

UNIVERSIDADE DO MINHO

LICENCIATURA EM ENGENHARIA INFORMÁTICA

Aprendizagem e Decisão Inteligentes

Conceção de modelos de aprendizagem

Grupo 11

Duarte Parente (A95844)

Gonçalo Pereira (A96849)

José Moreira (A95522)

Santiago Domingues (A96886)

Ano Letivo 2022/2023

Índice

1	Introdução	6
2	Tarefa A	7
2.1	Definição do Problema	7
2.2	Metodologia Adotada	7
2.3	Descrição do <i>Dataset</i>	8
2.4	Tratamento dos Dados	8
2.4.1	<i>Missing Values</i>	8
2.4.2	Inconsistência dos Dados	9
2.4.3	Adaptação dos Dados	9
2.5	Matriz de Correlação	10
2.6	Exploração dos Dados	10
2.7	Modelos Desenvolvidos	13
2.7.1	Decision Tree	13
2.7.2	Random Forest	14
3	Tarefa B	15
3.1	Definição do Problema	15
3.2	Metodologia Adotada	15
3.3	Descrição do <i>Dataset</i>	15
3.4	Tratamento dos Dados	16
3.4.1	Erros de Escrita	16
3.4.2	Conversão de string para valor numérico	16

3.4.3	Conversão de valor numérico para string	17
3.4.4	Arredondamentos	17
3.5	Matriz de Correlação	18
3.6	Exploração dos Dados	18
3.7	Modelos Desenvolvidos	21
3.7.1	<i>Dataset</i> Completo	22
3.7.2	<i>Dataset</i> do Departamento de <i>Sewing</i>	24
3.7.3	<i>Dataset</i> do Departamento de <i>Finishing</i>	26
3.7.4	Reflexão sobre os Modelos	28
4	Conclusão	29

Lista de Figuras

1	Tratamento dos Dados	9
2	Matriz de Correlação	10
3	Exploração Univariada e Bivariada dos Dados	11
4	Relação entre a média dos valores das transferências e a idade dos jogadores, por liga	11
5	Relação entre a média dos valores das transferências e o ano das mesmas, por liga	12
6	Relação entre a média dos valores das transferências e o ano das mesmas, por mercado	12
7	Aplicação do algoritmo <i>Decision Tree</i>	13
8	Matriz de confusão com os resultados obtidos	13
9	Aplicação do algoritmo <i>Random Forest</i>	14
10	Matriz de confusão com os resultados obtidos	14
11	Nodos de Tratamento dos Dados	17
12	Matriz de Correlação	18
13	Análise Univariada	19
14	Análise Bivariada	19
15	Número de ocorrências de cada equipa	20
16	Média de produtividade esperada e produtividade observada, por equipa	20
17	Número de ocorrências de cada departamento	21
18	Média de produtividade esperada e produtividade observada, por departamento . .	21
19	Aplicação do algoritmo <i>Decision Tree</i> ao <i>dataset</i> completo	22
20	Matriz de Confusão do algoritmo <i>Decision Tree</i> aplicado ao <i>dataset</i> completo . . .	23
21	Aplicação do algoritmo <i>Linear Regression</i> ao <i>dataset</i> completo	23
22	<i>Numeric Scorer</i> do algoritmo <i>Linear Regression</i> aplicado ao <i>dataset</i> completo . . .	23

23	Aplicação do algoritmo RProp MLP ao <i>dataset</i> completo	24
24	Matriz de Confusão do algoritmo RProp MLP aplicado ao <i>dataset</i> completo	24
25	Aplicação do algoritmo Decision Tree ao <i>dataset</i> do departamento de <i>sewing</i> . . .	25
26	Matriz de Confusão do algoritmo Decision Tree aplicado ao <i>dataset</i> do departamento de <i>sewing</i>	25
27	Aplicação do algoritmo Linear Regression ao <i>dataset</i> do departamento de <i>sewing</i> .	25
28	<i>Numeric Scorer</i> do algoritmo Linear Regression aplicado ao <i>dataset</i> do departamento de <i>sewing</i>	26
29	Aplicação do algoritmo Decision Tree ao <i>dataset</i> do departamento de <i>finishing</i> . .	26
30	Matriz de Confusão do algoritmo Decision Tree aplicado ao <i>dataset</i> do departamento de <i>finishing</i>	27
31	Aplicação do algoritmo Linear Regression ao <i>dataset</i> do departamento de <i>finishing</i>	27
32	<i>Numeric Scorer</i> do algoritmo Linear Regression aplicado ao <i>dataset</i> do departamento de <i>finishing</i>	28

Capítulo 1

Introdução

O presente relatório visa apresentar o projeto desenvolvido no âmbito da Unidade Curricular de Aprendizagem e Decisão Inteligentes. Partindo de dois *datasets*, sendo o primeiro escolhido pelo grupo e o segundo atribuído pelos docentes, o principal objetivo do projeto passa pela conceção de modelos de *Machine Learning*, modelos estes criados recorrendo à plataforma KNIME e de forma a conseguir-se reforçar a aprendizagem dos vários temas envolventes, já lecionados. Como se sabe, o conceito de *Machine Learning* refere-se à capacidade da máquina, obviamente, recorrendo a algoritmos específicos e a conjuntos de dados, de criar modelos de representação de conhecimento e, em simultâneo, aperfeiçoá-los de forma iterativa, permitindo, portanto, que a mesma ganhe capacidade de "aprender". Realizou-se todo o projeto em simultaneidade com o estudo da teoria que envolve *Machine Learning*, abordando-se, portanto, tópicos importantes como: *Inteligência Artificial*, *aprendizagem*, *decisão*, etc.

Com o objetivo de se conseguir responder aos parâmetros de avaliação presentes no enunciado do trabalho prático, procedeu-se à escolha/atribuição dos *datasets*. O *dataset* escolhido pelo grupo, para a primeira tarefa do projeto, é relativo às transferências no futebol nas principais ligas. Para a segunda tarefa, o *dataset* atribuído pelos docentes diz respeito à produção de vestuário. Ambos os *datasets* irão ser abordados com mais detalhe numa fase posterior deste relatório.

Neste relatório serão apresentados os resultados obtidos em cada uma das tarefas, bem como as metodologias utilizadas para a construção dos modelos de *Machine Learning*. Serão discutidos também os desafios enfrentados durante o processo dos modelos desenvolvidos, sendo por fim apresentadas as conclusões.

Capítulo 2

Tarefa A

2.1 Definição do Problema

No contexto do futebol profissional, uma transferência traduz a movimentação de um jogador de um clube para o outro, envolvendo a aquisição total ou parcial do passe do mesmo por parte do clube comprador. Regra geral, esta compra apenas pode ser realizada no decorrer dos períodos de transferência definidos pelos campeonatos profissionais. O valor de transferência acordado entre clubes está relacionado em alguns aspetos como: momento de forma do jogador, futuro potencial e margem de progressão, tendo em consideração não só a vertente desportiva, mas também uma eventual desvalorização do valor teórico do ativo adquirido.

O objetivo deste projeto é desenvolver um modelo de *machine learning* capaz de classificar com precisão diferentes intervalos de valores de transferência nas 7 principais ligas europeias, desde 1993 até à janela de transferências do verão da época desportiva atual, tratando-se, portanto, de um problema de classificação. O desenvolvimento de um modelo desta natureza poderá trazer inúmeros benefícios, nomeadamente ao nível da tomada de decisão financeira, permitindo fazer investimentos mais informados.

2.2 Metodologia Adotada

De forma a solucionar o problema anteriormente definido, foi adotada a seguinte estratégia:

1. Procura de um *dataset* adequado às necessidades do problema, contendo dados detalhados sobre os jogadores envolvidos na transferência.
2. Leitura e processamento do *dataset*, assim como a análise e exploração dos dados, garantindo um conjunto de dados adequado à resolução do problema.
3. Desenvolvimento de modelos de *machine learning* num paradigma de aprendizagem supervisionada devido à natureza do problema de classificação em questão.
4. Análise dos resultados obtidos e respetivo desempenho dos modelos e aperfeiçoamento dos mesmos, nomeadamente através do ajuste de diferentes parâmetros na tentativa de redução do erro de classificação, assim como o teste de diferentes algoritmos.

2.3 Descrição do *Dataset*

De forma a constituir objeto de estudo, surgiu a necessidade de procura de um *dataset* que fosse ao encontro do problema em questão. Deste modo, e como proposto pelo enunciado do projeto, visitou-se a secção de *datasets* do *Kaggle* e estudou-se os vários ficheiros presentes no mesmo. Em consequência, escolheu-se o *dataset* **Football Transfers (Major Leagues)**, de onde foram usados todos os ficheiros **.csv**, à exceção dos English Championship (*championship.csv*) e Russian Premier Liga (*premier-liga.csv*), uma vez que se pretendia aplicar o estudo apenas às principais sete ligas europeias de futebol. Assim, dispõe-se de um *dataset* de 137085 linhas, sendo cada uma delas constituída pelos seguintes atributos:

- **club_name** - clube que realizou a transferência;
- **player_name** - nome do jogador envolvido na transferência;
- **age** - idade do jogador;
- **position** - posição em campo do jogador;
- **club_involved_name** - clube com o qual a transferência foi realizada;
- **fee** - valor;
- **transfer_movement** - fluxo da transferência tendo em conta o clube que a realizou;
- **transfer_period** - mercado de transferências onde a mesma se realizou;
- **fee_cleaned** - transformação do atributo "fee" para um valor numérico;
- **league_name** - nome da liga onde o clube que realizou a transferência se encontra inscrito;
- **year** - ano da transferência;
- **season** - época da transferência.

2.4 Tratamento dos Dados

Uma das etapas essenciais para a realização de uma análise a um conjunto de dados prende-se exatamente na sua preparação e pré-processamento, no sentido de tratar dados incompletos ou até mesmo inválidos.

Sendo assim, mostrou-se necessário tratar o *dataset* escolhido de forma a detetar essas anomalias, para além de o adaptar em função do problema definido anteriormente. Segue-se uma enumeração e descrição dos passos tomados, assim como a respetiva aplicação na plataforma KNIME.

2.4.1 *Missing Values*

Analisando o conjunto de dados em mão, rapidamente se percebeu a existência de alguns valores em falta, em particular com a existência de '?' e '-' no atributo **fee**, o que por sua vez origina um 'NA' no atributo **fee_cleaned**. Para além disso, foram também detetados 'NA' no atributo 'age'. Procedeu-se então à remoção destes dados através do nodo **Row Filter**.

2.4.2 Inconsistência dos Dados

Para além da deteção de valores em falta, foi também detetada uma inconsistência no atributo `transfer_period`, onde o período de transferências relativo ao Verão encontrava-se escrito de duas formas distintas, `'summer'` e `'Summer'`. De forma a uniformizar este atributo, utilizou-se o nodo **Rule Engine** que procurou substituir todos os valores pela versão com a letra maiúscula.

2.4.3 Adaptação dos Dados

Tratando-se de um *dataset* de transferências entre as principais ligas europeias, contendo registos tanto de entradas como saídas de jogadores, o mesmo mostrou possuir inúmeras duplicações de registos nos casos em que as transferências ocorrem entre ligas contidas no *dataset*. Sendo assim, decretou-se a remoção de todas as transações de saída de jogadores, ficando apenas o registo das contratações, isto é, todas as transferências relativas à entrada de jogadores. De notar que se usou o nodo **Row Filter** para realizar esta operação. Para além disso, retirou-se o atributo `fee`, dado que a informação relevante nele contida está transcrita na `fee_cleaned`.

Outra decisão tomada em relação ao processamento dos dados, tendo em vista a definição do problema de classificação de intervalos de valores de transferência, prendeu-se na remoção de todos os registos associados a empréstimos de jogadores, através do nodo **Row Filter**.

Finalmente, surgiu a necessidade de fazer a transformação dos valores de transferência de forma a que os mesmos perdessem a propriedade de continuidade, fazendo assim o ajuste final ao *dataset*. Para esse efeito, deu-se uso ao nodo **Numeric Binner**, que agregou os diferentes valores do atributo `fee_cleaned` em 8 intervalos diferentes: $[0, 2[$, $[2, 5[$, $[5, 10[$, $[10, 20[$, $[20, 40[$, $[40, 60[$, $[60, 100[$, $[100, +\infty[$.

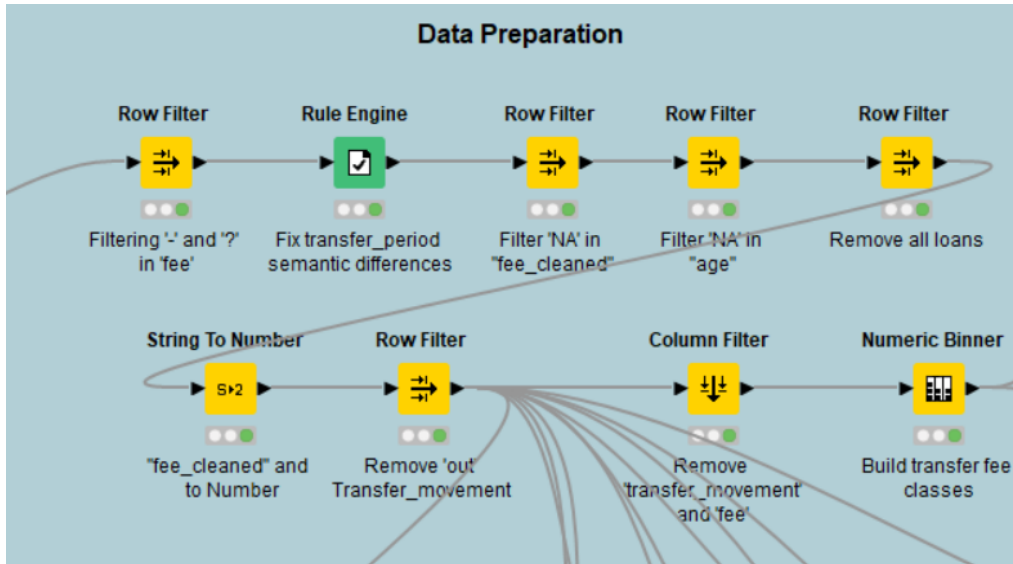


Figura 1: Tratamento dos Dados

Terminada a etapa de tratamento dos dados, o *dataset* passou, então, a conter 23742 registos e 10 colunas.

2.5 Matriz de Correlação

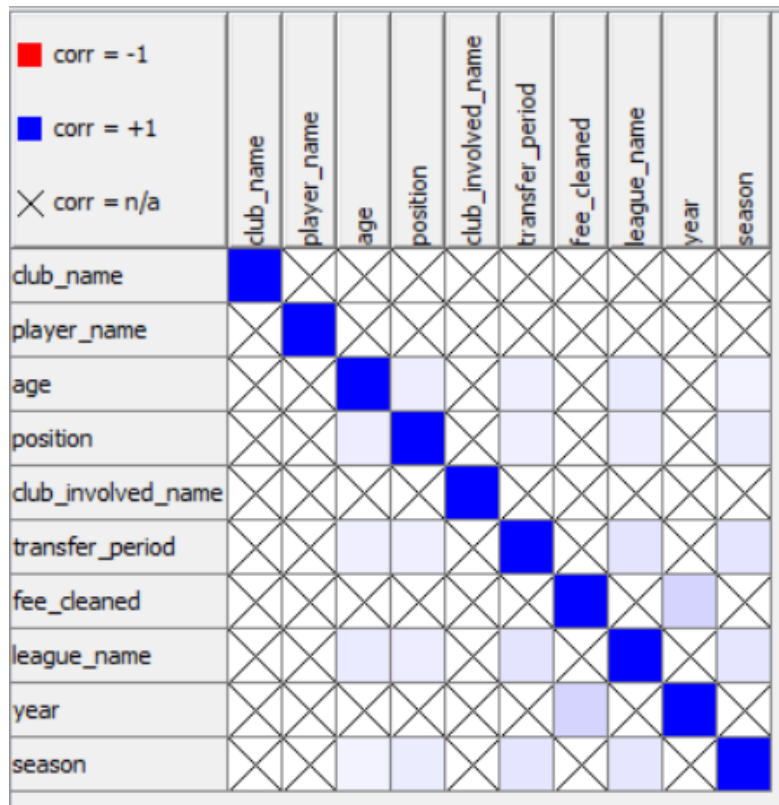


Figura 2: Matriz de Correlação

Analisando a matriz de correlação obtida através da aplicação do nodo **Linear Correlation**, conseguiu-se estudar a correlação linear dos atributos contidos no *dataset*.

A maior evidência a retirar da figura acima trata-se da clara falta de correlação dos atributos `club_name`, `player_name`, `club_involved_name`.

2.6 Exploração dos Dados

Nesta etapa, procedeu-se à exploração dos dados representados no *dataset* em estudo. Como é possível observar na imagem abaixo afixada, fez-se esta mesma exploração de duas formas distintas: análise univariada (que tem por base uma variável) e análise bivariada (que tem por base a influência/relação entre variáveis). Na primeira, usou-se o nodo **Statistics**, que, ligado a três nodos **Table View**, permite a visualização, estudo e análise do comportamento de cada um dos atributos do *dataset*. Por outro lado, recorrendo à análise bivariada, consegue-se perceber a distribuição dos dados com base em dois ou mais atributos, estudando-se, portanto, a relação entre os mesmos. Executou-se esta mesma análise recorrendo a nodos **Bar Chart**, **Histogram**, **Pivoting** e **Line Plot**.

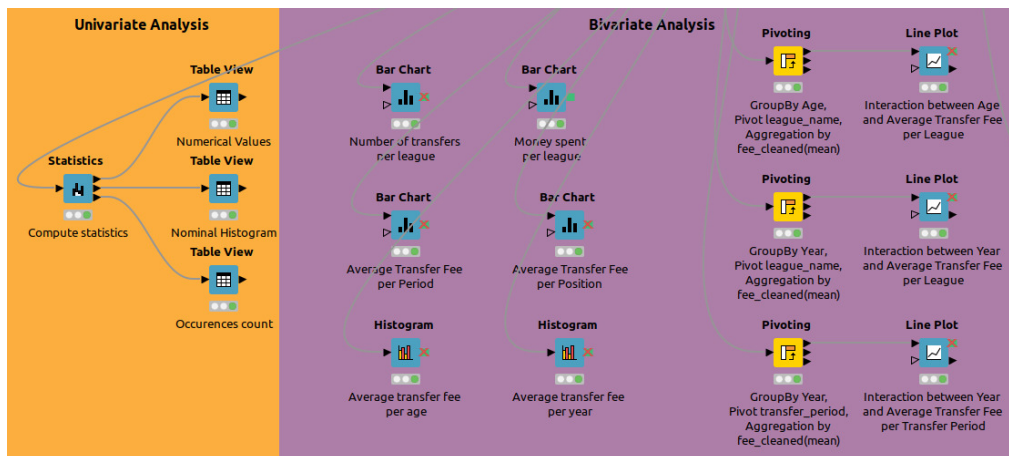


Figura 3: Exploração Univariada e Bivariada dos Dados

De forma a poder dar a entender a relação entre os atributos mais importantes do *dataset* em estudo, apresenta-se, de seguida, os resultados visuais dos nodos **Line Plot** acima mencionados.

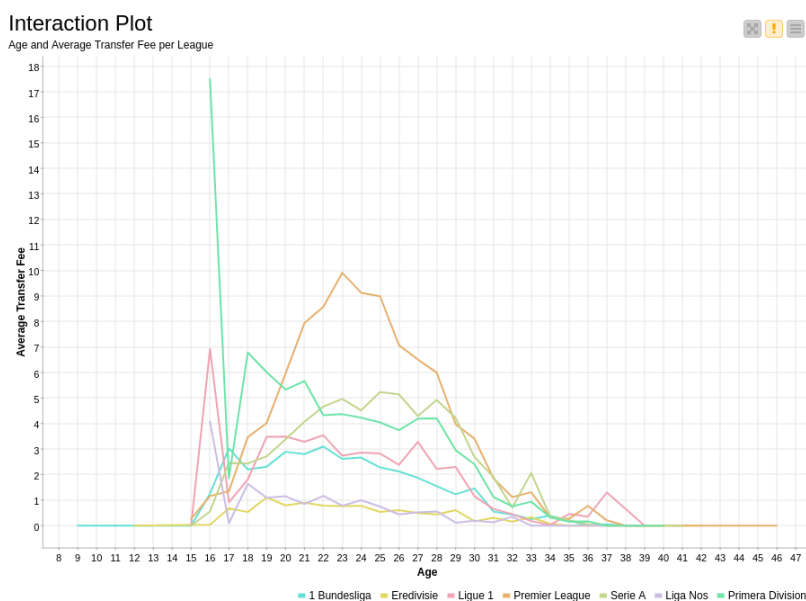


Figura 4: Relação entre a média dos valores das transferências e a idade dos jogadores, por liga

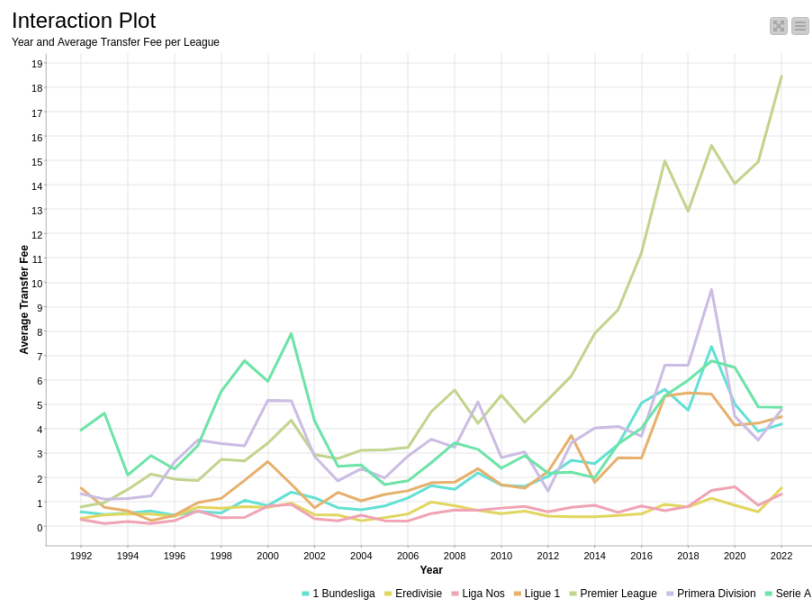


Figura 5: Relação entre a média dos valores das transferências e o ano das mesmas, por liga

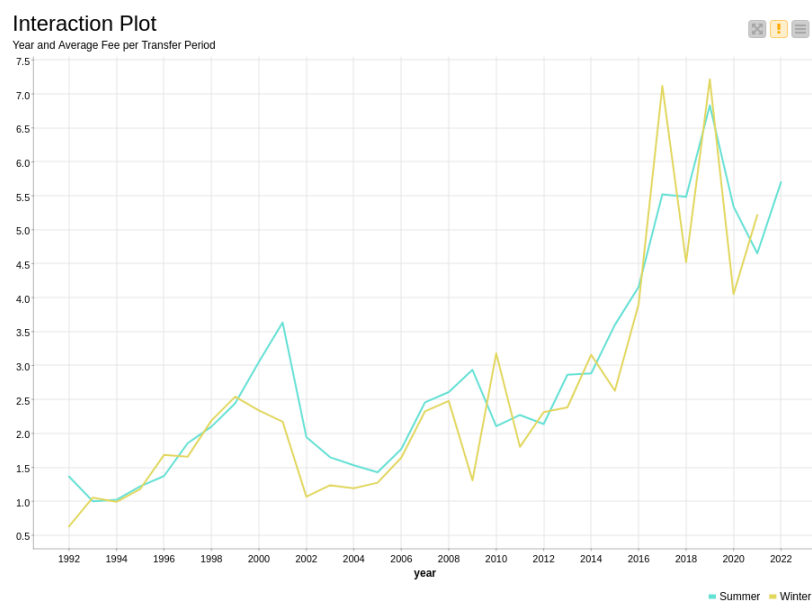


Figura 6: Relação entre a média dos valores das transferências e o ano das mesmas, por mercado

2.7 Modelos Desenvolvidos

Terminada a fase de leitura e processamento do *dataset* e respetiva análise exploratória dos dados, reuniu-se, portanto, as condições necessárias para o início do desenvolvimento dos modelos de *machine learning*. Tratando-se de um problema de classificação, optou-se pelo desenvolvimento de alguns dos algoritmos mais comuns neste tipo de problemas, como as *Decision Tree* e *Random Forest*.

De notar que houve também lugar à partição do *dataset*, através do nodo **Partitioning**, que possui a capacidade de o dividir em dois conjuntos, teste e treino. Escolheu-se os valores de separação **80%** e **20%** para o teste das previsões.

Todos os resultados apresentados neste capítulo correspondem à primeira interação com o respetivo algoritmo. Dessa forma, as definições usadas nos nodos correspondem à *default* do KNIME. Para além disso, não houve qualquer alteração efetuada, relativa ao *dataset*, no sentido de uma procura por um melhor desempenho do modelo.

2.7.1 Decision Tree

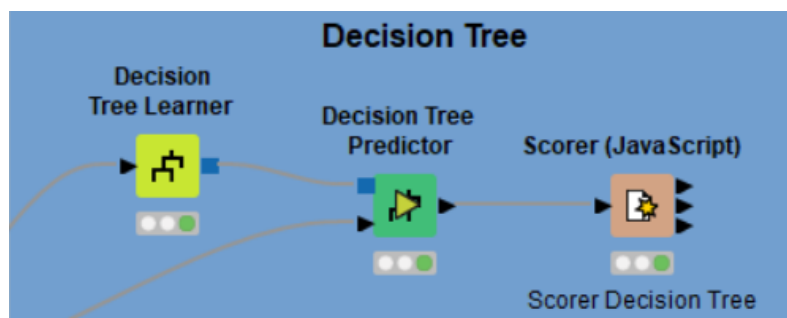


Figura 7: Aplicação do algoritmo *Decision Tree*

Decision Tree Scorer

Confusion Matrix

	[0,2[(Pr...	[10,20[(...	[100,-[(...	[2,5[(Pr...	[20,40[(...	[40,60[(...	[5,10[(P...	[60,100[...	
[0,2[(Ac...	3130	18	0	67	2	0	26	0	96.52%
[10,20[(...	189	24	0	20	7	0	11	0	9.56%
[100,-[(...	2	0	0	0	0	0	0	0	0.00%
[2,5[(Ac...	658	18	0	32	2	0	18	0	4.40%
[20,40[(...	72	18	0	9	3	0	6	0	2.78%
[40,60[(...	8	3	0	2	2	0	0	0	0.00%
[5,10[(A...	326	18	0	31	1	0	16	0	4.08%
[60,100[...	4	5	0	0	1	0	0	0	0.00%
	71.31%	23.08%	undefined	19.88%	16.67%	undefined	20.78%	undefined	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
67.49%	32.51%	0.100	3205	1544

Figura 8: Matriz de confusão com os resultados obtidos

2.7.2 Random Forest

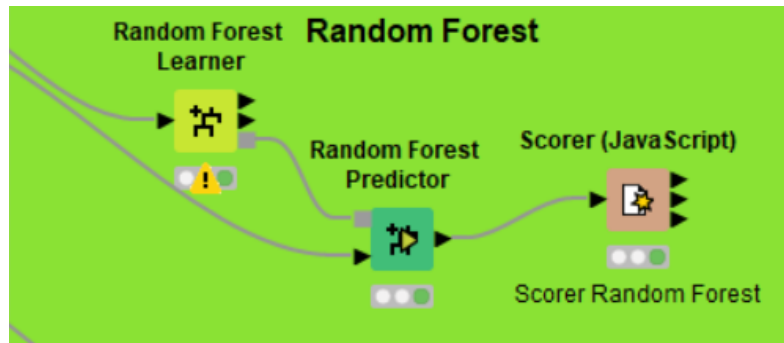


Figura 9: Aplicação do algoritmo *Random Forest*

Scorer View

Confusion Matrix

	[0,2[(Pr...	[10,20[(...	[100,-[(...	[2,5[(Pr...	[20,40[(...	[40,60[(...	[5,10[(P...	[60,100[...	
[0,2[(Ac...	3139	20	0	36	1	0	16	0	97.73%
[10,20[(...	183	28	0	17	1	0	11	0	11.67%
[100,-[(...	2	1	0	0	0	0	0	0	0.00%
[2,5[(Ac...	665	14	0	23	1	0	11	0	3.22%
[20,40[(...	97	30	0	2	0	0	5	0	0.00%
[40,60[(...	14	8	0	0	0	0	0	0	0.00%
[5,10[(A...	367	18	0	17	1	0	7	0	1.71%
[60,100[...	7	5	0	0	2	0	0	0	0.00%
	70.16%	22.58%	undefined	24.21%	0.00%	undefined	14.00%	undefined	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's Kappa (κ)	Correctly Classified	Incorrectly Classified
67.32%	32.68%	0.086	3197	1552

Figura 10: Matriz de confusão com os resultados obtidos

Capítulo 3

Tarefa B

3.1 Definição do Problema

A produção de vestuário é das indústrias mais influentes no mundo. Esta abrange desde a produção de fibras e tecidos até à manufatura de roupas, calçados e acessórios. No entanto, cada passo na produção de uma peça de vestuário envolve vários recursos, humanos e não humanos, o que, aliado à enorme escala desta indústria a nível mundial, faz com que seja crucial a constante análise da eficiência da mesma.

Nesta tarefa, pretendeu-se que se desenvolvesse um modelo de *machine learning* com o objetivo de prever a produção de vestuário, de forma a analisar a eficiência dos trabalhadores. Este estudo, uma vez que visa o estudo da produtividade dos funcionários, permite, assim, perceber quais os parâmetros e variáveis com maior peso e decisão naquilo que é a eficiência dos mesmos.

3.2 Metodologia Adotada

A abordagem utilizada para a resolução do problema foi a seguinte:

1. Explorar, analisar e preparar o *dataset*, procurando extrair conhecimento relevante no contexto do problema em questão.
2. Conceber modelos de *Machine Learning* de classificação e de regressão, abordando o problema em contexto de redes neuronais.
3. Realizar uma análise crítica de resultados e respetivo desempenho dos modelos. Tal como na tarefa anterior, esta análise permite o aperfeiçoamento dos modelos em questão.

3.3 Descrição do *Dataset*

Como referido anteriormente, este *dataset* foi atribuído pelos docentes e diz respeito à produção de vestuário. Trata-se de um ficheiro *.csv* composto por 1196 linhas, no qual cada linha contém os seguintes atributos:

- **rowID** - id da linha
- **date** - data, no formato dd/mm/aa hh:mm
- **department** - departamento responsável pela tarefa em questão
- **team** - número da equipa responsável pela tarefa em questão
- **targeted_productivity** - produtividade alvo definida para cada equipa, para cada dia
- **smv** - *Standard Minute Value*: tempo alocado para uma tarefa
- **wip** - *Work in progress*: número de itens inacabados para produtos
- **over_time** - quantidade de horas extras de cada equipa, em minutos
- **incentive** - quantidade de incentivo financeiro que permite ou motiva um determinado curso de ação
- **idle_time** - quantidade de tempo em que a produção foi interrompida (por diversas razões)
- **idle_men** - número de trabalhadores que ficaram inativos devido à interrupção da produção
- **no_of_style_change** - número de mudanças no estilo de um determinado produto
- **no_of_workers** - número de trabalhadores em cada equipa
- **actual_productivity** - percentagem real de produtividade entregue pelos trabalhadores

3.4 Tratamento dos Dados

Uma vez mais, à semelhança do primeiro *dataset*, surgiu a necessidade de fazer o tratamento dos dados, de forma a detetar/remover as anomalias encontradas. Segue-se a enumeração e descrição dos passos tomados, assim como a respetiva aplicação na plataforma KNIME.

3.4.1 Erros de Escrita

Numa fase inicial, detetou-se erros gramaticais no atributo **department**, onde apareciam palavras como "finishnig" em vez de "finishing", ou "swenig" no lugar de "sewing". De forma a colmatar estes erros, utilizou-se o nodo **Rule Engine**, que atuou no sentido de fazer a substituição destas palavras pelas corretas.

3.4.2 Conversão de string para valor numérico

Posteriormente, mostrou-se necessário/aconselhável, converter alguns dos parâmetros que se encontravam representados por string para o seu respetivo valor numérico. Com isto, refere-se os campos **no_of_workers**, **targeted_productivity**, **actual_productivity**, **smv** e **idle_time**. Toda esta conversão foi realizada com recurso ao nodo **String to Number**.

3.4.3 Conversão de valor numérico para string

Por outro lado, surgiu a necessidade de fazer o oposto, ou seja, converter dados numéricos para dados em representação do tipo string. Esta conversão deve-se ao facto desses mesmos valores numéricos não possuírem uma propriedade útil enquanto número. A mesma realizou-se no parâmetro **team**, através do nodo **Number to String**.

3.4.4 Arredondamentos

Aquando a fase de tratamento de dados, reparou-se na inconsistência entre o tipo de dados no parâmetro **no_of_workers** e o seu verdadeiro valor. Como explicado anteriormente, este campo refere-se ao número de trabalhadores em cada equipa e, portanto, mostra-se impossível em representação decimal. Uma vez que o dataset apresentava essas inconsistências, resolveu-se transformar esses mesmos valores no seu valor teto. O nodo **Round Double** encarregou-se de realizar esta transformação.

Por outro lado, decidiu-se reduzir os valores de **actual_productivity** para duas casas decimais, de forma a haver uma comparação mais correta entre os mesmos e os valores de **targeted_productivity** e de forma a haver um estudo mais certo e compacto por parte dos modelos de *machine learning* em desenvolvimento. Este mesmo processo realizou-se recorrendo a **Rule Engine**.

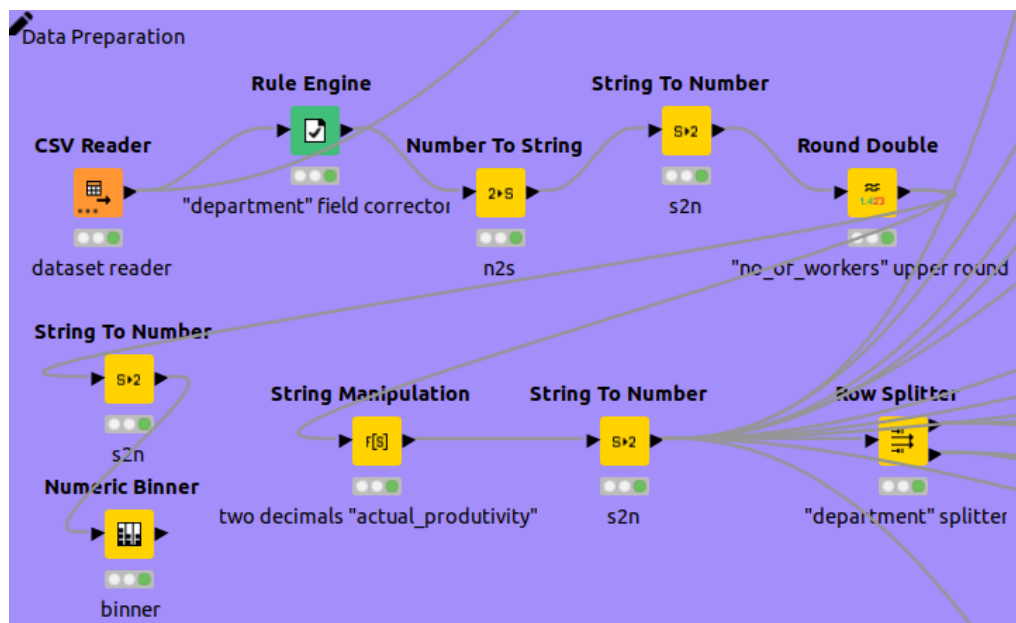


Figura 11: Nodos de Tratamento dos Dados

3.5 Matriz de Correlação

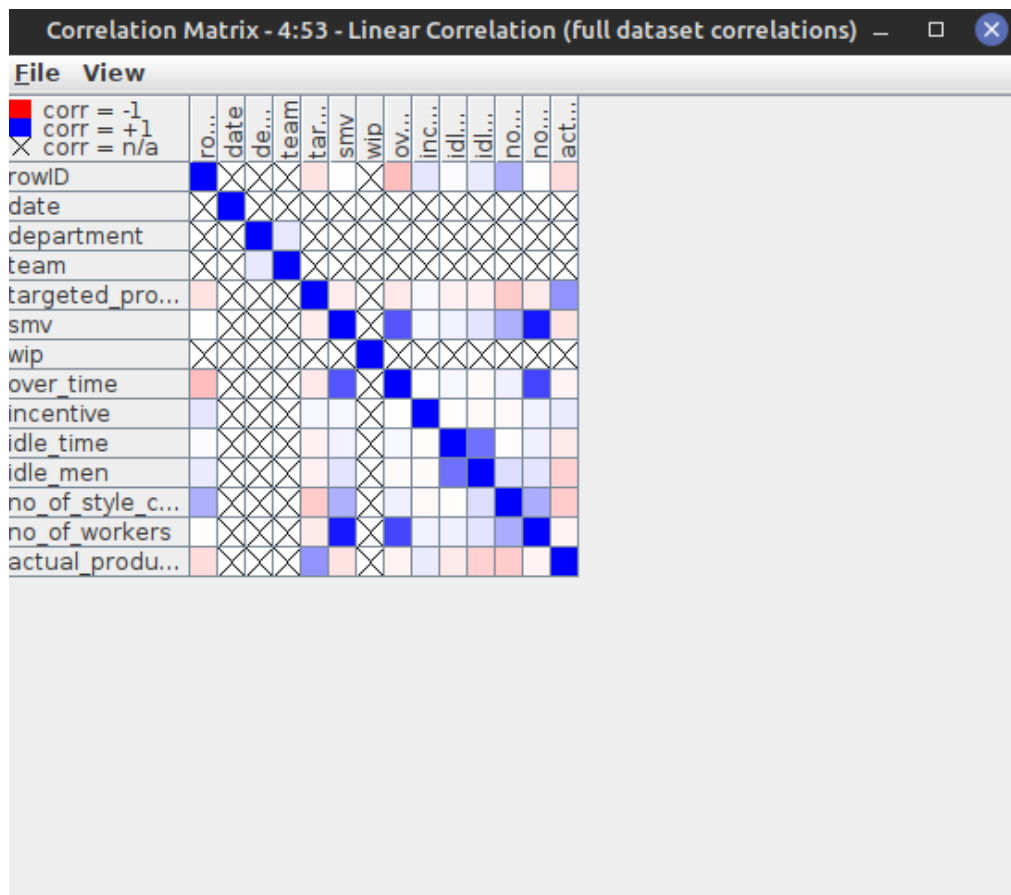


Figura 12: Matriz de Correlação

Por análise da matriz de correlação apresentada, conseguiu-se estudar e perceber, com maior eficiência, todas as correlações entre os diversos atributos. Possibilitou-se toda esta análise por aplicação do nodo **Linear Correlation**.

Como observável pela imagem, percebe-se, facilmente, a correlação positiva entre a variável **targeted_productivity** e a variável independente em estudo, **actual_productivity**.

3.6 Exploração dos Dados

Com a etapa de exploração do *dataset*, permitiu-se uma percepção melhor dos parâmetros em estudo. Deste modo, mostrou-se possível efetuar um melhor tratamento dos dados e consequente estudo dos mesmos. Com isto, realça-se o facto de se ter realizado a fase de exploração antes, durante e após a fase de tratamento dos dados. De forma a ser possível um estudo mais próximo do conteúdo em questão, procedeu-se à aplicação de nodos de exploração, o que permitiu uma investigação mais completa, suportada pela vertente visual das características dos dados. À semelhança do problema anterior, realizou-se esta mesma análise com base em duas vertentes:

análise univariada e análise bivariada (termos esses já explicados, anteriormente). A primeira das duas efetuou-se através dos nodos **Bar Chart**, **Data Explorer**, **Statistics** e **Table View**, enquanto que a segunda se realizou recorrendo aos nodos **Crosstab (local)**, **Linear Correlation**, **GroupBy** e **Bar Chart**.

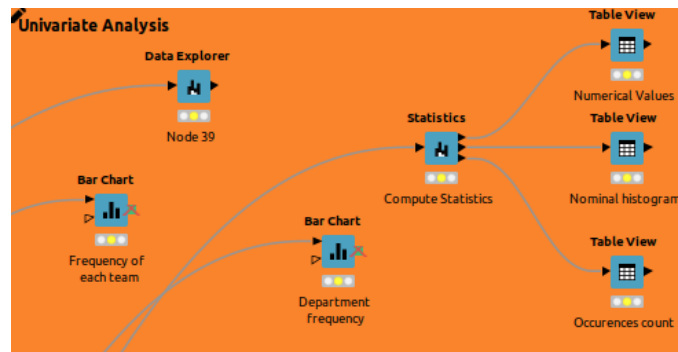


Figura 13: Análise Univariada

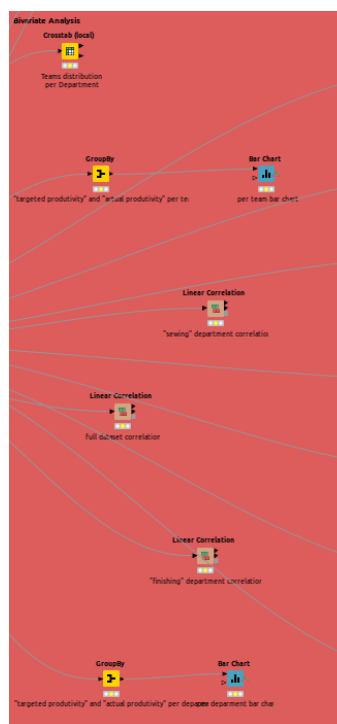


Figura 14: Análise Bivariada

Através das exploração e análise de dados anteriormente referidas e explicadas, conseguiu-se produzir resultados visuais que se revelaram úteis na construção e desenvolvimento do resto do processo de *machine learning*, quer nos modelos, quer no tratamento dos dados e suas relações. De seguida, apresentam-se alguns dos gráficos mais expressivos relativos a ambos os tipos de análise.

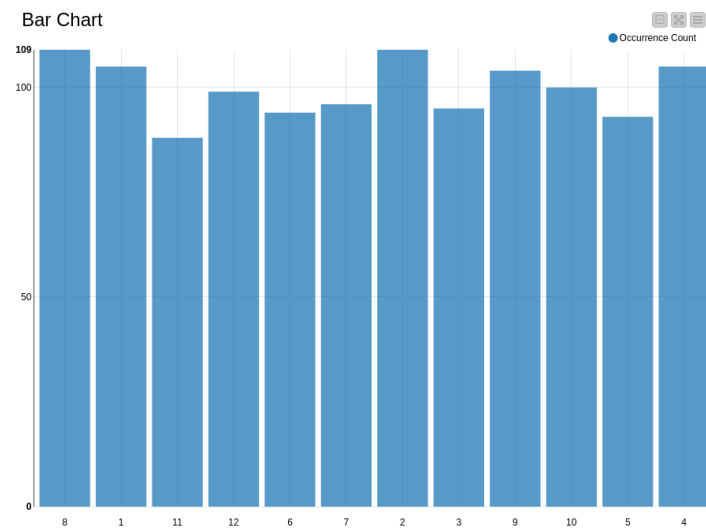


Figura 15: Número de ocorrências de cada equipa

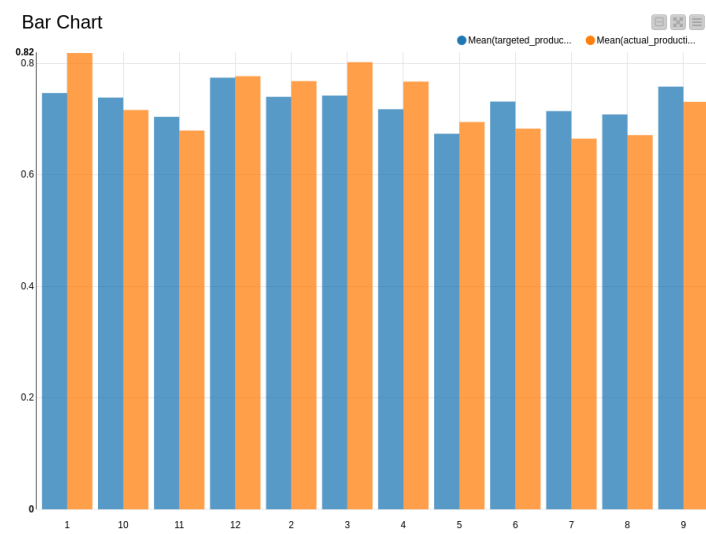


Figura 16: Média de produtividade esperada e produtividade observada, por equipa

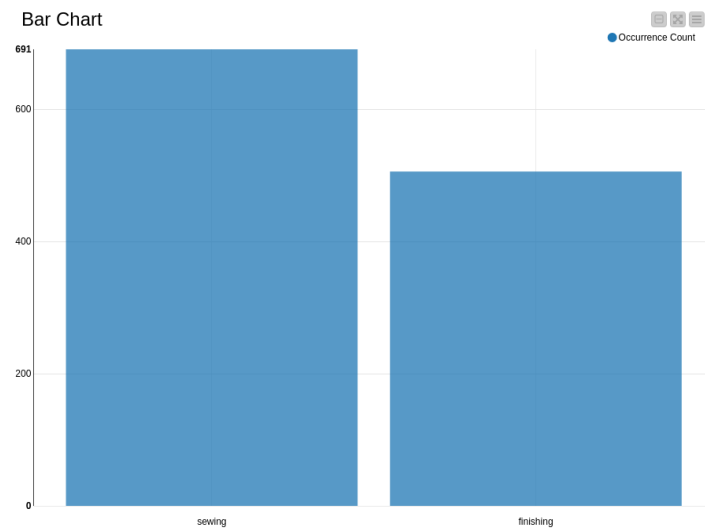


Figura 17: Número de ocorrências de cada departamento

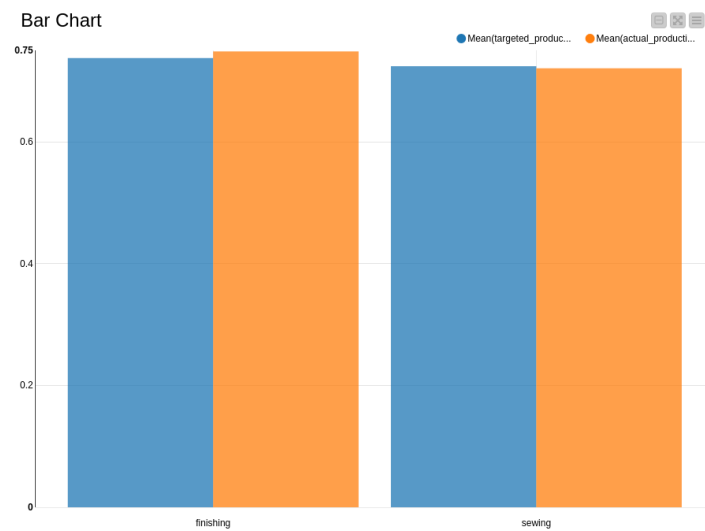


Figura 18: Média de produtividade esperada e produtividade observada, por departamento

3.7 Modelos Desenvolvidos

Após as fases de tratamento e exploração dos dados, procedeu-se ao desenvolvimento e afinação dos modelos de *machine learning*. Encarou-se o problema, inicialmente, como um problema de regressão e aplicou-se a teoria necessária para avançar com o seu estudo. Porém, decidiu-se, conjuntamente, retratar o mesmo também como um problema de classificação, em que a variável independente (em estudo) representa uma classe (uma vez que, referindo-se a uma percentagem em formato decimal, encontra-se definida numa escala de 0 a 1). Por outro lado, aplicou-se os nodos necessários para que o estudo fosse abordado por redes neuronais, o que, consequentemente,

fez com que o mesmo se tornasse mais completo. Porém, verificou-se que os resultados, tanto no modelo de regressão como no modelo de classificação e no modelo de redes neurais mostraram-se muito insatisfatórios e optou-se por estudar o problema dividido em duas partes idênticas, separando o *dataset* por ambos os departamentos, ao mesmo tempo que se descartou o uso de redes neurais para estudo do problema. Com isto, passa-se a explicar a abordagem do problema com base no objeto concreto de estudo: *dataset completo*, *dataset* referente ao departamento de "sewing" e *dataset* referente ao departamento de "finishing".

3.7.1 *Dataset* Completo

O estudo referente ao *dataset* completo encontra-se já explicado na parte introdutória da atual secção. Este mesmo estudo revelou-se insatisfatório e originou os dois estudos a explicar numa fase posterior do relatório. Por outro lado, apresenta-se, de seguida, os algoritmos usados e os seus respetivos resultados, alcançados através da exclusão dos parâmetros **rowID** e **date**, recorrendo ao nodo **Column Filter**.

Decision Tree

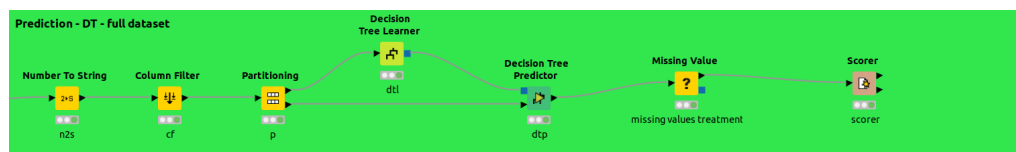


Figura 19: Aplicação do algoritmo Decision Tree ao *dataset* completo

File	Hilite					
actual_pr...	0.8	0.75	0.94	0.93	0.85	
0.8	32	0	0	0	1	
0.75	5	17	0	0	0	
0.94	0	0	0	0	0	
0.93	2	0	0	0	0	
0.85	0	0	0	0	6	
0.68	0	0	0	0	0	
0.95	0	0	0	0	0	
0.88	0	0	0	0	1	
0.98	0	0	0	0	0	
0.67	1	0	0	0	0	
0.9	2	0	1	0	0	
0.87	0	0	0	0	0	
0.72	0	0	0	0	0	
0.65	0	1	0	0	0	
0.5	0	0	0	0	0	
0.7	1	1	0	0	0	
0.35	1	0	0	0	0	
0.66	1	0	0	0	1	
0.6	0	0	0	0	0	
0.97	0	0	0	0	1	
0.82	0	0	0	0	0	
0.96	0	0	0	0	0	
0.64	0	0	0	0	0	
0.89	0	0	0	1	0	
0.86	1	0	0	0	0	

Correct classified: 89 Wrong classified: 151
 Accuracy: 37.083% Error: 62.917%
 Cohen's kappa (κ): 0.328%

Figura 20: Matriz de Confusão do algoritmo Decision Tree aplicado ao *dataset* completo

Linear Regression

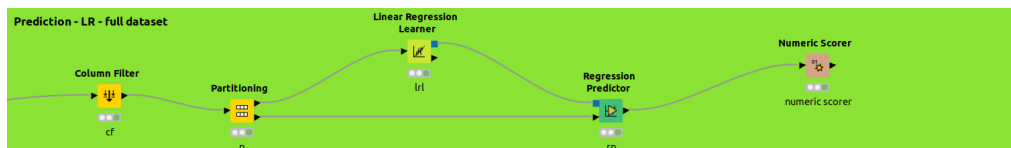


Figura 21: Aplicação do algoritmo Linear Regression ao *dataset* completo

File	
R ² :	0.319
Mean absolute error:	0.115
Mean squared error:	0.025
Root mean squared error:	0.157
Mean signed difference:	0.007
Mean absolute percentage error:	0.214
Adjusted R ² :	0.319

Figura 22: *Numeric Scorer* do algoritmo Linear Regression aplicado ao *dataset* completo

RProp MLP

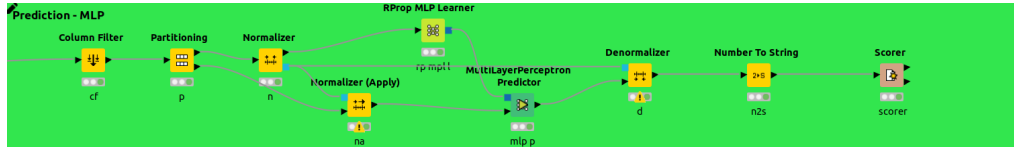


Figura 23: Aplicação do algoritmo RProp MLP ao *dataset* completo

File	Hilite					
actual_pr...		0.8	0.72	0.7	0.59	0.899999...
0.8		0	0	0	0	0
0.72		0	0	0	0	0
0.7		0	0	0	0	0
0.59		0	0	0	0	0
0.899999...		0	0	0	0	0
0.690000...		0	0	0	0	0
0.62		0	0	0	0	0
0.93		0	0	0	0	0
0.909999...		0	0	0	0	0
0.85		0	0	0	0	0
0.79		0	0	0	0	0
0.749999...		0	0	0	0	0
0.66		0	0	0	0	0
0.6		0	0	0	0	0
0.959999...		0	0	0	0	0
0.87		0	0	0	0	0
0.809999...		0	0	0	0	0
0.41		0	0	0	0	0
0.33		0	0	0	0	0
0.89		0	0	0	0	0
0.83		0	0	0	0	0
0.77		0	0	0	0	0
0.23		0	0	0	0	0
0.98		0	0	0	0	0
0.67		0	0	0	0	0

Correct classified: 0 Wrong classified: 360
 Accuracy: 0% Error: 100%
 Cohen's kappa (κ): 0%

Figura 24: Matriz de Confusão do algoritmo RProp MLP aplicado ao *dataset* completo

3.7.2 *Dataset* do Departamento de *Sewing*

Também se realizou o estudo referente às linhas do *dataset* em que o valor do parâmetro **department** é igual a *sewing*. Deste modo, aplicou-se os algoritmos respetivos a modelos de regressão e classificação, à semelhança do que foi aplicado ao *dataset* original, onde se recorreu à filtragem das mesmas colunas (**rowID**, **date**).

Decision Tree

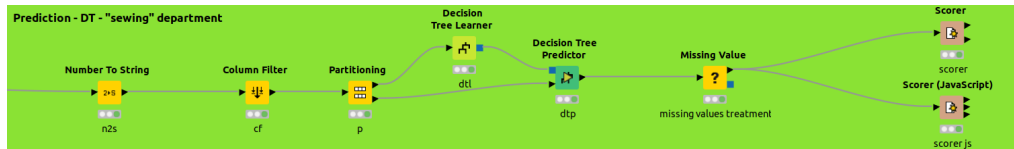


Figura 25: Aplicação do algoritmo Decision Tree ao *dataset* do departamento de *sewing*

File	Hilite					
actual_pr...		0.8	0.75	0.7	0.85	0.69
0.8		38	1	0	0	0
0.75		1	11	2	0	0
0.7		1	0	20	0	1
0.85		2	0	0	7	0
0.69		1	0	0	0	0
0.6		0	0	0	0	0
0.47		1	0	0	0	0
0.35		0	0	0	0	0
0.65		0	0	0	0	0
1.0		0	0	0	1	0
0.95		0	0	0	0	0
0.54		0	0	0	0	0
0.41		0	0	0	0	0
0.5		0	0	0	0	0
0.9		0	0	0	0	0
0.79		0	0	0	0	0
0.55		0	0	0	0	0
0.72		0	0	0	0	1
0.52		0	1	0	0	0
0.87		1	0	0	0	0
0.38		0	0	0	0	0
0.45		0	0	0	0	0
0.64		0	0	0	0	0
0.66		1	0	0	0	0
0.84		0	0	0	0	0

Correct classified: 91 Wrong classified: 48
 Accuracy: 65.468% Error: 34.532%
 Cohen's kappa (κ): 0.594%

Figura 26: Matriz de Confusão do algoritmo Decision Tree aplicado ao *dataset* do departamento de *sewing*

Linear Regression

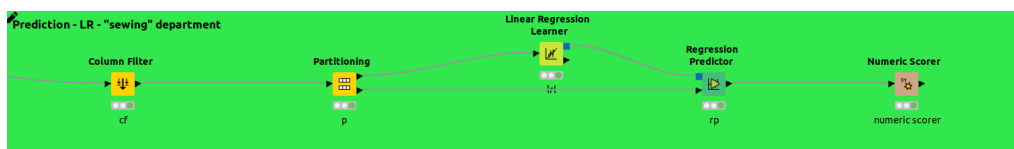


Figura 27: Aplicação do algoritmo Linear Regression ao *dataset* do departamento de *sewing*

Statistics - 3:123 - Numeric Sc... — □ ×	
File	
R ² :	0.85
Mean absolute error:	0.036
Mean squared error:	0.003
Root mean squared error:	0.055
Mean signed difference:	0.002
Mean absolute percentage error:	0.059
Adjusted R ² :	0.85

Figura 28: *Numeric Scorer* do algoritmo Linear Regression aplicado ao *dataset* do departamento de *sewing*

3.7.3 *Dataset* do Departamento de *Finishing*

Por outro lado, alargou-se o modelo de *machine learning* à fração correspondente às linhas do *dataset* em que o valor do parâmetro **department** é igual a *finishing*. De forma similar, usou-se os mesmos algoritmos que os usados para estudar os dados com o outro valor de *department*. Porém, houve a necessidade de filtragem de mais um parâmetro, ou seja, excluiu-se as colunas respetivas a **rowID**, **date** e **wip**, uma vez que a última possuía o mesmo valor (**NaN**) para todos os registos, não oferecendo assim conteúdo útil de estudo ao modelo.

Decision Tree

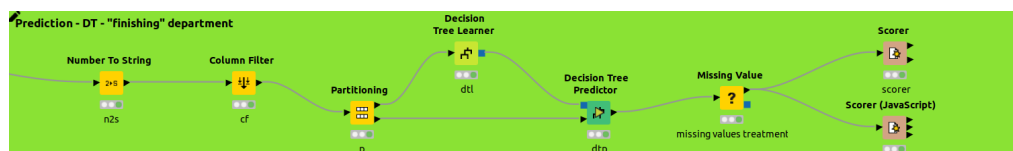


Figura 29: Aplicação do algoritmo Decision Tree ao *dataset* do departamento de *finishing*

File Hilite		0.7	0.62	0.91	0.8	0.89
actual_pr...						
0.7	1	0	0	0	0	0
0.62	1	0	0	0	0	0
0.91	0	0	0	0	0	0
0.8	0	0	0	0	1	1
0.89	0	0	0	0	0	1
0.68	0	0	0	0	0	1
0.82	0	0	0	0	0	0
0.95	0	0	0	0	0	0
0.94	0	0	0	0	0	0
0.88	0	0	0	0	0	1
0.85	0	0	0	0	0	0
0.93	0	0	0	0	0	0
0.9	0	0	1	0	1	1
0.86	0	0	0	0	0	1
0.98	0	0	0	0	0	0
0.59	0	0	1	0	0	0
0.74	0	0	0	0	0	0
0.55	0	0	0	0	0	0
0.97	0	0	0	0	0	1
0.87	0	0	1	0	0	0
0.54	0	0	0	0	0	0
0.71	0	0	0	0	0	0
0.77	0	1	0	0	0	0
0.81	0	0	0	0	0	0
0.72	0	0	0	0	0	1

Correct classified: 11 Wrong classified: 91
 Accuracy: 10.784% Error: 89.216%
 Cohen's kappa (κ): 0.088%

Figura 30: Matriz de Confusão do algoritmo Decision Tree aplicado ao *dataset* do departamento de *finishing*

Linear Regression

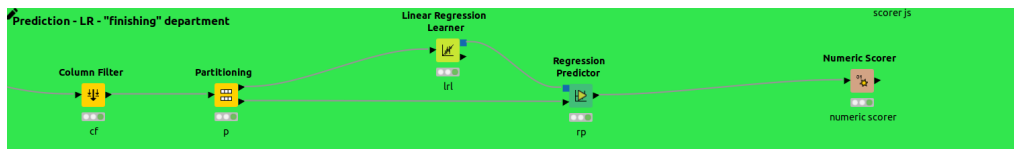
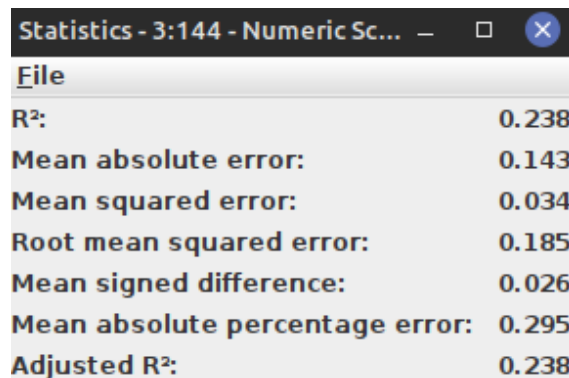


Figura 31: Aplicação do algoritmo Linear Regression ao *dataset* do departamento de *finishing*



File	
R ² :	0.238
Mean absolute error:	0.143
Mean squared error:	0.034
Root mean squared error:	0.185
Mean signed difference:	0.026
Mean absolute percentage error:	0.295
Adjusted R ² :	0.238

Figura 32: *Numeric Scorer* do algoritmo Linear Regression aplicado ao *dataset* do departamento de *finishing*

3.7.4 Reflexão sobre os Modelos

Como observável nas últimas frações do atual relatório, o estudo dividido proporcionou uma aprendizagem rica no sentido em que os resultados obtidos se mostram muito distantes. Porém, reitera-se o facto de todos os modelos terem sido preparados com as configurações *default* do **KNIME** e não ter havido, praticamente, manipulação das diversas variáveis. Por outro lado, confirma-se a importância da conversão das mesmas e das decisões de filtragem de certos parâmetros, dependendo do modelo em execução.

Capítulo 4

Conclusão

Dá-se assim por concluído o projeto desenvolvido no âmbito da Unidade Curricular de Aprendizagem e Decisão Inteligentes. Analisando o projeto, considera-se que o grupo correspondeu com satisfação aos objetivos pretendidos.

A realização do projeto permitiu ao grupo adquirir mais experiência na temática de *machine learning* e, conseqüentemente, alargar o grau de conhecimento em volta do assunto. Por outro lado, o estudo do tema em questão levou a que o grupo aumentasse o tempo de contacto com a ferramenta **KNIME**. Tudo isto resultou numa oportunidade de exploração das diversas funcionalidades da mesma, levando a uma utilização mais completa e a uma maior escala.

Relativamente à primeira tarefa, o primeiro grande obstáculo surgiu na escolha de um *dataset* adequado ao problema em estudo. Porém, ao ultrapassar esta fase, o grupo conseguiu avançar no desenvolvimento do projeto com relativa rapidez e um grau elevado de empenho. Em relação à segunda tarefa, houve claras dificuldades naquilo que foi o processo de manipulação e tratamento do *dataset*, uma vez que o mesmo é composto por uma grande quantidade de variáveis. Como é de prever, esta dificuldade resultou também num maior grau de complexidade em lidar com o desenvolvimento dos modelos, como por exemplo, na manutenção e configuração dos nodos. Sob outra perspetiva, ambos os problemas associados ao projeto desafiaram o grupo em relação à capacidade de conseguir desenvolver habilidades importantes relacionadas com a tomada de decisão.