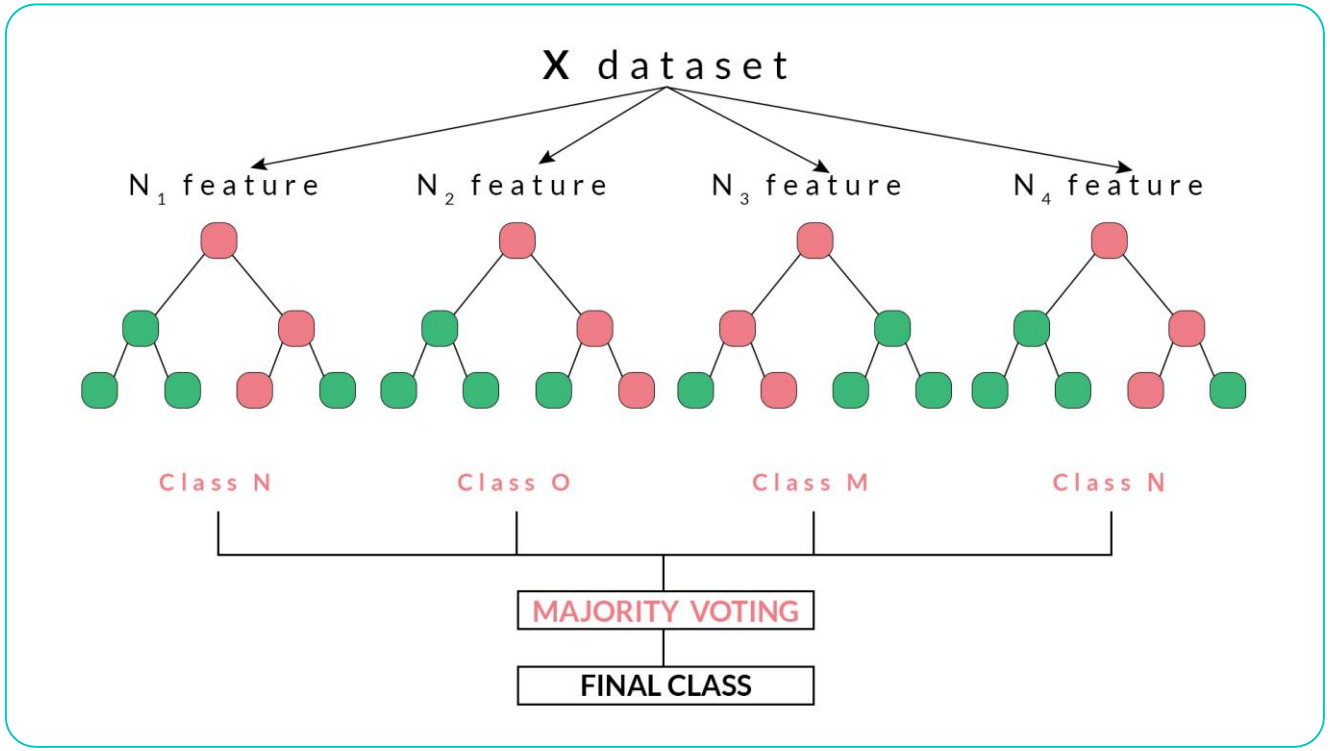


Tabela de Coorelações

Attributes	page_likes	page_interaction	page_category	tot_no_comments_bef	no_comments_24h	no_comments_48_24h	no_comments_24h_after_pub	delta_48_24h	character_count_post	no_shares_post	has_new_comments
page_likes	1	0.772	-0.010	0.059	0.046	0.040	0.059	0.007	-0.000	0.465	-0.035
page_interaction	0.772	1	-0.118	0.274	0.214	0.180	0.275	0.038	-0.006	0.463	-0.161
page_category	-0.010	-0.118	1	-0.142	-0.093	-0.092	-0.144	-0.006	0.033	-0.045	0.127
tot_no_comments_bef	0.059	0.274	-0.142	1	0.675	0.714	0.997	0.009	0.010	0.429	-0.304
no_comments_24h	0.046	0.214	-0.093	0.675	1	0.193	0.707	0.669	-0.005	0.359	-0.291
no_comments_48_24h	0.040	0.180	-0.092	0.714	0.193	1	0.701	-0.601	0.012	0.276	-0.199
no_comments_24h_after_pub	0.059	0.275	-0.144	0.997	0.707	0.701	1	0.045	0.009	0.431	-0.307
delta_48_24h	0.007	0.038	-0.006	0.009	0.669	-0.601	0.045	1	-0.013	0.084	-0.087
character_count_post	-0.000	-0.006	0.033	0.010	-0.005	0.012	0.009	-0.013	1	0.008	-0.028
no_shares_post	0.465	0.463	-0.045	0.429	0.359	0.276	0.431	0.084	0.008	1	-0.141
has_new_comments	-0.035	-0.161	0.127	-0.304	-0.291	-0.199	-0.307	-0.087	-0.028	-0.141	1

Resultados

Resultado mais alto obtido

- Algoritmo: Random Forest.
- Public Score: 0.85200
- Private Score: 0.81000
- Otimizado utilizando ferramentas do Rapidminer.
- Escolhido devido aos resultados provenientes de uma comparação de algoritmos realizada no rapidminer para o ficheiro fornecido.

Resultado que mais variou

- Algoritmo: Decision Tree.
- Public Score: 0.85400
- Private Score: 0.79200
- Optimizado utilizando ferramentas do rapidminer.
- Escolhido devido à representação da sua curva de ROC
- Submissão com Public Score mais elevada do grupo.

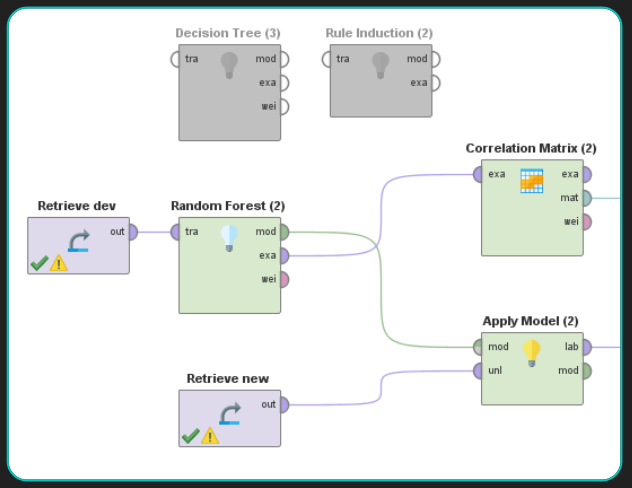
Conclusão

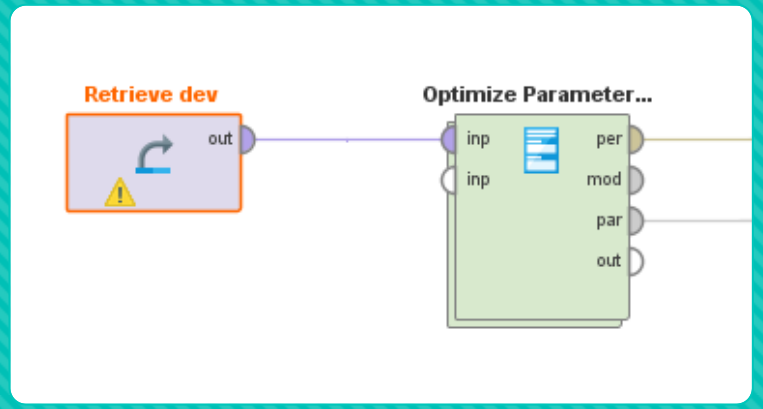
Através de uma análise dos resultados que obtivemos podemos concluir que, ambos os modelos submetidos estavam *overfitted* para os dados fornecidos, consideramos também que com algumas alterações não só aos modelos como também aos dados e parâmetros utilizados, como por exemplo a remoção de *outliers,* seria possível criar modelos mais corretos para este tipo de dados.

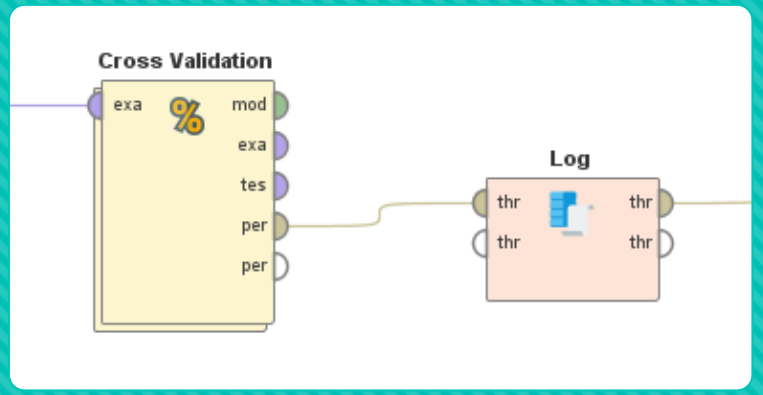
O estudo poderia, potencialmente, ter sido mais eficaz se fossemos a analisar a categoria de cada página. Apesar de não sabermos o tipo de categoria a que cada número corresponde, estas poderiam de qualquer forma ser utilizadas para estimar a variável alvo com mais êxito.

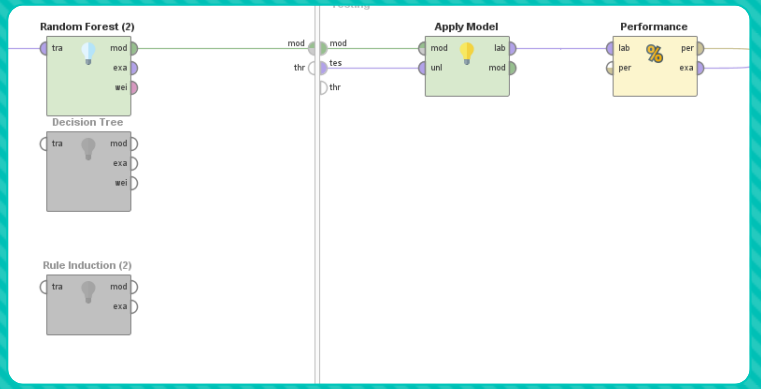
Apresentação privada

- Processo principal, composto por:
  - Ambos os conjuntos de dados dev e new.
  - Algoritmo de escolha (R. Forrest, D. Tree, …).
  - Criação de uma matriz de corelações para o modelo utilizado.
  - Aplicação do modelo aos dados não classificados.









Outros Processos

Processo que utiliza o operador de otimização de parâmetros para obter a melhor configuração para o algoritmos para os dados fornecidos no dev.csv.

Curva de ROC

- Inicialmente a curva de ROC do algoritmo Decision Tree era a que indicava melhor performance.
- Quanto mais otimizados eram os outros algoritmos mais estes ultrapassavam o algoritmo Decision Tree.
- Após as otimizações os melhores algoritmos de acordo com não só o gráfico mas também os resultados são Random Forest e Rule Induction.