

Universidade do Minho

Departamento de Informática

Mestrado em Engenharia Informática

Perfil de Machine Learning

Fundamentos e Aplicações Sistemas Baseados em Similaridade

1º Ano, 1º Semestre

Ano letivo 2020/2021

Enunciado Prático nº 2

26 de outubro de 2020

Joel Costa Carvalho

PG42837

Índice

[Tarefas - 4 -](#_Toc54615567)

[T1 – Carregar, no *Knime*, ambos os *datasets*. Utilizar um nodo *Joiner* para agregar, por “*area* *code*” e “*phone*”, os dados provenientes das duas *readers*. Transformar o atributo *Churn* em nominal; - 4 -](#_Toc54615568)

[T2 – Particionar os dados de forma estratificada (pela *feature* “*Churn*”), utilizando 75% para aprendizagem e 25% para teste. Aplicar um *Decision Tree Learner* e um *Decision Tree Predictor.* Avaliar a precisão (*accuracy*) do modelo utilizando o nodo *Scorer* e a respetiva matriz de confusão; - 4 -](#_Toc54615569)

[T3 – Remover, iterativamente, *features* do *dataset* e reavaliar a performance dos modelos candidatos. Descrever os resultados obtidos; - 5 -](#_Toc54615570)

[T4 – Seguir as práticas de bons-hábitos na construção de *workflows*; - 5 -](#_Toc54615571)

[T5 – Utilizar o *output* de um nodo *Decision Tree Learner* para criar uma imagem de uma Árvore de Decisão e guardar essa imagem no ambiente de trabalho; - 6 -](#_Toc54615572)

Índice de figuras

[Figura 1 – T1 - 4 -](#_Toc54615358)

[Figura 2 - T2 - 4 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP2/TP2_JoelCarvalho_PG42837.docx#_Toc54615359)

[Figura 3 - Matriz confusão - 4 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP2/TP2_JoelCarvalho_PG42837.docx#_Toc54615360)

[Figura 4 - Avaliação da precisão - 4 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP2/TP2_JoelCarvalho_PG42837.docx#_Toc54615361)

[Figura 5 - T4 - 5 -](#_Toc54615362)

[Figura 6 - T5 - 6 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP2/TP2_JoelCarvalho_PG42837.docx#_Toc54615363)

**Tarefas**

**T1 – Carregar, no *Knime*, ambos os *datasets*. Utilizar um nodo *Joiner* para agregar, por “*area* *code*” e “*phone*”, os dados provenientes das duas *readers*. Transformar o atributo *Churn* em nominal;**

Visível na Figura 1., esta tarefa foi subdividida em duas, uma primeira parte para a leitura dos *datasets* em *CSV* e *Excel*, utilizando os nodos *CSV Reader* e *Excel Reader* respetivamente. Para cruzar os dados de ambos, foi utilizado o nodo *Joiner*, comum noutros paradigmas de programação (SQL). Como conclusão, o atributo *Churn* foi transformado em *string,* isto porque o tipo nominal apenas labora com *strings*, e consequentemente em nominal.

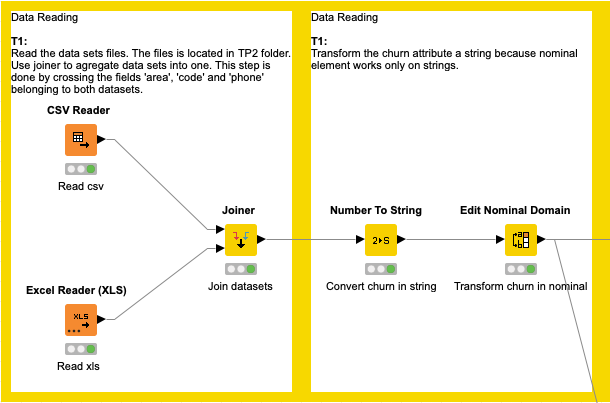


Figura 1 – T1

**T2 – Particionar os dados de forma estratificada (pela *feature* “*Churn*”), utilizando 75% para aprendizagem e 25% para teste. Aplicar um *Decision Tree Learner* e um *Decision Tree Predictor.* Avaliar a precisão (*accuracy*) do modelo utilizando o nodo *Scorer* e a respetiva matriz de confusão;**

Aplicação de uma partição de 75% vs 25% e treinar o modelo para obter a precisão e a respetiva matriz confusão.

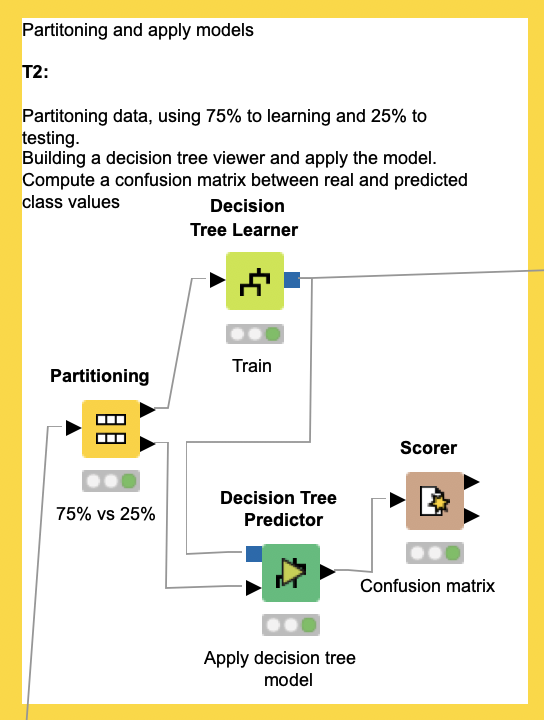


Figura 2 - T2



Figura 3 - Matriz confusão

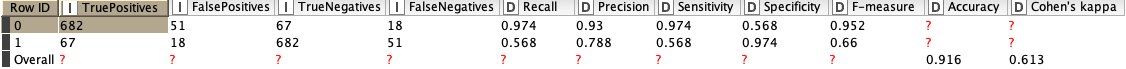


Figura 4 - Avaliação da precisão

**T3 – Remover, iterativamente, *features* do *dataset* e reavaliar a performance dos modelos candidatos. Descrever os resultados obtidos;**

Analisando a Tabela 1. podemos concluir que a remoção da *feature* *“State”* e *“Accont Length”* são onde obtemos os melhores resultados. Relativamente à precisão, ambos são similares com uma diferença de 0.1%, contudo podemos concluir que a *feature* *“State”* é um ligeiramente melhor devido ao número elevado de resultados bem classificados. Agregando a eliminação destes elementos, alcançamos os resultados mais precisos com 93.2%, apenas com 57 resultados mal classificados.

Concluindo, apenas uma análise pessoal, se for excluída uma *feature* relacionada com os minutos diários de comunicação dos clientes, a precisão padece de uma queda e regista uma percentagem abaixo dos 90%.

|  |  |  |  |
| --- | --- | --- | --- |
| *Feature* removida | *Accuracy (~)* | Bem classificados | Mal classificados |
| - | 91.6% | 749 | 69 |
| Phone | 91.6% | 727 | 67 |
| State | **92.6%** | **772** | **62** |
| Area Code | 90.5% | 733 | 77 |
| Account Length | **92.7%** | **746** | **59** |
| Account Length & State | **93.2%** | **777** | **57** |
| Day Mins | 89.8% | 737 | 83 |

Tabela 1 - Resultados mediante remoção de features

**T4 – Seguir as práticas de bons-hábitos na construção de *workflows*;**

Descrição de cada tarefa, atribuição de nomes ao nodos e delimitação colorida com a aplicação de anotações.



Figura 5 - T4

**T5 – Utilizar o *output* de um nodo *Decision Tree Learner* para criar uma imagem de uma Árvore de Decisão e guardar essa imagem no ambiente de trabalho;**

Veja-se a Figura 6., para guardar a imagem gerada pelo nodo *Decision Tree Learner* foi necessário realizar a respetiva transformação e salvaguardá-la na raiz do projeto como “imageOutput.png”.

Para a execução desta tarefa é necessário definir as cores para cada valor do “*Churn”* (0 ou 1).

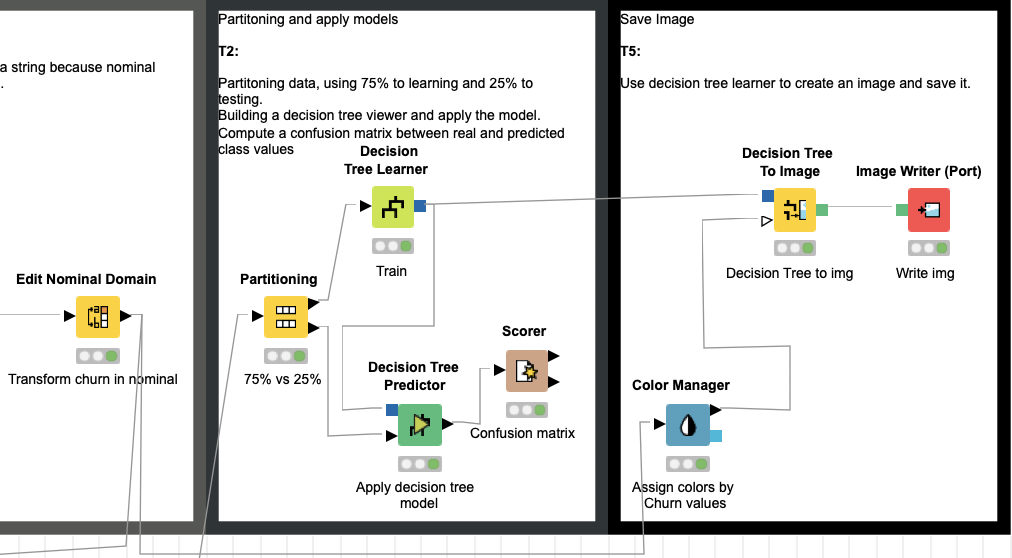
****

Figura 6 - T5