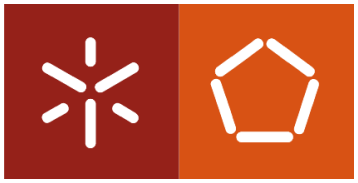

Universidade do Minho
Mestrado em Engenharia Informática
SBS

Trabalho Prático Nº1

Pedro Ribeiro - PG42848

Professor Doutor César Analide Freitas Silva Costa Rodrigues
Professor Doutor Bruno Filipe Martins Fernandes

Mestrado em Engenharia Informática
2021

Conteúdo

Capítulo 1

Introdução

No âmbito na unidade curricular "Sistemas Baseados em Similaridade", do perfil "Machine Learning: Fundamentos e Aplicações"foi desenvolvido este trabalho pratico que se foca na exploração de dois datasets, um fornecido pelos docentes da UC e outro escolhido pelo autor do trabalho pratico.

1.1 Conjuntos de dados

O dataset fornecido consta com os registos de incidentes rodoviários na cidade de Braga no período de 2019, contendo dados como a magnitude de atraso causado perante as estradas afetadas, o atraso em segundos provocado pelo incidente, entre outros.

O objetivo é prever a magnitude de incidentes de cada registo com o resultado possível associado, ou seja, que englobe o mínimo de erro possivel na previsão,

O segundo dataset foi criado por Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) e Paulo Rita (ISCTE-IUL) em 2014, e consta com informações sobre campanhas de marketing de um Banco de Portugal, conduzidas particularmente através de chamadas de telemóvel.[2] Este conjunto de dados consta com um total de 21 colunas, com informações pessoais dos inquiridos, dados relativos a campanha no ato do registo e dados económicos como por exemplo a taxa de euribor naquele momento.

Capítulo 2

Metodologia

2.1 *CRISP-DM*

Para alcançar uma qualidade nos modelos produzidos foi utilizada a metodologia *Cross Industry Standard Process for Data Mining*, ou *CRISP-DM*. Publicado em 1999, tornou-se a metodologia mais comum para *data mining* e projetos de *data science*, até ao dia 30 de junho de 2020 este foi o modelo mais pesquisado no google por uma margem bastante grande.[1]

O CRISP-DM é dividido em 6 fases:

- **Compreensão do modelo de negócio**, qualquer projeto começa com uma compreensão profunda das necessidades do cliente. Projetos de mineração de dados não são exceção e o CRISP-DM reconhece isso.
- **Compreensão dos dados**, depois da compreensão do modelo de negócio o CRISP-DM direciona o foco para identificar, coletar e analisar o conjunto de dados que podem ajudar a cumprir os objetivos do projeto.
- **Preparação dos dados**, uma regra comum é que 80% do projeto é preparação de dados, esta tarefa consiste em selecionar, limpar, construir, integrar e formatar os dados.
- **Modelação**, o que é considerado o trabalho mais empolgante de *data science* também costuma ser a fase mais curta do projeto, nesta fase consiste em construir e avaliar vários modelos com base em várias técnicas de modelagem diferentes.
- **Avaliação dos resultados**, esta fase analisa de forma mais ampla qual modelo que apresenta os melhores resultados e são definidos os próximos passos.
- **Implementação**, dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados.

Capítulo 3

Exploração e modelação

3.1 Incidentes em Braga - 2019

O conjunto de dados fornecidos pelos docentes da UC vesam para os incidentes rodoviários na cidade de Braga no ano de 2019, onde consta com as seguintes colunas e o respectivo tratamento de dados na respetiva coluna:

- city_name
  - Retirado devido a ser uma variável com pouca importância,o seu valor é sempre a respetiva cidade de Braga.
- magnitude_of_delay
  - devido a baixa incidência de casos *"Minor"* e *"UNKNOWN_DELAY"* estas foram mapeadas para um caso moderado
- delay_in_seconds
- affected_roads
  - Foi transformada em uma Lista e de seguida foi executado uma contagem do numero total de ruas afectadas.
- record_date
  - foi mapeada para um campo de data e hora e de seguida apenas foi extraído o numero do mês, o dia do ano , o numero do dia da semana e a Hora; de seguida o campo é retirado por já ter sido extraído toda a informação necessária
- luminosity
- avg_temperature
- avg_humidity
- avg_wind_speed
- avg_atm_pressure, avg_precipitation, avg_rain
  - Retidados por falta de importância para o modelo
- accidents

Modelação

3.2 Banco de Portugal - Campanhas de Marketing

O conjunto de dados escolhido pelo autor do trabalho incide em resultados de campanhas de marketing do **Banco de Portugal**, onde estas são resultantes principalmente através de chamadas telefónicas, oferecendo ao cliente um termo de deposito. A variável alvo categórica de decisão consta com 2 hipóteses: "yes","no

. Este conjunto de dados foi recolhido através da plataforma online *Kaggle*, uma comunidade focada na resolução de problema de *data science* e de *Machine Learning*.[2]

Este conjunto consta com as seguinte colunas e os seus respetivos pré-processamentos:

- age
- day_of_week
- duration
- campaign
- emp.var.rate: employment variation rate
- cons.price.idx: consumer price index
- cons.conf.idx: consumer confidence index
- euribor3m: euribor 3 month rate
- nr.employed
- education, marital, job
  - tratamento dos valores "Unknown"e a aplicação do *One Hot Encoding*
- y, loan, housing, contact, month, default, previous, pdays, poutcome
  - transformação dos valores nominais para valores numéricos

Modelação

Capítulo 4

Workflows Desenvolvidos

4.1 Incidentes Braga - 2019

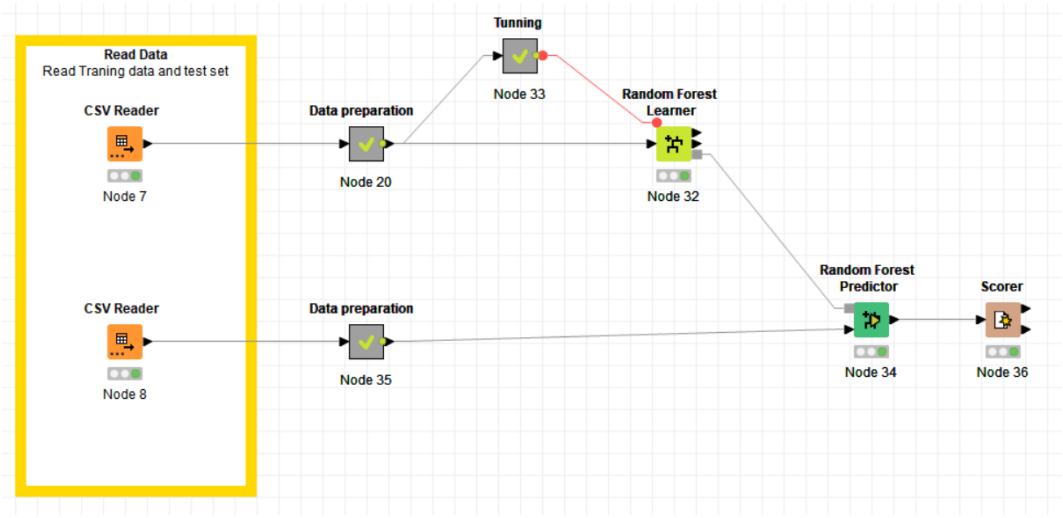

Figura 4.1: Visão geral do workflow gerado

- Read Data:
  leitura do dataset de treino e do dataset de teste nos *NODE 7 e 8*

- Data preparation
  Os *NODE 20 e 35* consistem no mesmo processamento de dados porque a estrutura dos dois conjuntos é a mesma

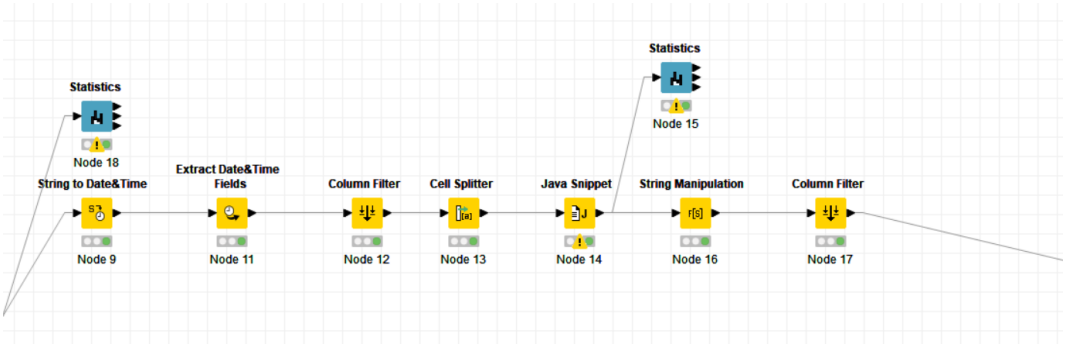

Figura 4.2: Pre-processamento

- Tunning

  De forma a obter o melhor modelo possível é executado um tunning para recorrer a loops em que estes testam varias combinações possíveis entre os critérios de divisão de Random Forests e múltiplos valores para o número de modelos e *TreeDepth*. Desta forma é obtido melhor modelo possível perante as opções de modificações

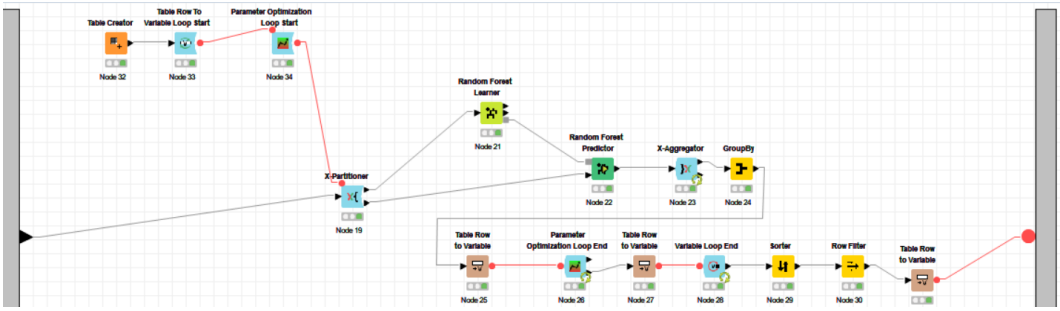

Figura 4.3: Tunning

- Previsões

  Após as optimizações é feito o treino utilizando uma Random Forest Learner e o respectivo teste do conjunto de teste através do Random Forest Predictor.

Através do *Scorrer* é obtido a matriz de confusão que devolve como as combinações entre Falso e Positivos relativamente ao conjunto de teste, podemos observar na figura 4.4.

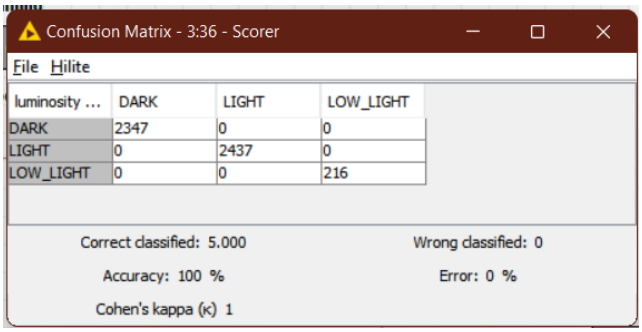

Figura 4.4: Matriz de confusão conjunto de teste

4.1.1 Descrição do Tunning

Após o loop de optimização foram obtidos os seguintes valores (figura 4.5):

maxIterations	10
currentIteration	9
Loop-Execute	
Loop (0)	
iteration	24
depth	30
NumMaxModels	800
Loop-Execute	
Loop (1)	
Creteria	Gini
RowID	Row2
currentIteration	2
maxIterations	3
Loop-Execute	
Loop (2)	
knime.workspace	D:\GitHub\SBS

Figura 4.5: Matriz de confusão conjunto de teste

4.2 Banco de Portugal - Campanhas de Marketing

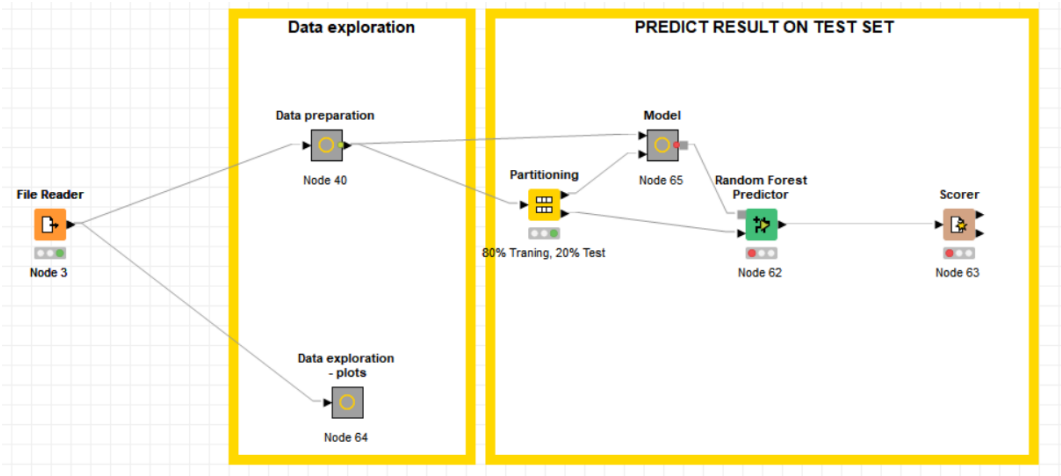

Figura 4.6: Visão geral do workflow gerado

- Read Data:
  leitura do dataset de treino e do dataset de teste nos *NODE 7 e 8*

- Data preparation
  Os *NODE 20 e 35* consistem no mesmo processamento de dados porque a estrutura dos dois conjuntos é a mesma

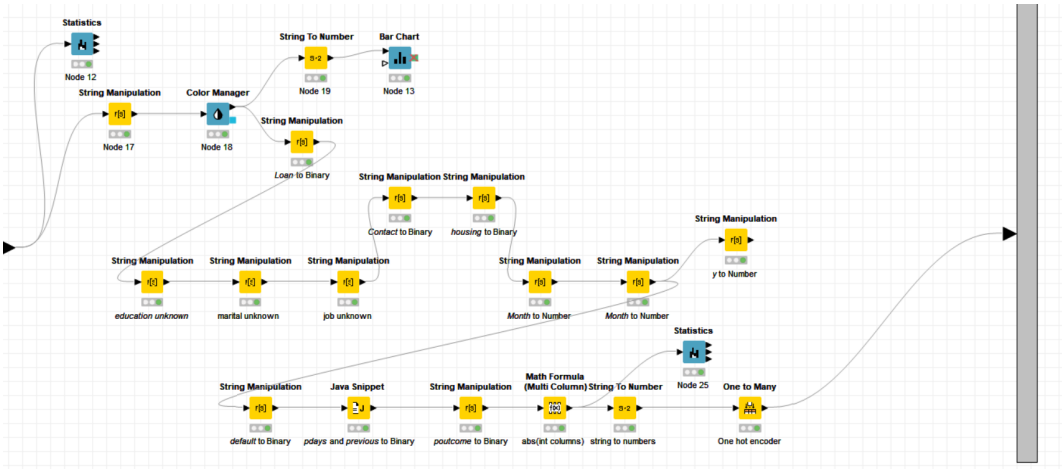

Figura 4.7: Pre-processamento

- Tunning

  De forma a obter o melhor modelo possível é executado um tunning para recorrer a loops em que estes testam varias combinações possíveis entre os critérios de divisão de Random Forests e múltiplos valores para o número de modelos e *TreeDepth*. Desta forma é obtido melhor modelo possível perante as opções de modificações

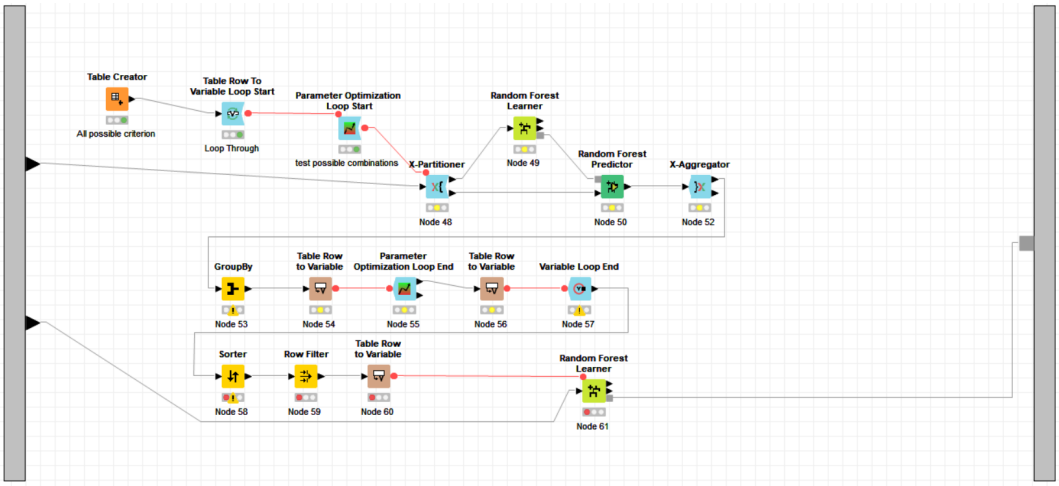

Figura 4.8: Tunning

- Previsões

  Após as optimizações é feito o treino utilizando uma Random Forest Learner e o respectivo teste do conjunto de teste através do Random Forest Predictor.

4.2.1 Descrição do Tunning

Após o loop de optimização foram obtidos os seguintes valores (figura 4.9):

Name	Value
maxIterations	10
currentIteration	3
Loop-Execute	
Loop (0)	
iteration	45
numModels	500
treeDepth	10
stopingCriteria	2
Loop-Execute	
Loop (1)	
splitCriterion	InformationGain
RowID	Row0
currentIteration	0
maxIterations	3
Loop-Execute	
Loop (2)	
knime.workspace	D:\GitHub\SBS

Figura 4.9: Configurações utilizados para o treino do modelo

Capítulo 5

Conclusões

Umas formas mais intuitivas e mais úteis é a utilização da metodologia CRISP-DM, ao longo deste projeto foi um ponto fulcral devido a correlação do aumento da complexidade do trabalho e o aumento do conhecimento sobre a UC. Este foi essencial não só para garantir a qualidade dos modelos como também para estruturar e organizar o desenvolvimento do projeto.

O CRISP-DM foi essencial quer a a exploração do dataset relativo aos incidentes em Braga em 2019, quer na exploração do dataset relativo Banco de Portugal.

Relativamente modelos criados, de uma forma geral o resultado é satisfeitos com os resultados obtidos. Foi possivel uma precisão alta com o modelo gerado para prever a magnitude de incidentes rodoviários na cidade de Braga.

Relativamente ao modelo desenhado para prever o Banco de Portugal, há uma necessidade de um reforço no modelo para poder obter resultados em concreto, alguma falta de conhecimento na utilização da ferramente KNIME tornou a solução difícil de conseguir

Sem duvida há muito ponto que poderiam ser trabalhados/modificados de forma a conseguir resultados ainda melhores e ter um trabalho ainda mais completo

Bibliografia

[1] Data Science Project Management
https://www.datascience-pm.com/crisp-dm-2/

[2] Dataset Banco de Portugal
https://www.kaggle.com/volodymyrgavrysh/bank-marketing-campaigns-dataset