

# Aprendizagem e Decisão Inteligentes

## Trabalho Prático

LEI - 2022/2023

Martim Ribeiro  
A96113



Afonso Bessa  
A95225



João Silva  
A91671



Luís Vilas  
A91697



Grupo 9



*Universidade do Minho*

---

## Resumo

O presente documento descreve sucintamente os objetos de avaliação e de análise ao longo de um projeto de análise inserido na Unidade Curricular **Aprendizagem e Decisão Inteligentes**. Este projeto teve como principais objetivos a análise, tratamento e previsão de dados de dois *datasets* distintos, usando a plataforma **KNIME**. Ao longo dos vários capítulos e secções presentes no relatório são expostas todas as decisões tomadas por parte da equipa de trabalho relativas aos métodos utilizados para o alcance do objetivo do projeto.

# Índice

1	Introdução .....	6
1	Características do <i>Dataset</i> .....	7
2	<i>KNIME Workflow</i> .....	8
3	Análise Preliminar dos Dados .....	9
3.1	Análise Geral .....	9
3.2	Análise da Correlação .....	11
3.3	Análise das <i>features</i> .....	14
4	Pré-Processamento .....	17
5	Análise dos Dados após o Pré-Processamento .....	18
6	Validação dos Modelos de Machine Learning .....	18
7	Modelos de Machine Learning .....	18
7.1	Decision Tree .....	19
7.2	Naive Bayes .....	20
7.3	Random Forest .....	21
7.4	Tree Ensemble .....	22
7.5	Gradient Boosted Tree .....	23
7.6	DL4J FeedForward .....	24
8	Análise de Resultados .....	24
1	Características do <i>Dataset</i> .....	25
2	<i>KNIME Workflow</i> .....	26
3	Análise Preliminar dos Dados .....	26
4	Pré-Processamento .....	28
5	Análise dos Dados após o Pré-Processamento .....	29
6	Validação dos Modelos de Machine Learning .....	31
7	Modelos de Machine Learning .....	32
7.1	<b>Decision Tree</b> .....	32
7.2	<b>Random Forest</b> .....	33
7.3	<b>Naive Bayes</b> .....	33
7.4	<b>Tree Ensemble</b> .....	33
7.5	<b>Gradient Boosted Trees</b> .....	34
7.6	<b>DL4J FeedForward</b> .....	34
7.7	<b>RProp MLP</b> .....	36
7.8	<b>Simple Regression Tree</b> .....	36
7.9	<b>Tree Ensemble (Regression)</b> .....	37
7.10	<b>Random Forest (Regression)</b> .....	37
7.11	<b>Gradient Boosted Trees (Regression)</b> .....	38
7.12	<b>Polynomial Regression</b> .....	38
7.13	<b>Linear Regression</b> .....	39
8	Resultados Obtidos .....	40
9	Conclusão .....	40

## **Lista de Figuras**

1	Exerto do Dataset .....	8
2	Workflow desenvolvido .....	8
3	Exploração geral dos dados numéricos .....	9
4	Exploração geral dos dados numéricos .....	9
5	Exploração geral dos dados nominais .....	10
6	Distribuição da satisfação pelo <i>dataset</i> .....	11
7	Correlação de <i>features</i> nominais .....	11
8	Correlação de <i>features</i> não nominais .....	12
9	Correlação de todas as <i>features</i> .....	12
10	Scatter plot entre a <i>feature</i> Arrival Delay in minutes e Departure Delay in minutes .....	13
11	Distribuição das <i>features</i> .....	14
12	Distribuição das <i>features</i> .....	15
13	Distribuição do tipo de viagem por tipo de cliente .....	16

---

14 Distribuição do nível de satisfação por tipo de cliente .....	16
15 Distribuição do tipo de classe por clientes satisfeitos ou não satisfeitos .....	16
16 Workflow de pré-processamento .....	17
17 Tratamento de missing values .....	17
18 Tratamento de missing values .....	17
19 Modelos de classificação .....	18
20 Decision Tree .....	19
21 Decision Tree .....	20
22 Random Forest .....	21
23 Tree Ensemble .....	22
24 Tree Ensemble .....	23
25 FeedForward .....	24
1 Excerto do Dataset .....	25
2 Workflow .....	26
3 Converter para <i>Missing Values</i> .....	26
4 Análise Estatística e Estudo de Correlação .....	27
5 <i>Data Explorer</i> - Numéricos .....	27
6 <i>Data Explorer</i> - Nominais .....	27
7 <i>Statistics - Department</i> .....	27
8 <i>Bar Chart - Work in Progress</i> .....	28
9 Correlação .....	28
10 Pré-Processamento .....	28
11 Pré-Processamento .....	29
12 Pré-Processamento .....	29
13 <i>Department</i> Antigo .....	29
14 <i>Department</i> Novo .....	29
15 Análise Estatística e Estudo de Correlação .....	30
16 <i>Data Explorer</i> - Numéricos .....	30
17 <i>Data Explorer</i> - Nominais .....	30
18 <i>Box Plot</i> .....	30
19 Correlação .....	31
20 Validação dos Modelos .....	32
21 Tipos de Problemas .....	32
22 Valores Obtidos .....	33
23 Valores Obtidos .....	33
24 Valores Obtidos .....	33
25 Valores Obtidos .....	33
26 Valores Obtidos .....	34
27 <i>DL4J FeedForward</i> .....	34
28 Valores Obtidos .....	36
29 Valores Obtidos .....	37
30 Valores Obtidos .....	37
31 Valores Obtidos .....	38
32 Valores Obtidos .....	38
33 Valores Obtidos .....	39
34 Valores Obtidos .....	39

## **Lista de Tabelas**

1 Variáveis Independentes .....	7
2 Variável Dependente .....	7
3 Resultados do Tree Leaner Hold-out validation .....	19
4 Resultados do Tree Leaner Cross Validation .....	19
5 Resultados do Naive Bayes Hold-out Validation .....	20
6 Resultados do Naive Bayes Cross Validation .....	20
7 Random Forest Hold Out Validation .....	21
8 Random Forest Cross Validation .....	21
9 Resultados do Tree Ensemble Hold-out Validation .....	22
10 Resultados do Tree Ensemble Cross Validation .....	22
11 Resultados do Tree Ensemble Hold-out Validation .....	23
12 Resultados do Tree Ensemble Cross Validation .....	23

---

13 Variáveis Independentes .....	25
14 Variável Dependente .....	25
15 Resultados dos experimentos com Tree Ensemble.....	34
16 Resultados dos experimentos com Naïve Bayes.....	35
17 Resultados dos experimentos com Decision Tree.....	35
18 Resultados dos experimentos com Random Forest. ....	35
19 Resultados dos experimentos com Gradient Boosted Trees. ....	36
20 Resultados dos experimentos com DL4J.....	36
21 Resultados dos experimentos com diferentes algoritmos de aprendizado de máquina.....	40

---

## 1 Introdução

No âmbito da Unidade Curricular de **Aprendizagem e Decisão Inteligentes** foi proposto o desenvolvimento e conceção de um projeto cujo objetivo consistiu no desenvolvimento de um modelo de *Machine Learning*, através da utilização de modelos de aprendizagem que foram abordados ao longo do semestre em ambiente **KNIME**.

**KNIME** é uma plataforma livre e de código aberto de análise de dados, construção de relatórios e integração de dados, que tem como principal objetivo ajudar os utilizadores a organizar e entender os mesmos, através dos diversos nós e linguagens de programação disponíveis. Para além disto, o **KNIME**, ainda possui uma interface gráfica que permite a visualização dos fluxos de trabalho, do início ao fim, de maneira fácil e estruturada.

Neste sentido, o trabalho prático encontra-se dividido em duas tarefas: A com um *dataset* escolhido pelo grupo de trabalho e B com um *dataset* fornecido pela equipa docente:

1. **Tarefa A - Satisfação\_de\_Voos**, um *dataset* que revela os dados recolhidos sobre a satisfação de viagens de avião
2. **Tarefa B - Producao\_Vestuario**, um *dataset* cujo foco é prever a produção de vestuário através do *feature actual\_productivity*

**Siglas:**

**KNIME** - Konstanz Information Miner

**Keywords:**

*KNIME, Machine Learning*

---

# Tarefa A - Dataset Satisfação\_de\_Voos

Após uma procura detalhada o grupo decidiu selecionar um *dataset* que fosse robusto quer em termos de tamanho quer na quantidade de *features*. Por isso foi escolhido um *dataset* sobre a Satisfação de Voos.

## 1 Características do Dataset

Primitivamente devemos por começar em tirar algumas apreciações e notas do *dataset* que foi escolhido, para podermos inicializar a um planeamento das tarefas a realizar.

O *dataset* **Satisfação\_de\_Voos** apresenta cerca de 129880 linhas e 24 *features*, cujas estas são apresentadas pela seguinte tabela:

Tabela 1: Variáveis Independentes

Number	Name	Description	Data Type
1	<b>id</b>	ID da entrada	int
2	<b>Gender</b>	Sexo do cliente	string
3	<b>Costumer Type</b>	Tipo de cliente (leal ou não leal)	string
4	<b>Age</b>	Idade do cliente	int
5	<b>Type of Travel</b>	Propósito de viagem (pessoal ou de negócio)	string
6	<b>Class</b>	Classe de viagem (económico, económico mais ou de negócio)	string
7	<b>Flight Distance</b>	Distância do voo	int
8	<b>Inflight wifi service</b>	Satisfação com o serviço de wifi durante o voo (0: não avaliado; 1-5)	int
9	<b>Departure/Arrival time convenient</b>	Conveniência de partida ou chegada do voo (0: não avaliado; 1-5)	int
10	<b>Ease of online booking</b>	Facilidade da marcação do voo online (0: não avaliado; 1-5)	int
11	<b>Gate location</b>	Facilidade de acesso à porta de entrada para o avião (0: não avaliado; 1-5)	int
12	<b>Food and Drink</b>	Satisfação da comida e bebidas distribuídas durante o voo (0: não avaliado; 1-5)	int
13	<b>Online boarding</b>	Satisfação com o embarque online (0: não avaliado; 1-5)	int
14	<b>Seat comfort</b>	Satisfação com o conforto do assento (0: não avaliado; 1-5)	int
15	<b>Inflight entertainment</b>	Satisfação com o entretenimento a bordo do avião (0: não avaliado; 1-5)	int
16	<b>On-board service</b>	Satisfação com o serviço prestado durante o voo (0: não avaliado; 1-5)	int
17	<b>Leg room service</b>	Satisfação com o serviço para o espaço para as pernas dos clientes (0: não avaliado; 1-5)	int
18	<b>Baggage handling</b>	Satisfação com o manuseio das bagagens (0: não avaliado; 1-5)	int
19	<b>Checkin service</b>	Satisfação com o serviço de checkin (0: não avaliado; 1-5)	int
20	<b>Inflight service</b>	Satisfação com o serviço durante o voo (0: não avaliado; 1-5)	int
21	<b>Cleanliness</b>	Satisfação com a limpeza do avião (0: não avaliado; 1-5)	int
22	<b>Departure delay in minutes</b>	Atrasos de partida em minutos	int
23	<b>Arrival delay in minutes</b>	Atrasos de chegada em minutos	int

Tabela 2: Variável Dependente

Number	Name	Description	Data Type
24	<b>Satisfaction</b>	Nível de satisfação do voo	string

Este *dataset* contém os resultados dum questionário sobre a satisfação dos passageiros de viagens de avião. O seguinte problema de **Classificação**. É necessário prever qual dos dois níveis de satisfação de viagens um passageiro pertence:

- Satisfeito
- Neutro ou não satisfeito

Foi utilizado o nodo **CSV Reader** para carregar e inicializar o *dataset*. A figura 1 apresenta um excerto do *dataset*.

Row ID	ColumnID	id	Gender	Customer...	Age	Type of ...	Class	Flight D...	Inf...	De...	Ea...	Fo...	On...	...I...	In...	I...	B...	C...	In...	Depart...	Arrival...	satisfaction			
Row0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	1	5	3	4	3	4	4	5	5	25	18	neutral or dissatisfied		
Row1	1	5047	Male	loyal Cust...	25	Business travel	Business	235	3	2	3	3	1	1	5	3	1	4	1	1	6	6	neutral or dissatisfied		
Row2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	5	5	5	4	3	4	4	5	0	0	0	satisfied		
Row3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	2	2	2	2	2	5	3	1	4	2	11	9	neutral or dissatisfied	
Row4	4	22029	Male	Loyal Customer	23	Business travel	Business	214	3	3	3	4	4	3	3	4	4	3	3	3	0	0	0	satisfied	
Row5	5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1390	3	4	2	1	1	2	1	3	4	4	4	1	0	0	neutral or dissatisfied		
Row6	6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2	4	2	3	2	2	2	2	3	3	4	3	2	9	23	neutral or dissatisfied	
Row7	7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4	3	4	4	5	5	5	5	5	5	4	4	4	0	0	satisfied	
Row8	8	79485	Female	Loyal Customer	41	Business travel	Business	853	1	2	2	2	4	3	3	1	1	2	1	4	1	2	0	0	neutral or dissatisfied
Row9	9	25725	Male	loyal Cust...	20	Business travel	Eco	1061	3	3	4	2	2	2	2	2	3	4	3	2	0	0	0	neutral or dissatisfied	
Row10	10	34691	Female	loyal Cust...	24	Business travel	Eco	1382	4	5	5	4	2	5	2	2	3	3	5	3	2	0	0	neutral or dissatisfied	
Row11	11	51412	Female	Loyal Customer	12	Personal Travel	Eco Plus	308	2	4	2	1	2	1	1	1	2	5	5	1	0	0	neutral or dissatisfied		
Row12	12	98628	Male	Loyal Customer	53	Business travel	Eco	834	1	4	4	1	1	1	1	1	3	4	4	1	28	8	neutral or dissatisfied		
Row13	13	83502	Male	Loyal Customer	33	Personal Travel	Eco	946	4	2	4	3	4	4	4	4	4	5	2	2	4	0	0	satisfied	
Row14	14	95769	Female	Loyal Customer	26	Personal Travel	Eco	453	3	3	3	2	2	3	2	2	3	3	2	2	45	35	neutral or dissatisfied		
Row15	15	100580	Male	loyal Cust...	13	Business travel	Eco	496	2	1	3	4	2	1	4	2	1	4	1	3	4	0	0	neutral or dissatisfied	
Row16	16	71142	Female	Loyal Customer	26	Business travel	Business	2123	3	3	3	4	4	4	4	5	3	4	5	4	4	49	51	satisfied	
Row17	17	127461	Male	Loyal Customer	41	Business travel	Business	2075	4	4	2	4	4	4	5	5	5	5	5	0	10	0	satisfied		
Row18	18	70354	Female	Loyal Customer	45	Business travel	Business	2466	4	4	4	4	3	4	5	5	5	5	3	5	4	7	5	satisfied	
Row19	19	66246	Male	Loyal Customer	38	Personal Travel	Eco	460	2	3	3	2	5	3	5	5	1	2	4	3	2	5	17	18	neutral or dissatisfied
Row20	20	39676	Male	Loyal Customer	9	Business travel	Eco	1174	2	4	2	4	2	1	2	1	5	3	4	3	2	0	4	neutral or dissatisfied	

Figura 1: Excerto do Dataset

## 2 KNIME Workflow

O Workflow desenvolvido apresentasse dividido em diversas secções:

- Leitura de Dados;
- Estudos de Correlação;
- Estatística Descritiva;
- Pré-Processamento de Dados;
- Validação e Geração de Modelos de Machine Learning.

Esta segmentação reflete um caminho ideal para o estude e desenvolvimento de modelos **Machine Learning**.

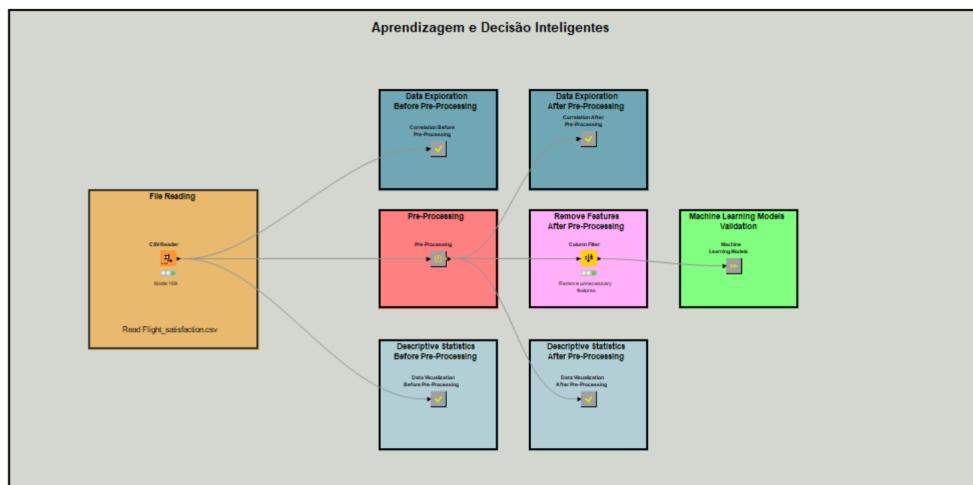


Figura 2: Workflow desenvolvido

### 3 Análise Preliminar dos Dados

#### 3.1 Análise Geral

Como refletido anteriormente foi feita uma análise detalhada dos dados antes de começarmos a desenvolver e manipular o próprio *dataset*.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞	No. -∞	No. Histogram
Column0	<input type="checkbox"/>	0	103903	44158.700	31207.377	973900383.073	0.332	-1.222	5735331956	2	0	0	0	0	
id	<input type="checkbox"/>	1	129880	64940.500	37493.271	1405745356.667	0.000	-1.200	8434472140	0	0	0	0	0	
Age	<input type="checkbox"/>	7	85	39.428	15.119	228.595	-0.004	-0.719	5120903	0	0	0	0	0	
Flight Distance	<input type="checkbox"/>	31	4983	1190.316	997.452	994911.445	1.108	0.266	154598293	0	0	0	0	0	
Inflight wifi service	<input type="checkbox"/>	0	5	2.729	1.329	1.767	0.040	-0.849	354403	3916	0	0	0	0	
Departure/Arrival time convenient	<input type="checkbox"/>	0	5	3.058	1.527	2.331	-0.332	-1.041	397121	6681	0	0	0	0	
Ease of Online booking	<input type="checkbox"/>	0	5	2.757	1.402	1.965	-0.019	-0.914	358063	5682	0	0	0	0	
Gate location	<input type="checkbox"/>	0	5	2.977	1.279	1.635	-0.058	-1.032	386643	1	0	0	0	0	
Food and drink	<input type="checkbox"/>	0	5	3.205	1.330	1.769	-0.155	-1.145	416236	132	0	0	0	0	
Online boarding	<input type="checkbox"/>	0	5	3.253	1.351	1.824	-0.457	-0.699	422452	3080	0	0	0	0	
Seat comfort	<input type="checkbox"/>	0	5	3.441	1.319	1.741	-0.486	-0.923	446964	1	0	0	0	0	

Figura 3: Exploração geral dos dados numéricos

Inflight entertainment	<input type="checkbox"/>	0	5	3.358	1.334	1.780	-0.366	-1.061	436147	18	0	0	0	0	
On-board service	<input type="checkbox"/>	0	5	3.383	1.287	1.657	-0.421	-0.889	439387	5	0	0	0	0	
Leg room service	<input type="checkbox"/>	0	5	3.351	1.316	1.733	-0.348	-0.963	435212	598	0	0	0	0	
Baggage handling	<input type="checkbox"/>	1	5	3.632	1.180	1.392	-0.677	-0.384	471739	0	0	0	0	0	
Checkin service	<input type="checkbox"/>	0	5	3.306	1.266	1.603	-0.367	-0.830	429418	1	0	0	0	0	
Inflight service	<input type="checkbox"/>	0	5	3.642	1.177	1.385	-0.692	-0.358	473048	5	0	0	0	0	
Cleanliness	<input type="checkbox"/>	0	5	3.286	1.314	1.726	-0.301	-1.015	426828	14	0	0	0	0	
Departure Delay in Minutes	<input type="checkbox"/>	0	1592	14.714	38.071	1449.411	6.822	100.645	1911017	73356	0	0	0	0	
Arrival Delay in Minutes	<input type="checkbox"/>	0	1584	15.091	38.466	1479.606	6.670	95.117	1954105	72753	393	0	0	0	

Figura 4: Exploração geral dos dados numéricos

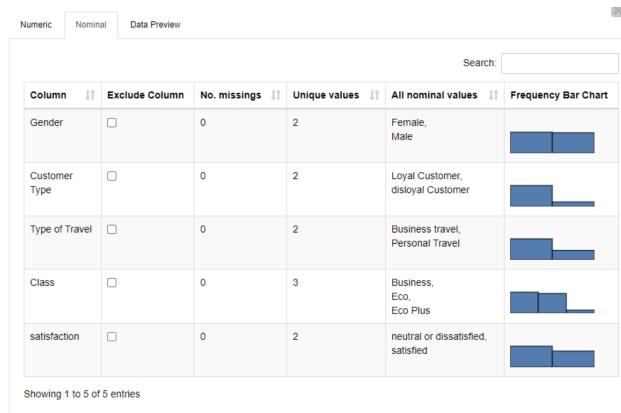


Figura 5: Exploração geral dos dados nominais

Como é possível ver nas últimas duas imagens podemos tirar as seguintes conclusões:

- A feature *Arrival delay in minutes* é a única que apresenta *missing values*;
- As duas primeiras *features* (column0 e id) são inúteis para um problema de classificação e por isso podem ser removidas;
- Podemos dividir as *features* em variáveis quantitativas e categóricas;

### 3.2 Análise da Correlação

Como a *feature* objetivo é a Satisfaction vamos ver como é a distribuição dela pelo dataset:

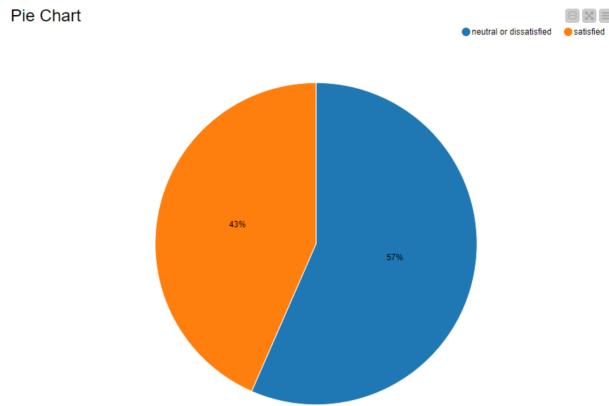
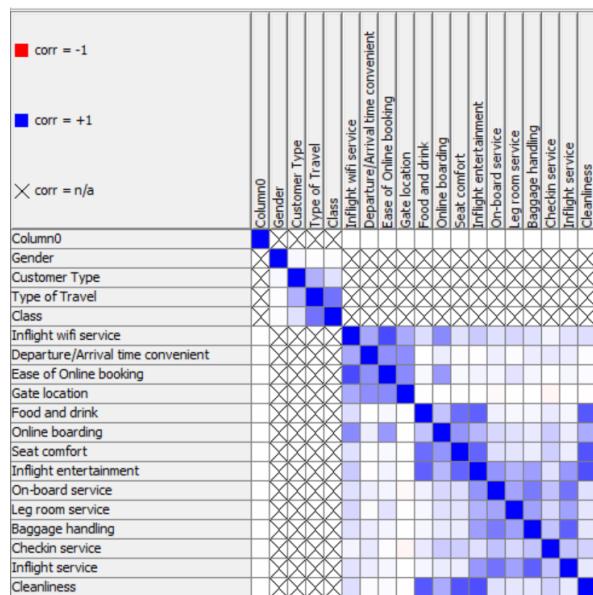
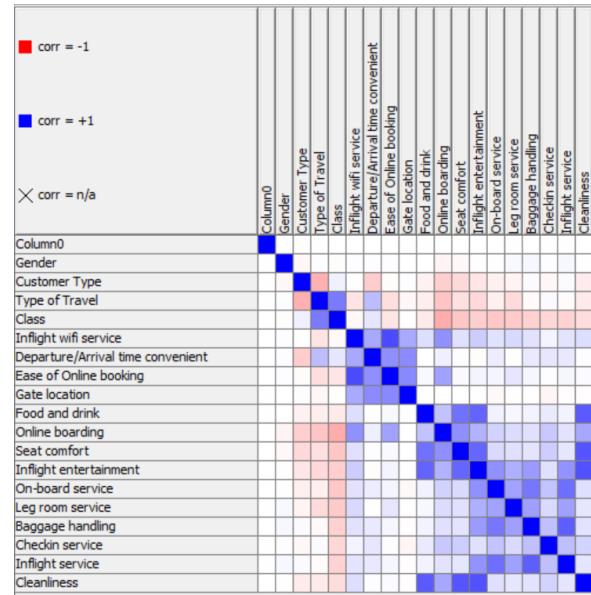


Figura 6: Distribuição da satisfação pelo *dataset*

Com isto vemos que a distribuição é mais ou menos balanceada. Agora podemos ver a correlação entre *features* nominais e não nominais para retirarmos mais informação que nos possa dar dicas para o **pré-processamento** dos dados.

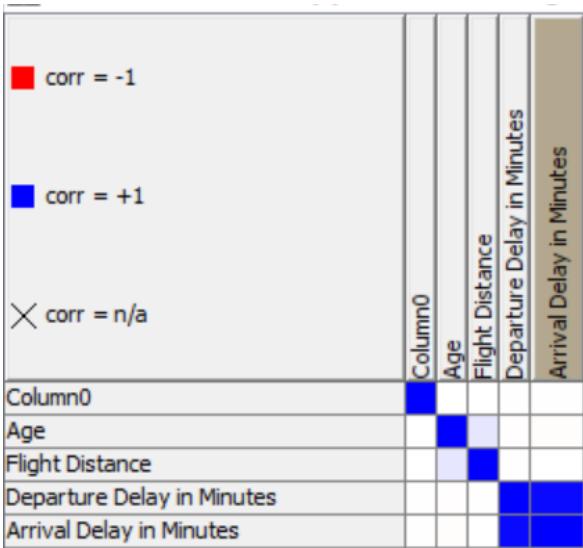


(a) Linear correlation de *features* nominais

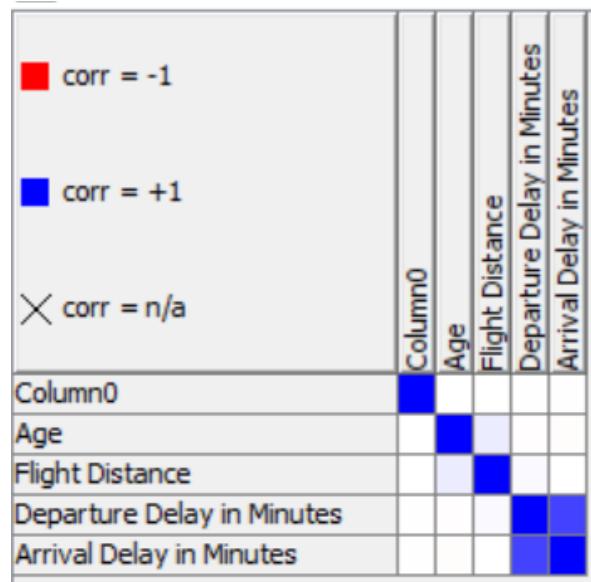


(b) Rank correlation de *features* nominais

Figura 7: Correlação de *features* nominais

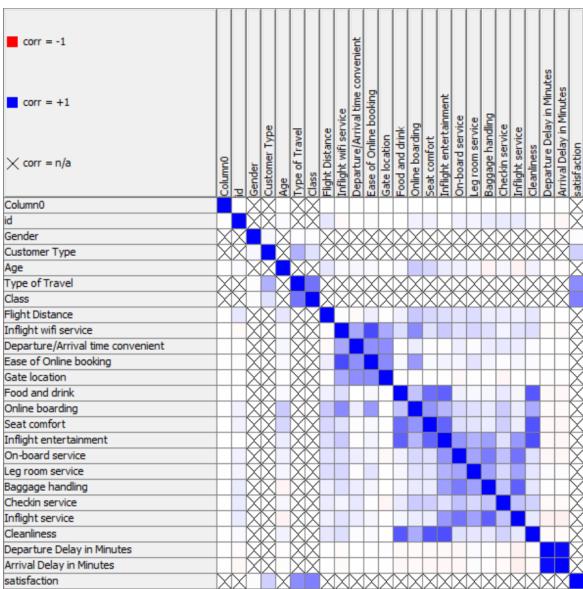


(a) Linear correlation de *features* não nominais

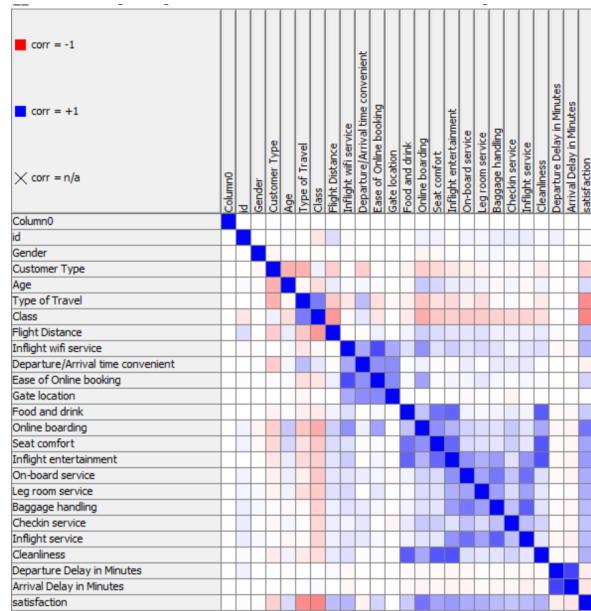


(b) Rank correlation de *features* não nominais

Figura 8: Correlação de *features* não nominais



(a) Linear correlation entre todas as *features*



(b) Rank correlation entre todas as *features*

Figura 9: Correlação de todas as *features*

Como é de esperar existe uma correlação muito forte entre o atraso de partida e de chegada dos voos, cerca de 0.9653 de correlação linear. Isto porque se acontece um atraso de partida dum voo, é muito provável que, mantendo-se tudo normal, haja um atraso de chegada. Podemos agora, também, ver graficamente esta relação a partir dum nodo de *Scatter Plot*:

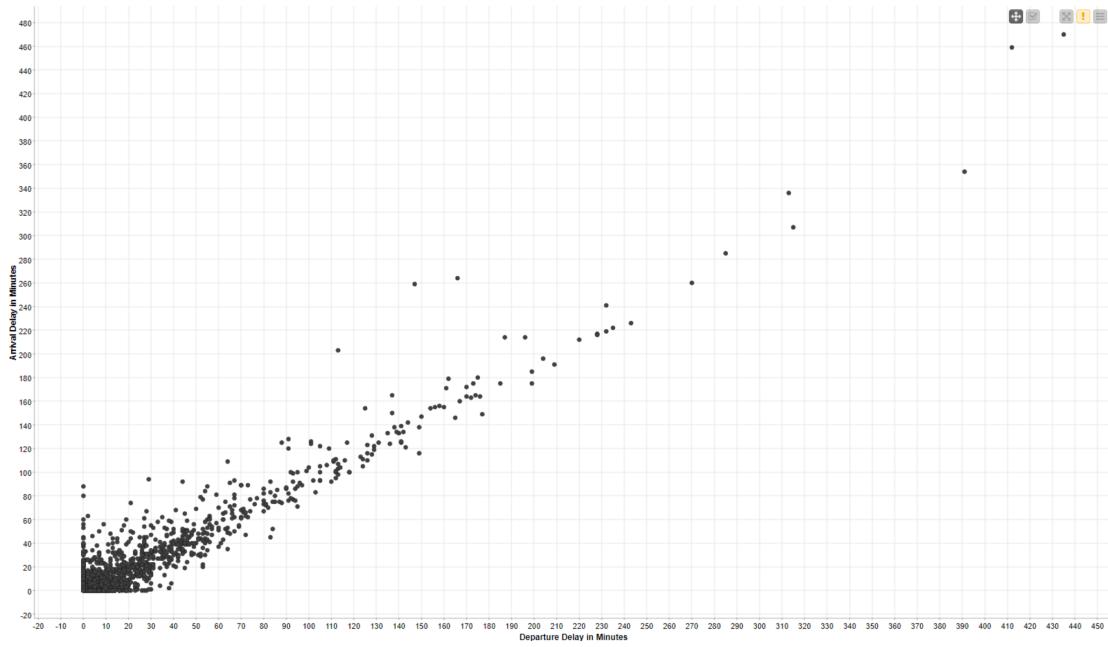


Figura 10: Scatter plot entre a *feature* Arrival Delay in minutes e Departure Delay in minutes

Isto revela que as *feature* **Arrival Delay in minutes** e **Departure Delay in minutes** são muito parecidas, e por isso, podemos considerar a eliminação de uma delas. Sendo assim, com esta análise já temos feitas algumas decisões em relação ao pré-processamento dos dados, iremos remover 3 *features* por se demonstrar que a sua existência não ajuda no estudo que estamos a fazer.

### 3.3 Análise das features

Agora podemos começar por fazer uma breve análise de cada feature

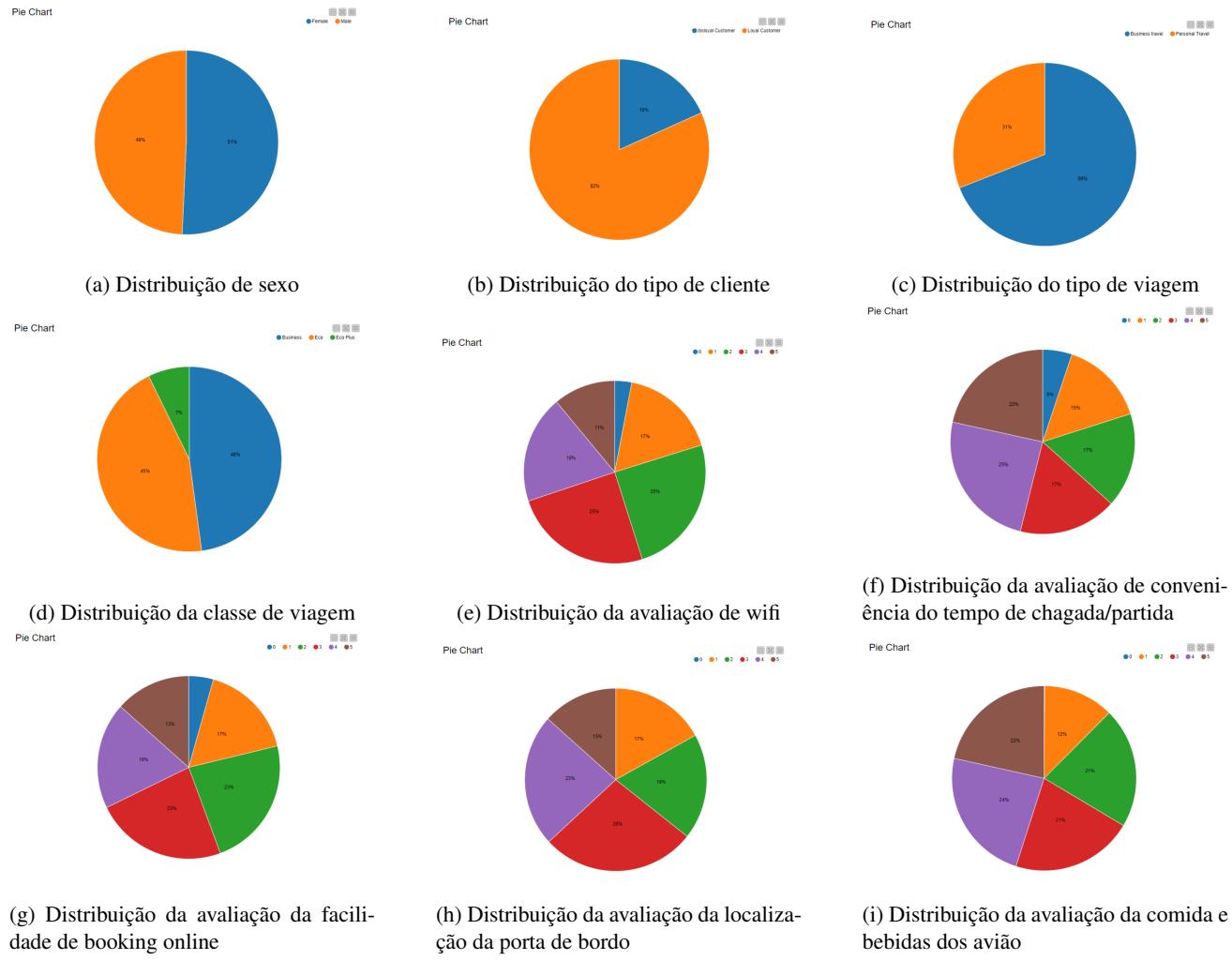


Figura 11: Distribuição das features

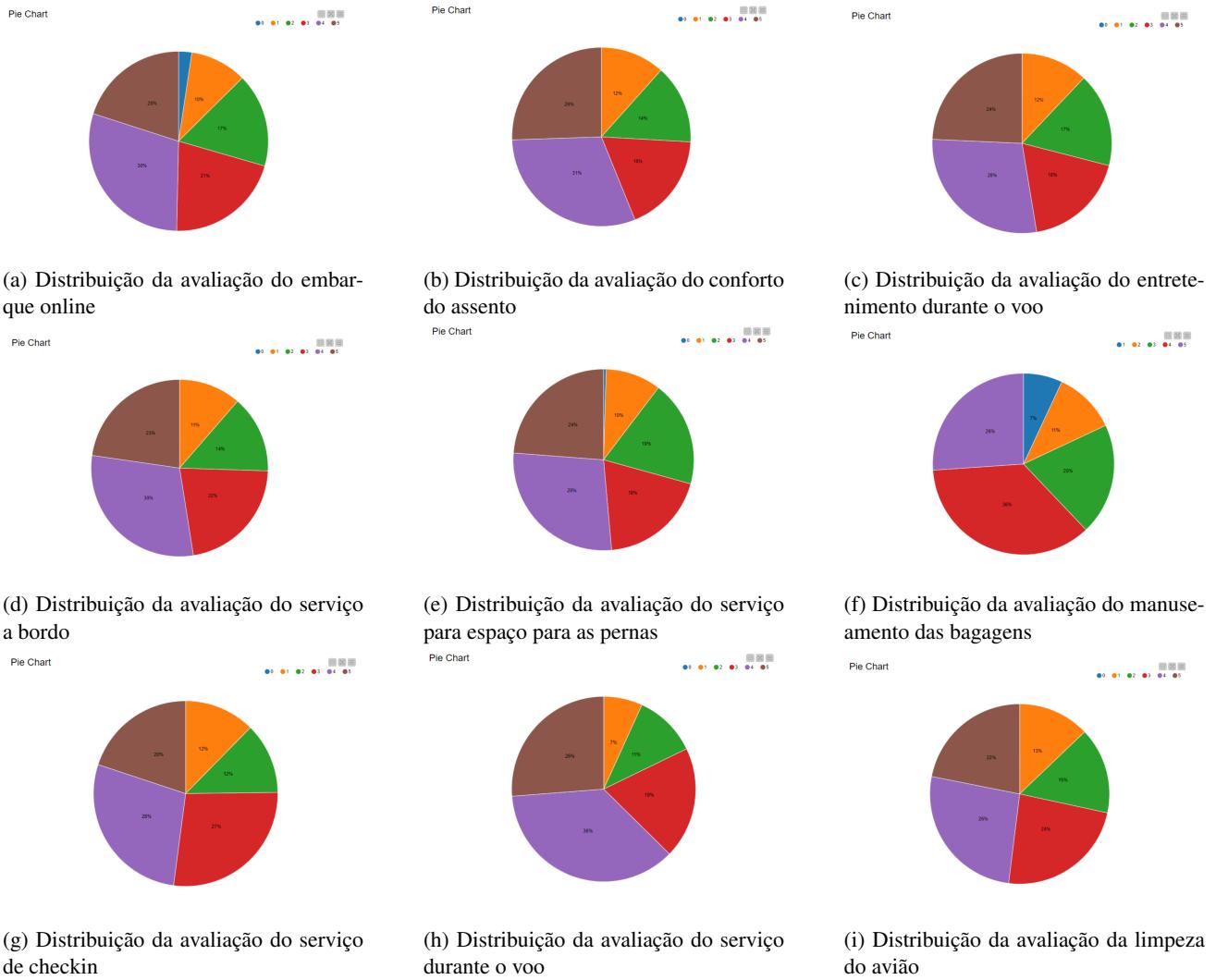
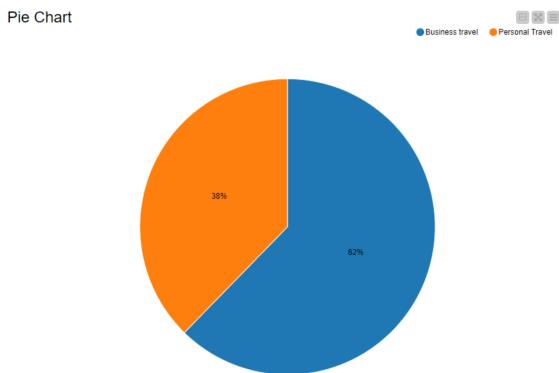


Figura 12: Distribuição das *features*

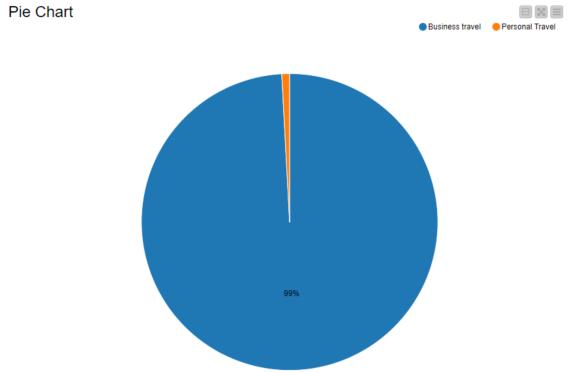
Podemos tirar as seguintes conclusões destes gráficos:

- O número de homens e mulheres é mais ou menos igual;
- A maioria dos clientes desta companhia aérea são clientes do costume;
- A maior parte dos clientes viajou por motivos profissionais e não pessoais;
- Por volta de metade dos clientes viajaram em classe de negócios;
- Cerca de 60% dos clientes ficaram satisfeitos com o manuseio das bagagens;
- Mais de 50% dos clientes ficaram satisfeitos com o conforto dos assentos.

Também podemos analisar os dados por tipo de cliente por exemplo:

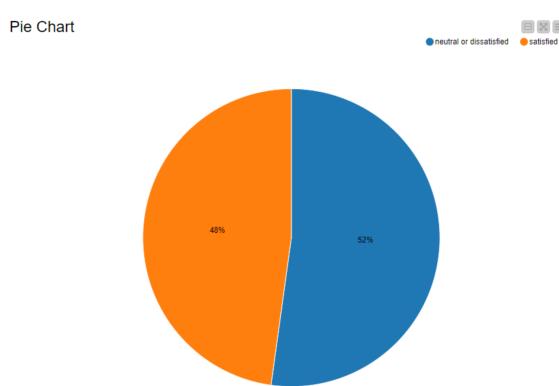


(a) Distribuição do tipo de viagem dos cliente leais

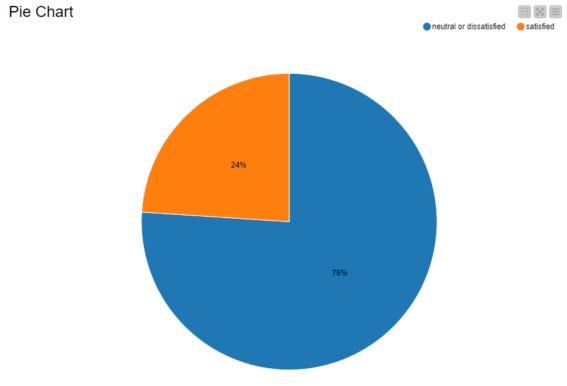


(b) Distribuição do tipo de viagem dos cliente não leais

Figura 13: Distribuição do tipo de viagem por tipo de cliente

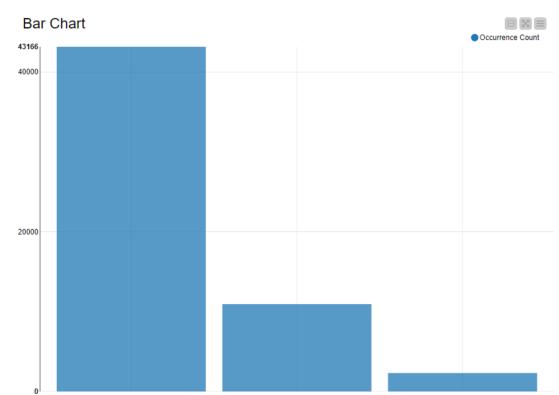


(a) Distribuição da satisfação dos cliente leais

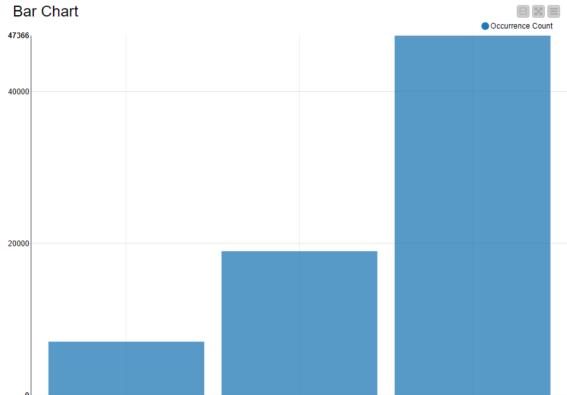


(b) Distribuição da satisfação dos cliente não leais

Figura 14: Distribuição do nível de satisfação por tipo de cliente



(a) Distribuição do tipo de classe dos clientes satisfeitos



(b) Distribuição do tipo de classe dos clientes não satisfeitos

Figura 15: Distribuição do tipo de classe por clientes satisfeitos ou não satisfeitos

## 4 Pré-Processamento

Agora que fizemos uma análise detalhada dos dados podemos começar com o processamento e limpeza destes. Já concluímos que podemos remover algumas *features* por estas não terem impacto significativo no estudo. Portanto, vamos remover as *features* *clomun0*, *id* e *Arrival Delay in minutes* com os seguintes nodos:

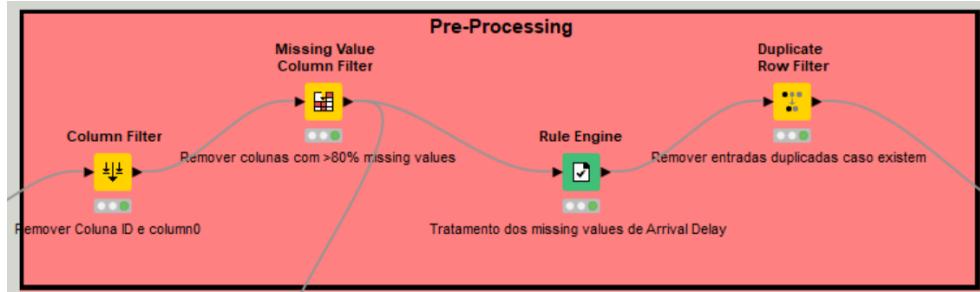


Figura 16: Workflow de pré-processamento

Também é importante mencionar que, como vimos nas secções anteriores, o *dataset* contém missing values, nomeadamente a *feature* *Arrival Delay in minutes*, apesar de esta *feature* ir ser removida o grupo pretende demonstrar que sabe resolver problemas de missing values. Neste caso, como os valores que desapareceram faziam parte do atraso da chegada podemos assumir que se existe um atraso na partida deverá haver um atraso semelhante na chegada, logo os missing values da coluna *Arrival Delay in minutes* foram substituídos pelos valores da coluna *Departure Delay in minutes*. Para tal foi usado o nodo **Rule Engine**:

O código no Rule Engine é:

```
Expression
?
1 // enter ordered set of rules, e.g.:
?
2 // $double column name$ > 5.0 => "Large"
?
3 // $string column name$ LIKE "*blue*" => "small and blue"
?
4 // TRUE => "default outcome"
?
5 MISSING $Arrival Delay in Minutes$ => $Departure Delay in Minutes$
D NOT MISSING $Arrival Delay in Minutes$ => $Arrival Delay in Minutes$
```

Figura 17: Tratamento de missing values

Também foram usados os nodos **Missing Value Column Filter** e **Duplicate Row Filter** para remover uma coluna caso a quantidade de missing values fosse superior a 80% e para remover linhas duplicadas respetivamente. Neste caso, os nodos não modificaram o *dataset* pois este não tinha essa características negativas.

A *feature* *Arrival Delay in minutes* é removida Pós-processamento pelo nodo **Column Filter**

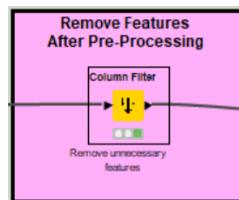


Figura 18: Tratamento de missing values

## 5 Análise dos Dados após o Pré-Processamento

Após o tratamento dos dados, foi feita uma análise destes para rever a eficácia do tratamento feito. Não foram identificados nenhum problema com o processamento feito, e agora que já foi tratado a existência dos missing values podemos prosseguir com o desenvolvimento e treinamento de modelos de machine learning.

## 6 Validação dos Modelos de Machine Learning

A validação dos modelos é idêntica do *dataset* da Tarefa B, é usada o mesmo método mencionado nesta secção da Tarefa B.

## 7 Modelos de Machine Learning

Como o problema apresentado é um problema de **Classificação** podemos para já ter ideias de algoritmos que iremos utilizar. A figura seguinte apresenta a coleção de modelos implementados.

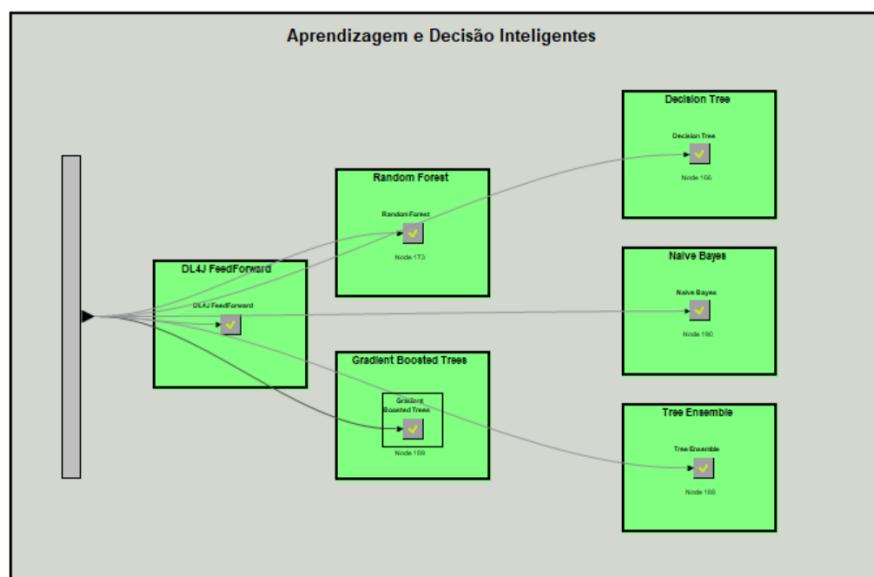


Figura 19: Modelos de classificação

Agora vamos demonstrar os modelos e mostrar o resultado deste com a *feature target* sendo a **Satisfaction**.

## 7.1 Decision Tree

Como mencionado anteriormente a validação dos modelos é feita usando as técnicas de **Hold-out validation** e **Cross validation**.

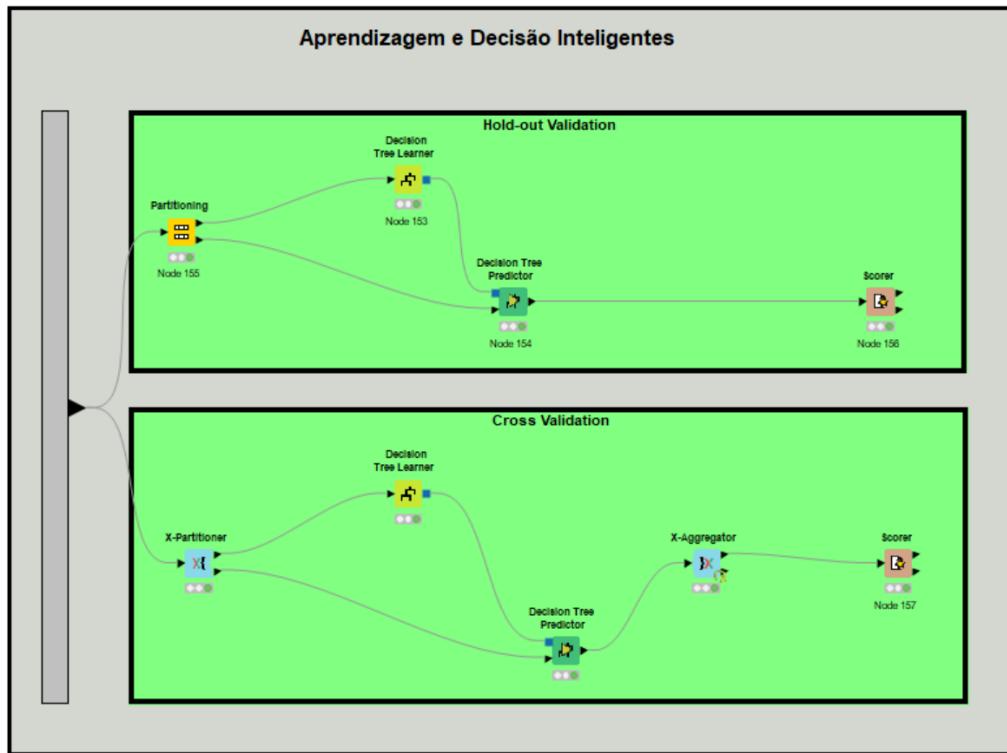


Figura 20: Decision Tree

Tabela 3: Resultados do Tree Leaner Hold-out validation

Quality Measure	Pruning Method	Min number records per node	Number records to store for view	Accuracy	Cohen's Kappa(k)
Gain ratio	No pruning	3	10000	94.999%	0.898%
Gini index	No pruning	3	10000	94.776%	0.894%
Gain ration	MDL	3	10000	96.008%	0.918%
Gini index	MDL	3	10000	95.846%	0.916%
Gain ratio	MDL	5	10000	96.108%	0.92%
Gain ratio	MDL	6	10000	96.166%	0.922%
Gain ratio	MDL	7	10000	96.158%	0.921%

Tabela 4: Resultados do Tree Leaner Cross Validation

Quality Measure	Pruning Method	Min number records per node	Number records to store for view	Accuracy	Cohen's Kappa(k)
Gain ratio	No pruning	3	10000	94.904%	0.896%
Gini index	No pruning	3	10000	94.828%	0.895%
Gain ration	MDL	3	10000	95.91%	0.917%
Gini index	MDL	3	10000	95.908%	0.916%
Gain ratio	MDL	5	10000	95.986%	0.918%
Gain ratio	MDL	6	10000	96.008%	0.918%
Gain ratio	MDL	7	10000	96.006%	0.918%

## 7.2 Naive Bayes

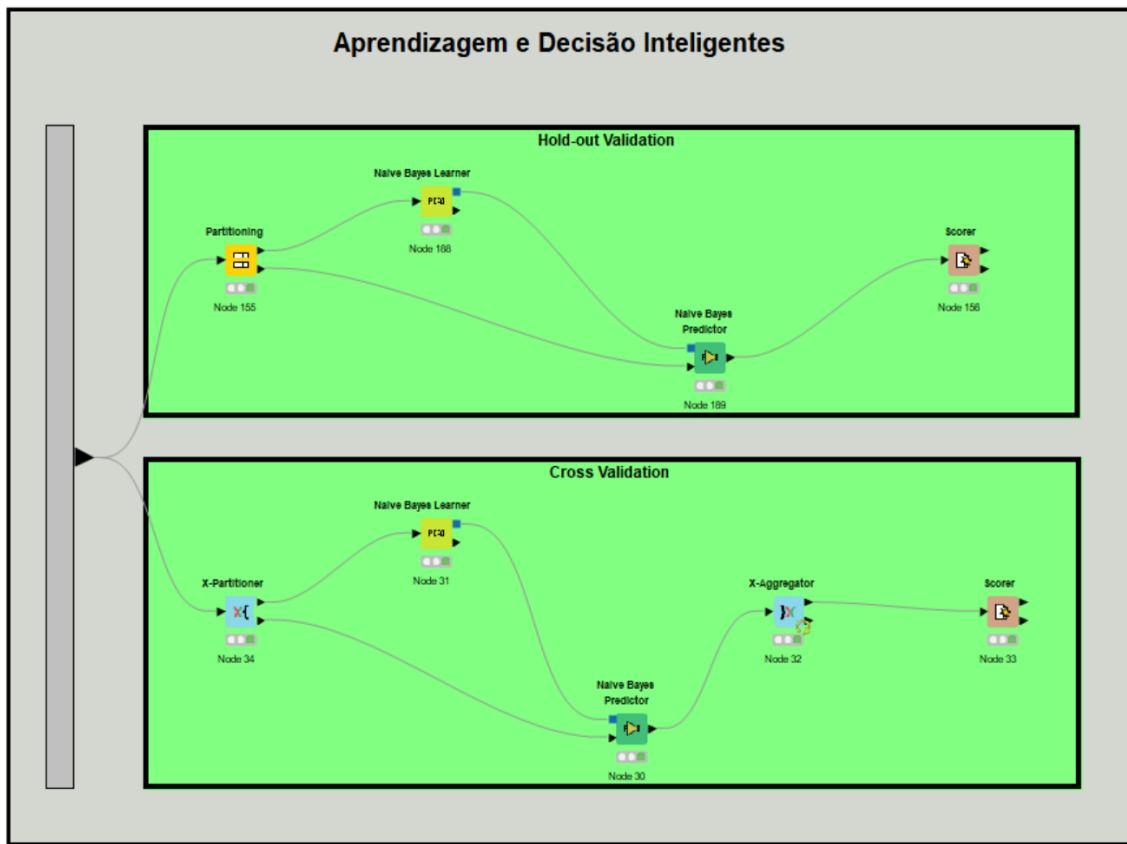


Figura 21: Decision Tree

Tabela 5: Resultados do Naive Bayes Hold-out Validation

Probability	Minimum Standard Deviation	Threshold standard deviation	Maximum number of unique n. values	Accuracy	Kappa
0.0001	0.0001	0.001	10	85.483%	0.704%
0.0003	0.0001	0.001	10	85.483%	0.702%
0.0001	0.0003	0.001	10	85.483%	0.704%
0.0001	0.0001	0.002	10	85.483%	0.704%
0.0001	0.0001	0.001	20	85.483%	0.704%
0.0003	0.0003	0.003	10	85.356%	0.702%

Tabela 6: Resultados do Naive Bayes Cross Validation

Probability	Minimum Standard Deviation	Threshold standard deviation	Maximum number of unique n. values	Accuracy	Kappa
0.0001	0.0001	0.0	20	85.326%	0.701%
0.0002	0.0001	0.0	20	85.333%	0.701%
0.0003	0.0001	0.0	20	85.205%	0.699%
0.0001	0.0002	0.0	20	85.236%	0.701%
0.0001	0.0001	0.0001	20	85.326%	0.701%

### 7.3 Random Forest

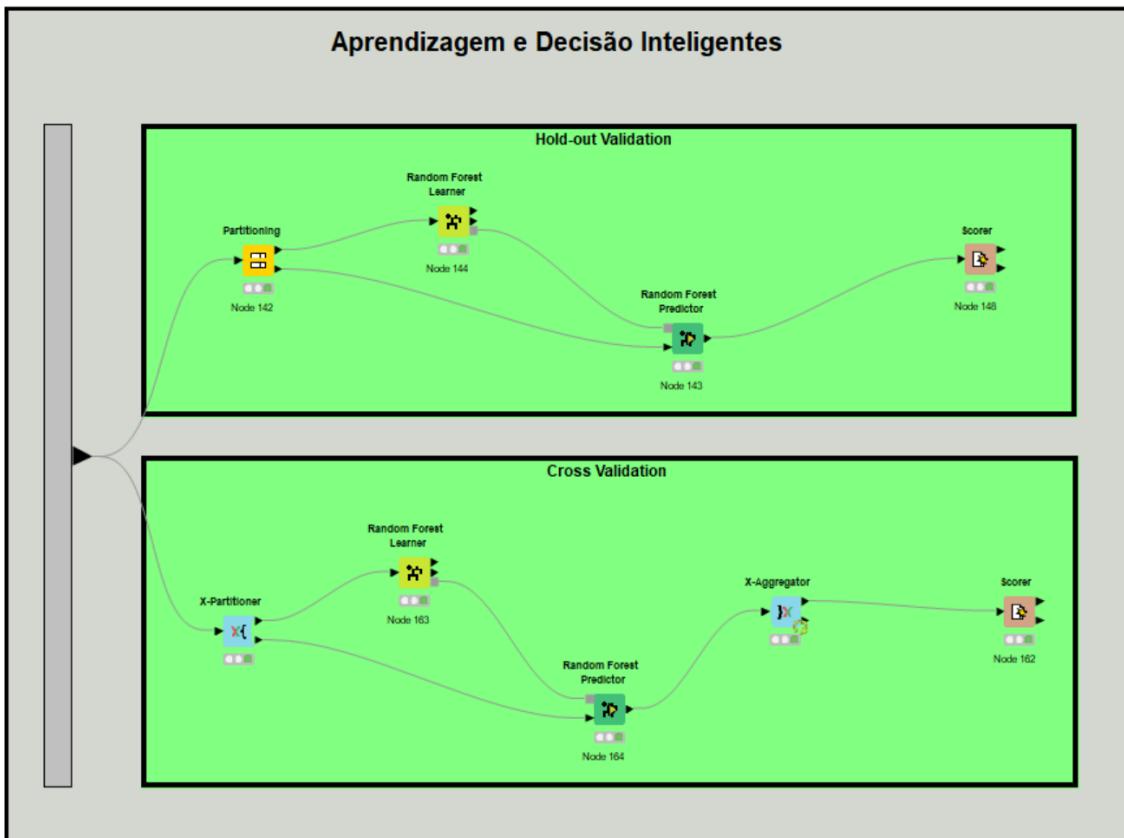


Figura 22: Random Forest

Tabela 7: Random Forest Hold Out Validation

Split Criterion	Score	Cohen's Kappa (k)
Information Gain Ration	96.451 %	0.927 %
Information Gain	96.377 %	0.926 %
Gini Index	96.316 %	0.925 %

Tabela 8: Random Forest Cross Validation

Split Criterion	Score	Cohen's Kappa (k)
Information Gain Ration	96.457 %	0.928 %
Information Gain	96.389 %	0.926 %
Gini Index	96.398 %	0.925 %

## 7.4 Tree Ensemble

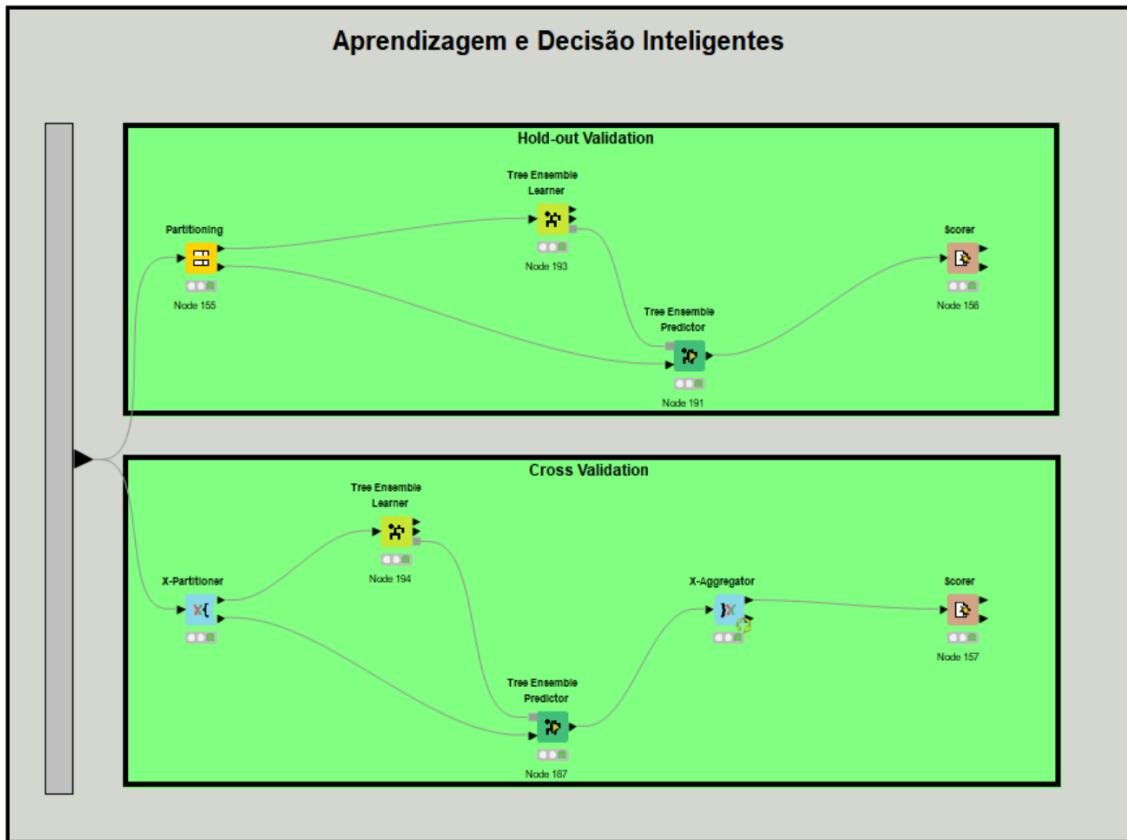


Figura 23: Tree Ensemble

Tabela 9: Resultados do Tree Ensemble Hold-out Validation

Split Criterion	Accuracy	Cohen's Kappa(k)
Information Gain Ratio	96.451%	0.927%
Informatio Gain	96.377%	0.926%
Gini Index	96.254%	0.924%

Tabela 10: Resultados do Tree Ensemble Cross Validation

Split Criterion	Accuracy	Cohen's Kappa(k)
Information Gain Ratio	96.223%	0.923%
Informatio Gain	96.405%	0.927%
Gini Index	96.256%	0.923%

## 7.5 Gradient Boosted Tree

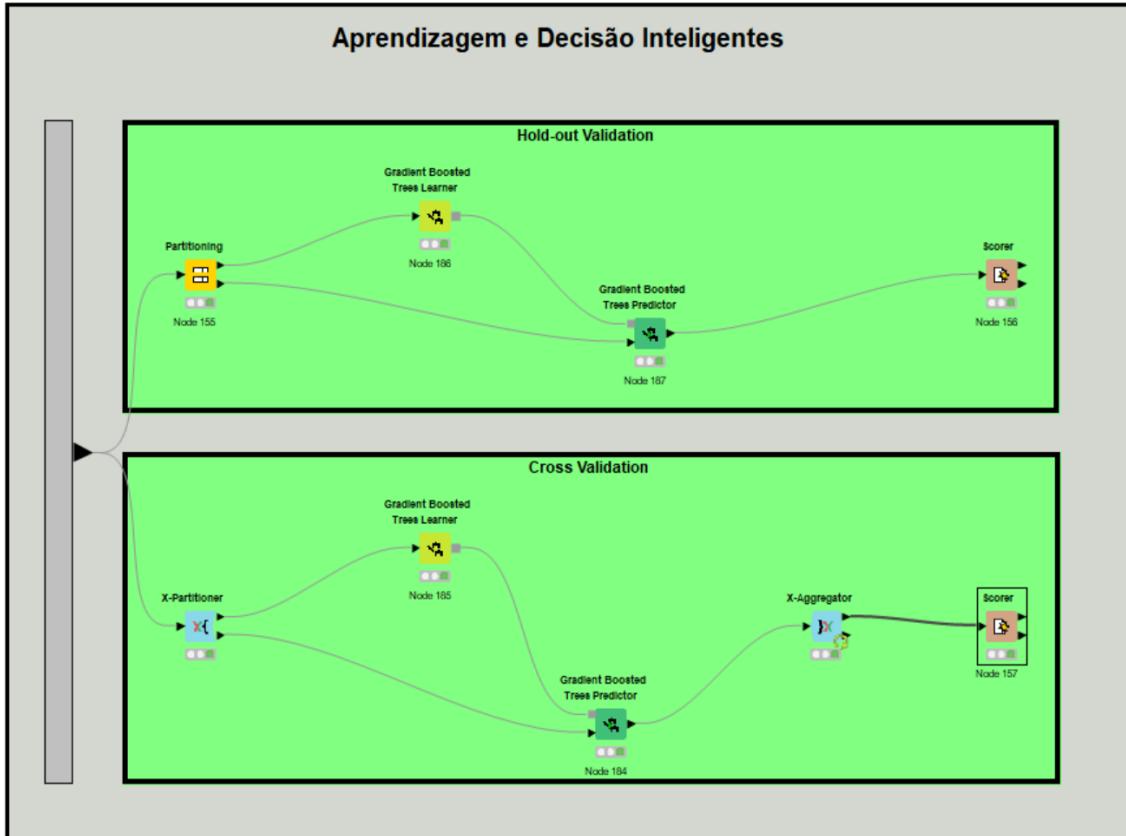


Figura 24: Tree Essemble

Tabela 11: Resultados do Tree Essemble Hold-out Validation

Number of models	Learning rate	Accuracy	Cohen's Kappa(k)
160	0.1	94.618%	0.891%
275	0.1	94.662%	0.891%
275	0.2	94.681%	0.892%
160	0.2	94.668%	0.892%

Tabela 12: Resultados do Tree Essemble Cross Validation

Number of models	Learning rate	Accuracy	Cohen's Kappa(k)
160	0.05	94.158%	0.882%
160	0.2	95.749%	0.913%

Foi difícil realizar vários teste do modelo **Tree Essemble** em modo de Cross Validation devido ao tempo demorado para executar o modelo.

## 7.6 DL4J FeedForward

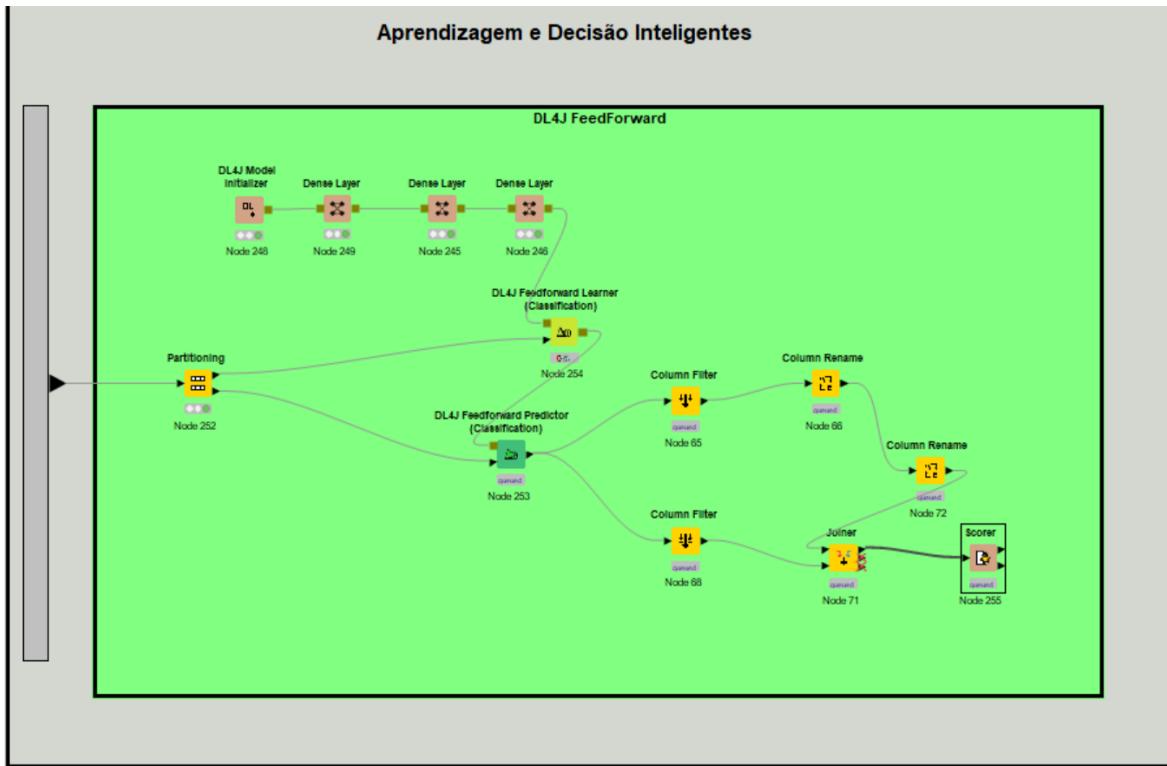


Figura 25: FeedForward

Foi encontrado um problema com este modelo, o predictor da satisfação é sempre o mesmo fazendo com que a precisão do modelo desça para cerca de 50%.

## 8 Análise de Resultados

Deste forma, consideramos que este *dataset* continha uma quantidade de dados bastante proveitosa para efeitos de aprendizagem e que apenas foi necessário de um breve tratamento de dados e erros para obter bons resultados. Sendo que este é um problema de classificação e após os testes feitos podemos afirmar que o modelo que obteve melhores resultados foi o **Random Forest** usando a técnica de **Cross Validation** com uma precisão de 96.457% e um kappa de Cohen de 0.928%.

# Tarefa B - Dataset Producao\_Vestuario

Dado que o número do nosso grupo é o nove foi atribuído, por parte da equipa docente, o **dataset Producao\_Vestuario** com o objetivo de prever a produção de vestuário, ou seja, a variável ***actual\_productivity***.

## 1 Características do Dataset

Inicialmente, foi necessário realizar uma análise cuidada e extensiva do *dataset*, de modo a perceber os dados contidos no mesmo e que informações podemos a priori extrair.

**Produção\_Vestuário** apresenta 1197 linhas e 14 *features*, *features* essas que são explicadas através das tabelas seguintes:

Tabela 13: Variáveis Independentes

Number	Name	Description	Data Type
1	<b>rowID</b>	Register ID	int
2	<b>date</b>	Date in DD/MM/YYYY format	string
3	<b>department</b>	Associated department with the instance	string
4	<b>team</b>	Associated team number with the instance	int
5	<b>targeted_productivity</b>	Targeted productivity set by the Authority for each team for each day	string
6	<b>smv</b>	Standard Minute Value, it is the allocated time for a task	string
7	<b>wip</b>	Work in progress. Includes the number of unfinished items for products	double
8	<b>over_time</b>	Represents the amount of overtime by each team in minutes	int
9	<b>incentive</b>	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action	int
10	<b>idle_time</b>	The amount of time when the production was interrupted due to several reasons	string
11	<b>idle_men</b>	The number of workers who were idle due to production interruption	int
12	<b>no_of_workers</b>	Number of workers in each team	int
13	<b>no_of_style_change</b>	Number of changes in the style of a particular product	string

Tabela 14: Variável Dependente

Number	Name	Description	Data Type
14	<b>actual_productivity</b>	The actual percentage of productivity that was delivered by the workers	string

O carregamento de dados para a plataforma **KNIME** foi efetuado com recurso ao nodo **CSV Reader**. A Figura 1 apresenta um excerto do *dataset*, com dados para as diversas *features* anteriormente descritas.

Row ID	S date	S depart...	I team	S targeted...	S smv	D wip	I over_time	I incentive	S idle_time	I idle_men	I no_of...	S no_of...	S actual...
1	01/01/2015 00:00	swinging	8	0,8	26,16	1,108	7080	98	0	0	0	59	0,940725424
2	01/01/2015 00:00	finishing	1	0,75	3,94	NaN	960	0	0	0	0	8	0,8865
3	01/01/2015 00:00	swinging	11	0,8	11,41	968	3660	50	0	0	0	30,5	0,800570492
4	01/01/2015 00:00	swinging	12	0,8	11,41	968	3660	50	0	0	0	30,5	0,800570492
5	01/01/2015 00:00	swinging	6	0,8	25,9	1,170	1920	50	0	0	0	56	0,800381944
6	01/01/2015 00:00	swinging	7	0,8	25,9	984	6720	38	0	0	0	56	0,800125
7	01/01/2015 00:00	finishing	2	0,75	3,94	NaN	960	0	0	0	0	8	0,75166667
8	01/01/2015 00:00	swinging	3	0,75	28,08	795	6900	45	0	0	0	57,5	0,753683478
9	01/01/2015 00:00	swinging	2	0,75	19,87	733	6000	34	0	0	0	55	0,753097531
10	01/01/2015 00:00	swinging	1	0,75	28,08	681	6900	45	0	0	0	57,5	0,750427826
11	01/01/2015 00:00	swinging	9	0,7	28,08	872	6900	44	0	0	0	57,5	0,721126957
12	01/01/2015 00:00	swinging	10	0,75	19,31	578	6480	45	0	0	0	54	0,712025247
13	01/01/2015 00:00	swinging	5	0,8	11,41	668	3660	50	0	0	0	30,5	0,707045902
14	01/01/2015 00:00	finishing	10	0,65	3,94	NaN	960	0	0	0	0	8	0,705916667
15	01/01/2015 00:00	finishing	8	0,75	2,9	NaN	960	0	0	0	0	8	0,676666667
16	01/01/2015 00:00	finishing	4	0,75	3,94	NaN	2160	0	0	0	0	18	0,593055556

Figura 1: Excerto do Dataset

## 2 KNIME Workflow

O trabalho desenvolvido, na plataforma **KNIME**, para o *dataset* Producao\_Vestuario encontra-se dividido em diversas secções, tal como é apresentado na Figura 2:

- Leitura de Dados;
- Estudos de Correlação;
- Estatística Descritiva;
- Pré-Processamento de Dados;
- Validação e Geração de Modelos de **Machine Learning**.

Esta organização procurou refletir os principais passos a executar aquando da geração de modelos de **Machine Learning**.

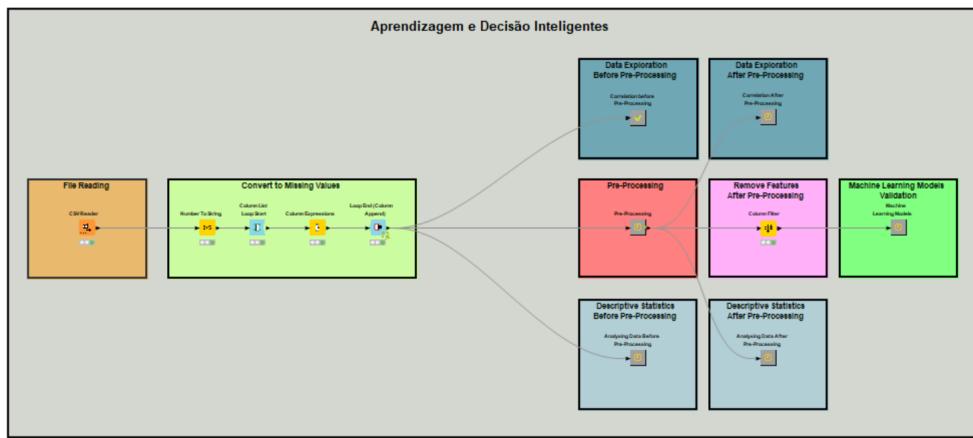


Figura 2: Workflow

Assim sendo, o primeiro passo consiste na leitura do *dataset*, com recurso ao nodo **CSV Reader**, tal como foi mencionado anteriormente.

## 3 Análise Preliminar dos Dados

Ao introduzir o ficheiro *producao\_vestuario.csv* no nodo de leitura verificou-se imediatamente a ausência de dados na *feature wip*. No entanto, como esses campos vazios estavam preenchidos com **Nan** em formato de *double*, não sendo reconhecidos como valores ausentes. Para solucionar o problema, primeiramente converteu-se de *double* para *String* e depois aplicou-se um ciclo, ilustrado na Figura 3, que itera por cada uma das colunas e substitui qualquer string *Nan* por um valor nulo (*null*).

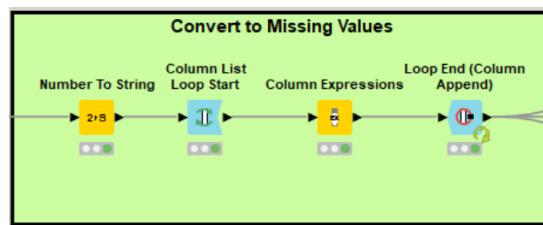
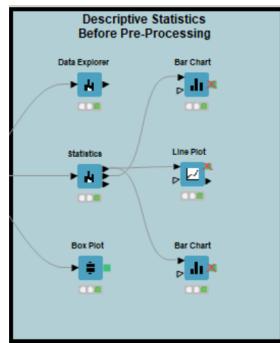
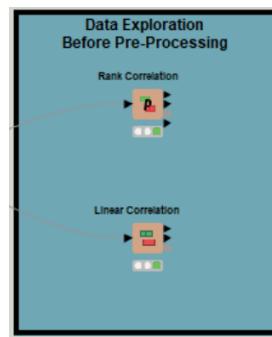


Figura 3: Converter para *Missing Values*

De seguida, elaboramos um estudo geral do estado inicial dos dados presentes no *dataset*, de modo a determinar o tratamento necessário a conduzir para a utilização dos dados na criação de modelos de aprendizagem automática. Para conduzir esta análise inicial, foram utilizadas as secções referentes aos **estudos de correlação** e à **análise estatística descritiva**. A Figura 4 apresenta a expansão dos metanodos presentes nas referidas secções.



(a) Estatísticas descritivas



(b) Estudo de Correlação

Figura 4: Análise Estatística e Estudo de Correlação

A exploração estatística efetuada permitiu conhecer, de um modo geral, o conteúdo de todas as *features* que integram o conjunto de dados anteriormente apresentado. A Figura 5 apresenta o *output* do nodo **Data Explorer** para os dados numéricos. Este nodo permitiu obter informação acerca dos valores extremos das diversas *features*, da sua média e desvio padrão e ainda a presença, ou não, de valores omissos – valores que devem ser tratados aquando da fase de pré-processamento. Conseguimos já observar através das duas seguintes figuras que os tipos das *features* se encontram errados.

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis
team	<input type="checkbox"/>	1	12	6.427	6	3.464	11.999	0.010	-1.224
wip	<input type="checkbox"/>	7	23122	1190.466	1039	1837.455	3376240.881	9.742	101.702
over_time	<input type="checkbox"/>	0	25920	4567.460	3960	3348.624	11214619.255	0.673	0.424
incentive	<input type="checkbox"/>	0	3600	38.211	0	160.183	25658.479	15.791	299.032
ide_men	<input type="checkbox"/>	0	45	0.369	0	3.269	10.686	9.855	102.963
no_of_style_change	<input type="checkbox"/>	0	2	0.150	0	0.428	0.183	2.943	8.181

Figura 5: Data Explorer - Numéricos

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
date	<input type="checkbox"/>	0	59	11/03/2015 00:00, 31/01/2015 00:00, 12/01/2015 00:00, 10/03/2015 00:00, 24/01/2015 00:00, 11/02/2015 00:00, 05/02/2015 00:00, 16/02/2015 00:00, 09/02/2015 00:00, 14/02/2015 00:00, 20/01/2015 00:00	
department	<input type="checkbox"/>	0	5	sweing, finishing, finishng, finnishing, sweng	
targeted_productivity	<input type="checkbox"/>	0	9	0.8, 0.7, 0.75, 0.65, 0.6, ~	

Figura 6: Data Explorer - Nominais

Com recurso a esta variedade de nodos, como, já mencionado, *Data Explorer* ou *Statistics* conseguimos extrair a priori algumas conclusões:

1. Conseguimos, logo, identificar que a *feature Department* continha valores errados, neste caso, os dados encontravam-se mal escritos como é possível ver na Figura seguinte:

department
No. missings: 0
<b>Top 20:</b> sweing : 688 finishing : 496 finishng : 6 finnishing : 4 swenig : 3

Figura 7: Statistics - Department

2. Observamos, da mesma maneira, que a *feature Work in Progress* tinha valores 506 missing values em 1197 linhas.  
 3. Tal como foi mencionado anteriormente, os tipos de algumas *features* encontram-se errados.

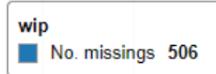
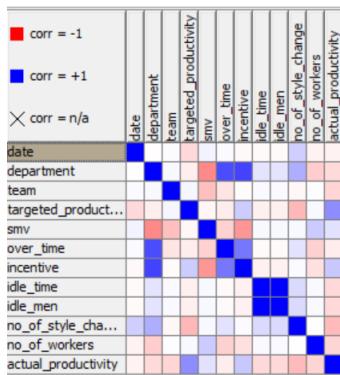


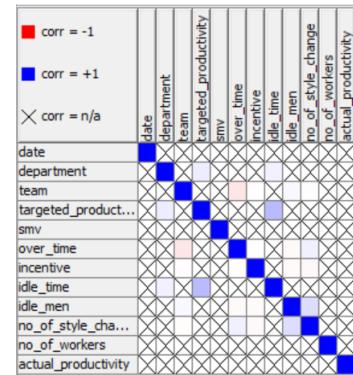
Figura 8: Bar Chart - Work in Progress

A análise de correlação apresentada na Figura X permite medir a força e direção da associação entre duas variáveis, o que pode fornecer informações úteis sobre as *features* que devem ser incorporadas nos modelos de aprendizagem automática. De um modo geral, pode-se observar que existem *features* com um alto grau de correlação, como é o caso do par *idle\_time - idle\_men*. Isso pode indicar que a informação fornecida por essas *features* pode ser redundante.

Focando na variável *target*, *actual\_productivity*, pode-se observar que existe uma correlação positiva com a coluna *target\_productivity* e uma correlação negativa com a coluna *no\_of\_style\_changes*.



(a) Rank Correlation



(b) Linear Correlation

Figura 9: Correlação

## 4 Pré-Processamento

O super-nodo **Pre-Processing**, inserido na secção com o mesmo nome, contempla o trabalho desenvolvido em termos de tratamento e limpeza inicial do conjunto de dados. Após ter sido feito um estudo cuidadoso ao conjunto de dados, desde a existência de *missing values* até a relação entre duas variáveis e a identificação das *features* relevantes para os modelos de aprendizagem automática, foi possível idealizar uma forma de tratar e preparar o *dataset* para a criação de modelos de **Machine Learning**, conforme representado na Figura X.

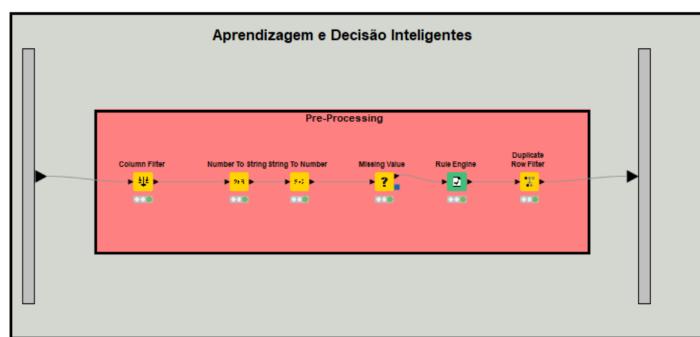


Figura 10: Pré-Processamento

Primeiramente, fez-se uma limpeza geral aos dados, começando por remover:

-> todas as *features* que não influenciam *actual\_productivity*.

-> todos os registos (linhas) repetidos;

Para isso, foram usados os nodos **Column Filter** e **Duplicate Row Filter**.

Após isso, algumas *features* tinham os tipos errados logo foi necessário corrigir os mesmos e alterar para o seu valor correto e para isso usamos os nodos **Number to String** e o **String to Number**.

Depois de alguma exploração dos dados verificamos que os *missing values* relativos a *feature* wip aconteciam sempre que a *feature department* correspondia a *finishing* logo o seu valor tinha que 0. Assim, para resolver isso usamos o nodo **Missing Values** para atribuir o valor 0 a todos os *missing values* presente nessa *feature*.

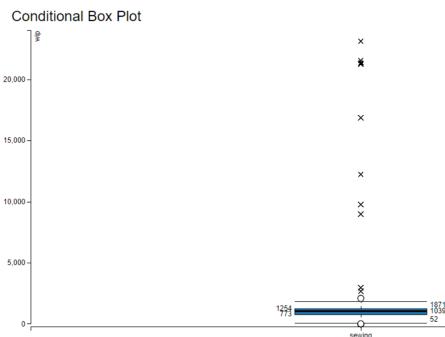


Figura 11: Pré-Processamento

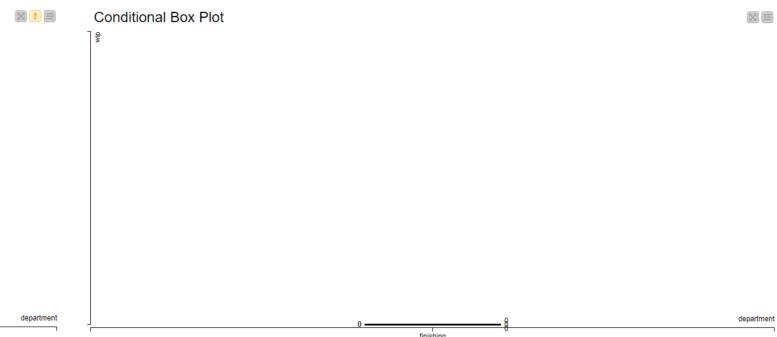


Figura 12: Pré-Processamento

Por fim, corrigimos a *feature Department*. Para isso, usamos o nodo **Rule Engine** com o seguinte código:

```
$department$ MATCHES "finnishing" => "finishing"
$department$ MATCHES "swenig" => "sewing"
$department$ MATCHES "finishnig" => "finishing"
$department$ MATCHES "finishing" => "finishing"
$department$ MATCHES "sweing" => "sewing"
```

Desta maneira foi possível, corrigir os dados que se encontravam escritos de forma errada.

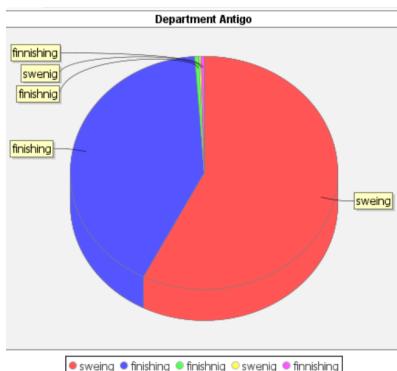


Figura 13: Department Antigo

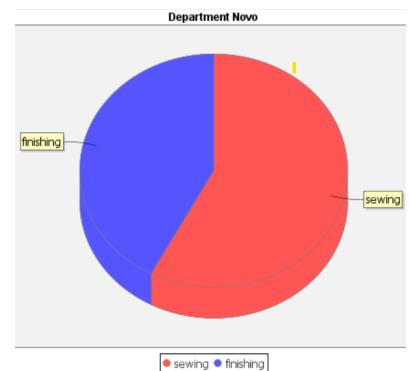


Figura 14: Department Novo

## 5 Análise dos Dados após o Pré-Processamento

Uma vez terminado o tratamento inicial do dataset, foi novamente conduzida uma análise dos dados, que procurou, em parte, confirmar que o trabalho conduzido sobre os dados foi eficaz em termos de, por exemplo, tratamento de valores ausentes. Adicionalmente, foi conduzida uma avaliação correlacional, a fim de observar eventuais diferenças que possam surgir entre as variáveis após o seu tratamento. A Figura 15 apresenta o trabalho executado nesta secção.

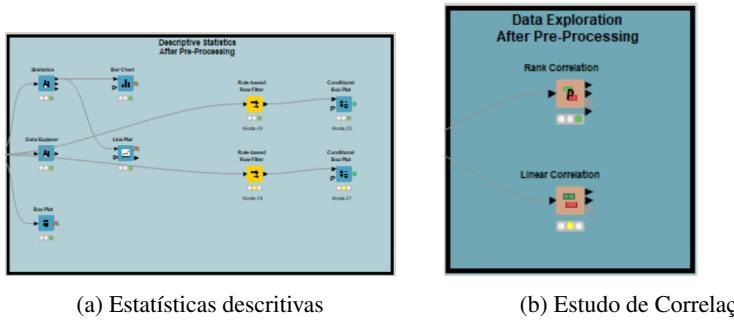


Figura 15: Análise Estatística e Estudo de Correlação

As Figuras 16 e 17 apresentam as saídas geradas pelo nodo **Data Explorer** do **KNIME**, onde podem ser observadas as características das variáveis numéricas e nominais, respectivamente.

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance
targeted_productivity	□	0.070	0.800	0.730	0.750	0.098	0.010
smv	□	2.900	54.560	15.062	15.260	10.943	119.754
wip	□	0	23122	687.228	586	1514.582	2293959.666
incentive	□	0	3600	38.211	0	160.183	25658.479
idle_time	□	0	300	0.730	0	12.710	161.538
idle_men	□	0	45	0.369	0	3.269	10.686
no_of_style_change	□	0	2	0.150	0	0.428	0.183
actual_productivity	□	0.234	1.120	0.735	0.773	0.174	0.030

Figura 16: Data Explorer - Numéricos

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
date	□	0	59	11/03/2015 00:00, 31/01/2015 00:00, 13/02/2015 00:00, 10/03/2015 00:00, 24/01/2015 00:00, ..., 31/02/2015 00:00, 16/02/2015 00:00, 09/02/2015 00:00, 14/02/2015 00:00, 20/01/2015 00:00	
department	□	0	2	sewing, finishing	
team	□	0	12	2, 8, 1, 4, ...	

Figura 17: Data Explorer - Nominais

O grupo optou por manter os *outliers* do *dataset* de maneira a preservar a integridade dos dados, uma vez que todos os *outliers* parecem ser informação válida, ou por outras palavras, representam pontos de dados válidos e legítimos que refletem a realidade do estudo, ao removê-los estaríamos distorcer a representação dos dados e a compreensão do problema.

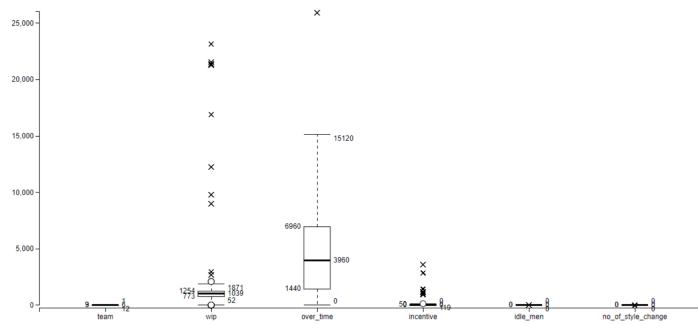


Figura 18: Box Plot

Por fim, como mencionado anteriormente, foi avaliada a correlação entre as diversas variáveis do conjunto de dados. Para isso, foram utilizados os nodos **Rank Correlation** e **Linear Correlation**, cujos resultados estão apresentados na Figura 19, respectivamente.



(a) Rank Correlation



(b) Linear Correlation

Figura 19: Correlação

Como podemos observar através da Figura 19, `idle_time` e `idle_men` são muito correlacionadas entre si por esse motivo decidimos remover a *feature* `idle_time`, uma vez que esta só acontece quando existe `idle_men`, ou seja, quando não existe nenhum trabalho para fazer. Ao removemos esta mesma *feature* estamos a garantir que não agregamos valor adicional à análise e pode até introduzir viés ou redundância. Por outro lado, as *features* `date` e `wip` foram removidas pois apresentação muito baixa correlação com a *feature target*, uma vez que não contribuem significativamente para a capacidade preditiva do modelo.

## 6 Validação dos Modelos de Machine Learning

A validação dos modelos é de elevada importância aquando do desenvolvimento de modelos de aprendizagem automática. Tal acontecimento deve-se, essencialmente, pelo facto de que o uso dos dados do *dataset* na íntegra configura um caso de potencial enviesamento aquando da geração de modelos de **Machine Learning**, isto é, ao usar a totalidade dos dados, sem técnicas de validação, os modelos podem obter um bom desempenho por estarem demasiado ajustados aos dados usados e obter más performances quando deparados com um novo conjunto de dados. Este problema de *overfitting* pode ser reduzido através de técnicas de validação, procurando assim que o desempenho dos modelos criados seja reflexo de padrões encontrados nos dados e não de eventuais ruídos que possam ocorrer no *dataset*.

O trabalho conduzido recorreu a duas técnicas: **hold-out validation** e **cross-validation**. A técnica de validação **hold-out** consiste na segregação do conjunto inicial de dados em dois *datasets*: um usado para treino dos modelos e outro para teste, com vista a avaliar o comportamento do modelo com dados nunca vistos. A técnica de validação consistiu numa divisão de **80%** e **20%** para os *datasets* de treino e teste, respetivamente. A estipulação destes valores centrou-se no facto de:

1. Ser uma divisão usual para o método em uso;
2. A divisão dos *datasets* deve ser efetuada de modo a que sejam alocados dados suficientes tanto para treino como para teste, evitando assim elevada variação na performance dos modelos de **Machine Learning**.

Por outro lado, a técnica de **cross-validation** permite a divisão do conjunto de dados num número,  $k$ , de grupos - motivo pelo qual esta técnica também é conhecida como  **$k$ -fold cross-validation**. Um dos grupos é usado como conjunto de teste para o modelo, e os restantes usados como conjunto de treino. O processo repete-se iterativamente, até que os  $k$  grupos tenham sido usados como *dataset* de teste. A técnica de validação cruzada fixou o valor de  **$k=10$** , uma vez que:

1. É uma definição do número de *folds* usual em **Machine Learning**;
2. É um valor que resulta, geralmente, em modelos com baixo enviesamento.

Independentemente do método de validação, a separação dos dados para teste e treino foi realizada de modo aleatório, com uma *seed* igual a 2023. Esta decisão visa permitir que o desempenho dos modelos, dentro de cada método de validação, sejam comparáveis, uma vez que todos os modelos treinam e testam o mesmo conjunto de dados – apesar de estes conjuntos serem gerados aleatoriamente. As configurações usadas no *workflow* para os nós de **Partitioning** e **X-Partitioner** podem ser consultadas na Figura 20, respectivamente.

First partition Flow Variables Job Manager Selection Memory Policy

Choose size of first partition

Absolute

Relative[%]

Take from top

Linear sampling

Draw randomly

Stratified sampling

Use random seed

Standard settings Flow Variables Job Manager Selection Memory Policy

Number of validations

Linear sampling

Random sampling

Stratified sampling

Class column

Random seed

Leave-one-out

(a) Partitioning

(b) X-Partitioner

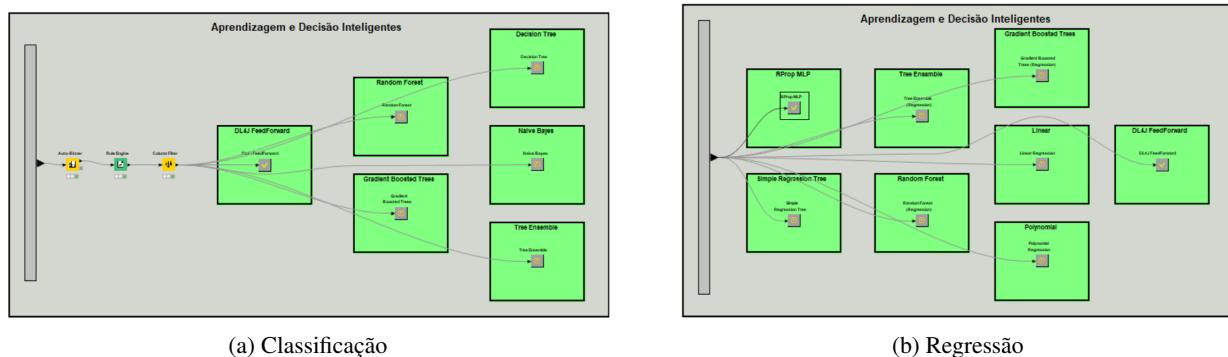
Figura 20: Validação dos Modelos

## 7 Modelos de Machine Learning

Podemos dividir o problema em análise em dois problemas diferentes:

1. Problema de **Classificação**
2. Problema de **Regressão**

A Figura seguinte mostra a expansão dos Modelos que de seguida irão ser abordados individualmente.



(a) Classificação

(b) Regressão

Figura 21: Tipos de Problemas

## Modelos de Classificação

Uma vez que a *feature target* é um número (*double*) optamos por dividir essa mesma *feature* em três *bins* de igual largura, usando o nodo **Auto-Binner**. Após isso, através do **Rule Engine**, mudamos o nome atribuído pelo Auto-Binner tal como é possível observar de seguida:

```
$actual_productivity [Binned]$ = "Bin 1" => "0.529-"
$actual_productivity [Binned]$ = "Bin 2" => "[0.529, 0.825]"
$actual_productivity [Binned]$ = "Bin 3" => "0.825+"
```

Por fim, eliminamos a *feature actual\_productivity* pois com a criação dos *bins* esta mesma está com uma correlação perfeita ou quase perfeita com a *feature binned*, fazendo com que o modelo fique com uma *accuracy* de 100% ou perto.

Seguidamente, iremos mostrar todos os modelos elaborados, bem como, o melhor valor obtido em cada um dos mesmos após várias tentativas e execuções com diferentes hiperparâmetros.

### 7.1 Decision Tree



Figura 22: Valores Obtidos

## 7.2 Random Forest

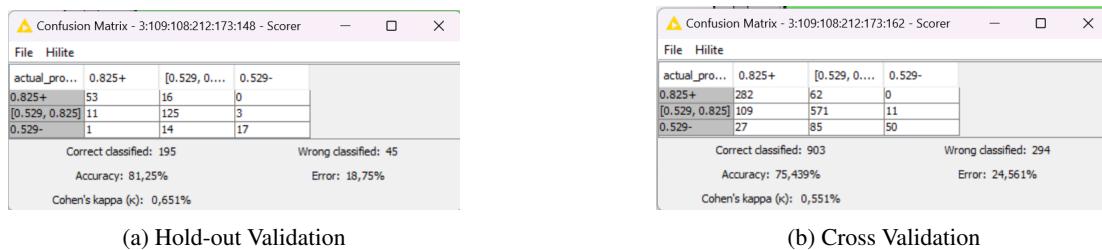


Figura 23: Valores Obtidos

## 7.3 Naive Bayes

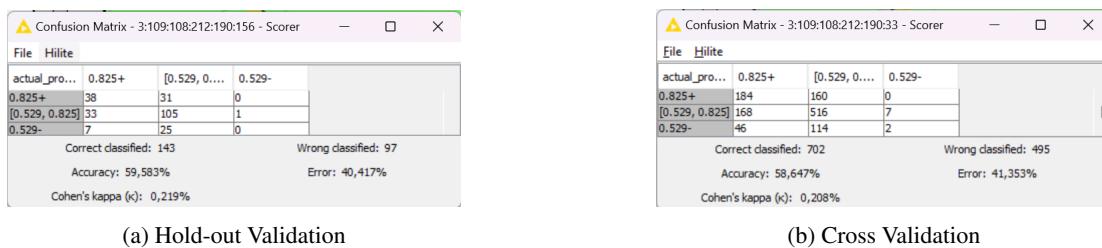


Figura 24: Valores Obtidos

## 7.4 Tree Ensemble

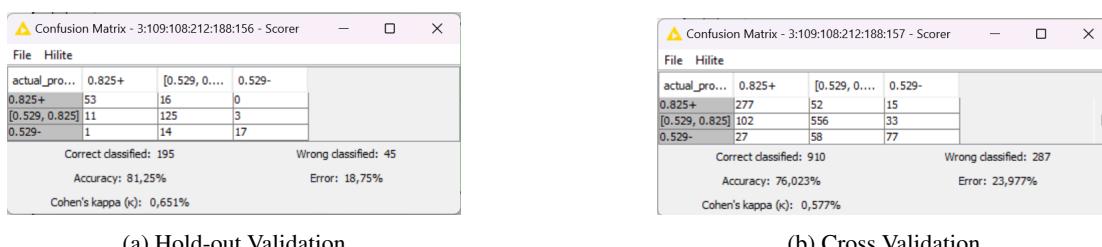
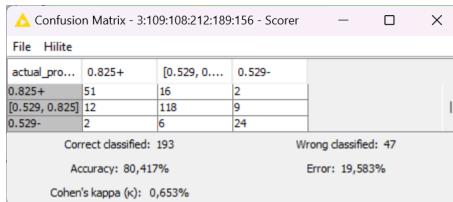
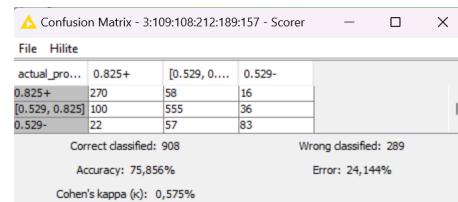


Figura 25: Valores Obtidos

## 7.5 Gradient Boosted Trees



(a) Hold-out Validation



(b) Cross Validation

Figura 26: Valores Obtidos

## 7.6 DL4J FeedForward

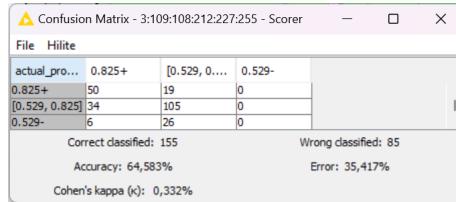


Figura 27: *DL4J FeedForward*

Tree Ensemble		
Validação	Scorer (%)	Detalhes
Hold Out	81.25	Information Gain Ratio Number of Models: 100 Tree Depth: 10 Data Sampling: Random / With Replacement Attribute Sampling: Sample (square root) Attribute Selection: Use different set for each tree node
Cross Validation	76.023	Information Gain Ratio Number of Models: 100 Tree Depth: 10 Data Sampling: Random / With Replacement Attribute Sampling: Sample (square root) Attribute Selection: Use different set for each tree node

Tabela 15: Resultados dos experimentos com Tree Ensemble.

<b>Naïve Bayes</b>		
<b>Validação</b>	<b>Scorer (%)</b>	<b>Detalhes</b>
Hold Out	59.583	Default Probability: 0.0001 Minimum Standard Deviation: 1 Threshold Standard Deviation: 1 Max. Number of Unique Nominal Values per Attribute: 100
Cross Validation	58.647	Default Probability: 0.0001 Minimum Standard Deviation: 1 Threshold Standard Deviation: 1 Max. Number of Unique Nominal Values per Attribute: 20

Tabela 16: Resultados dos experimentos com Naïve Bayes.

<b>Decision Tree</b>		
<b>Validação</b>	<b>Scorer (%)</b>	<b>Detalhes</b>
Hold Out	79.167	Quality Measure: Gain Ratio Pruning Method: No Pruning Min. Number of Records per Node: 3 Number of Records to Store for View: 10000 Number of Threads: 8
Cross Validation	73.434	Quality Measure: Gini Index Pruning Method: No Pruning Min. Number of Records per Node: 3 Number of Records to Store for View: 10000 Number of Threads: 8

Tabela 17: Resultados dos experimentos com Decision Tree.

<b>Random Forest</b>		
<b>Validação</b>	<b>Scorer (%)</b>	<b>Detalhes</b>
Hold Out	81.25	Tree Depth: 10 Number of Models: 100
Cross Validation	75.439	Tree Depth: 10 Number of Models: 275

Tabela 18: Resultados dos experimentos com Random Forest.

Gradient Boosted Trees		
Validação	Scorer (%)	Detalhes
Hold Out	80.417	Number of Models: 160 Learning Rate: 0.1 Missing Values Handling: XGBoost Attribute Sampling: Sample (linear fraction) Attribute Selection: Use same set for entire tree
Cross Validation	75.856	Number of Models: 160 Learning Rate: 0.1 Missing Values Handling: XGBoost Attribute Sampling: Sample (linear fraction) Attribute Selection: Use same set for entire tree

Tabela 19: Resultados dos experimentos com Gradient Boosted Trees.

DL4J		
Validação	Scorer (%)	Detalhes
Hold Out	64.583	Updater: RMSPROP Batch Size: 200 Epochs: 20

Tabela 20: Resultados dos experimentos com DL4J.

## Modelos de Regressão

Como a variável *target* é um número, não foi necessário fazer nenhuma alteração. No entanto, durante a avaliação dos modelos subsequentes, realizamos análises variando os *outliers* da variável *actual\_productivity* para investigar seu impacto na avaliação do modelo.

### 7.7 RProp MLP



Figura 28: Valores Obtidos

### 7.8 Simple Regression Tree

(a) Partitioning

	Value
R <sup>2</sup> :	0.322
Mean absolute error:	0.09
Mean squared error:	0.023
Root mean squared error:	0.15
Mean signed difference:	0.011
Mean absolute percentage error:	0.174
Adjusted R <sup>2</sup> :	0.322

(b) Partitioning com correção de Outliers

	Value
R <sup>2</sup> :	0.336
Mean absolute error:	0.087
Mean squared error:	0.02
Root mean squared error:	0.142
Mean signed difference:	0.009
Mean absolute percentage error:	0.151
Adjusted R <sup>2</sup> :	0.336

(c) X-Partitioner

	Value
R <sup>2</sup> :	0.387
Mean absolute error:	0.08
Mean squared error:	0.019
Root mean squared error:	0.137
Mean signed difference:	-0.001
Mean absolute percentage error:	0.145
Adjusted R <sup>2</sup> :	0.387

(d) X-Partitioner com correção de Outliers

	Value
R <sup>2</sup> :	0.301
Mean absolute error:	0.084
Mean squared error:	0.02
Root mean squared error:	0.141
Mean signed difference:	0.002
Mean absolute percentage error:	0.142
Adjusted R <sup>2</sup> :	0.301

Figura 29: Valores Obtidos

## 7.9 Tree Ensemble (Regression)

(a) Partitioning

	Value
R <sup>2</sup> :	0.399
Mean absolute error:	0.092
Mean squared error:	0.02
Root mean squared error:	0.141
Mean signed difference:	0.01
Mean absolute percentage error:	0.18
Adjusted R <sup>2</sup> :	0.399

(b) Partitioning com correção de Outliers

	Value
R <sup>2</sup> :	0.443
Mean absolute error:	0.084
Mean squared error:	0.017
Root mean squared error:	0.13
Mean signed difference:	0.007
Mean absolute percentage error:	0.147
Adjusted R <sup>2</sup> :	0.443

(c) X-Partitioner

	Value
R <sup>2</sup> :	0.439
Mean absolute error:	0.083
Mean squared error:	0.017
Root mean squared error:	0.131
Mean signed difference:	0.003
Mean absolute percentage error:	0.152
Adjusted R <sup>2</sup> :	0.439

(d) X-Partitioner com correção de Outliers

	Value
R <sup>2</sup> :	0.441
Mean absolute error:	0.084
Mean squared error:	0.016
Root mean squared error:	0.126
Mean signed difference:	0.004
Mean absolute percentage error:	0.144
Adjusted R <sup>2</sup> :	0.441

Figura 30: Valores Obtidos

## 7.10 Random Forest (Regression)

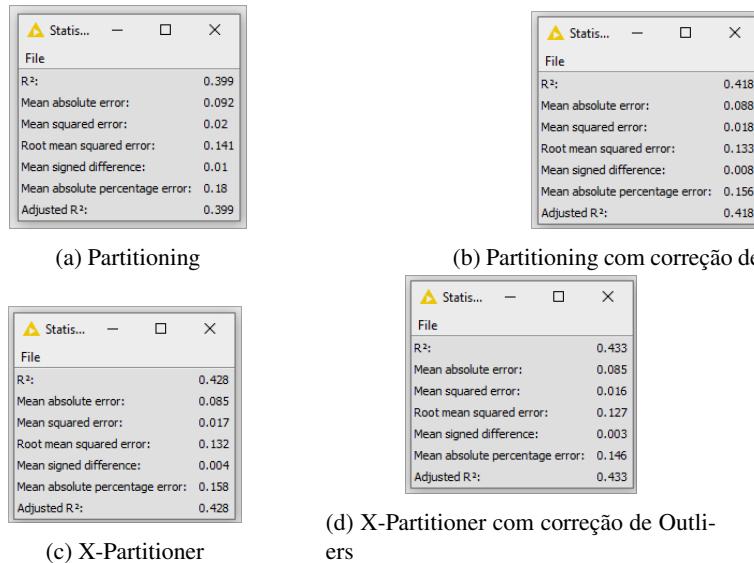


Figura 31: Valores Obtidos

## 7.11 Gradient Boosted Trees (Regression)

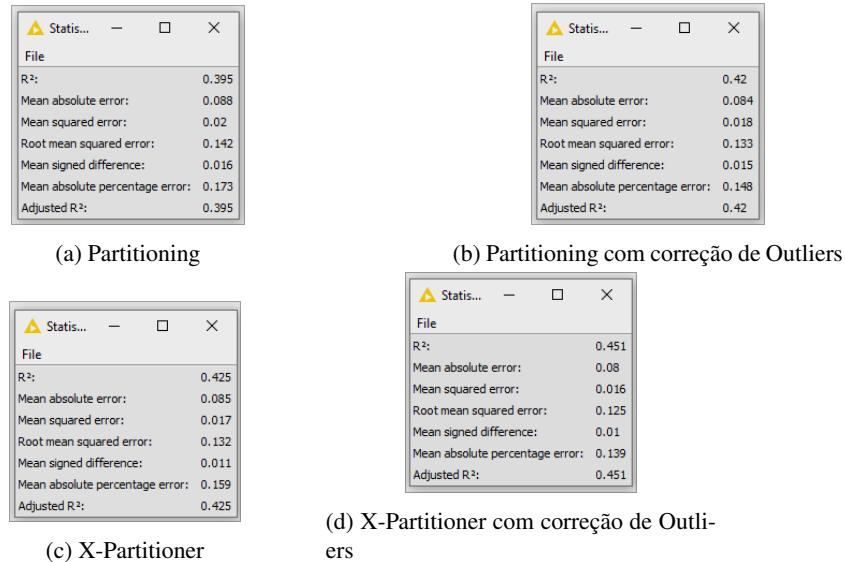


Figura 32: Valores Obtidos

## 7.12 Polynomial Regression

(a) Partitioning

	Value
R <sup>2</sup> :	0.27
Mean absolute error:	0.109
Mean squared error:	0.024
Root mean squared error:	0.156
Mean signed difference:	0.005
Mean absolute percentage error:	0.198
Adjusted R <sup>2</sup> :	0.27

(b) Partitioning com correção de Outliers

	Value
R <sup>2</sup> :	0.266
Mean absolute error:	0.107
Mean squared error:	0.022
Root mean squared error:	0.149
Mean signed difference:	0.004
Mean absolute percentage error:	0.177
Adjusted R <sup>2</sup> :	0.266

(c) X-Partitioner

	Value
R <sup>2</sup> :	0.24
Mean absolute error:	0.108
Mean squared error:	0.023
Root mean squared error:	0.152
Mean signed difference:	-0
Mean absolute percentage error:	0.19
Adjusted R <sup>2</sup> :	0.24

(d) X-Partitioner com correção de Outliers

	Value
R <sup>2</sup> :	0.25
Mean absolute error:	0.106
Mean squared error:	0.021
Root mean squared error:	0.146
Mean signed difference:	-0
Mean absolute percentage error:	0.175
Adjusted R <sup>2</sup> :	0.25

Figura 33: Valores Obtidos

## 7.13 Linear Regression

(a) Partitioning

	Value
R <sup>2</sup> :	0.251
Mean absolute error:	0.11
Mean squared error:	0.025
Root mean squared error:	0.158
Mean signed difference:	0.003
Mean absolute percentage error:	0.202
Adjusted R <sup>2</sup> :	0.251

(b) Partitioning com correção de Outliers

	Value
R <sup>2</sup> :	0.255
Mean absolute error:	0.107
Mean squared error:	0.023
Root mean squared error:	0.15
Mean signed difference:	0.002
Mean absolute percentage error:	0.178
Adjusted R <sup>2</sup> :	0.255

(c) X-Partitioner

	Value
R <sup>2</sup> :	0.256
Mean absolute error:	0.108
Mean squared error:	0.023
Root mean squared error:	0.15
Mean signed difference:	0
Mean absolute percentage error:	0.189
Adjusted R <sup>2</sup> :	0.256

(d) X-Partitioner com correção de Outliers

	Value
R <sup>2</sup> :	0.273
Mean absolute error:	0.105
Mean squared error:	0.021
Root mean squared error:	0.144
Mean signed difference:	0
Mean absolute percentage error:	0.173
Adjusted R <sup>2</sup> :	0.273

Figura 34: Valores Obtidos

Algoritmo		Método	Validação	Detalhes	Scorer			
					1°	2°	3°	4°
Tree Ensemble	Hold-out	Partitioning		0.399 0.092 0.020				
		Numeric Outliers		0.443 0.084 0.017 0.130				
	Cross	X-Partitioner		0.439 0.083 0.017 0.131				
		Numeric Outliers		0.441 0.084 0.016 0.126				
Random Forest	Hold-out	Partitioning		0.401 0.092 0.020 0.141				
		Numeric Outliers		0.418 0.088 0.018 0.133				
	Cross	X-Partitioner		0.440 0.084 0.016 0.126				
		Numeric Outliers		0.433 0.085 0.016 0.127				
Gradient Boosted	Hold-out	Partitioning		0.395 0.088 0.020 0.142				
		Numeric Outliers		0.418 0.085 0.018 0.133				
	Cross	X-Partitioner		0.425 0.085 0.017 0.132				
		Numeric Outliers		0.451 0.080 0.016 0.125				
Linear Regression	Hold-out	Partitioning		0.251 0.110 0.025 0.158				
		Numeric Outliers		0.255 0.107 0.023 0.150				
	Cross	X-Partitioner		0.256 0.108 0.023 0.150				
		Numeric Outliers		0.273 0.105 0.021 0.144				
Polynomial Regression	Hold-out	Partitioning		0.270 0.109 0.024 0.156				
		Numeric Outliers		0.266 0.107 0.022 0.149				
	Cross	X-Partitioner		0.240 0.108 0.023 0.152				
		Numeric Outliers		0.250 0.106 0.021 0.146				
Rprop MLP	Hold-out	Partitioning		0.405 0.093 0.018 0.135				
	Cross	X-Partitioner		0.413 0.087 0.018 0.134				
Simple Regression Tree	Hold-out	Partitioning		0.322 0.090 0.023 0.150				
		Numeric Outliers		0.336 0.087 0.020 0.142				
	Cross	X-Partitioner		0.336 0.087 0.020 0.142				
		Numeric Outliers		0.301 0.084 0.020 0.141				

Tabela 21: Resultados dos experimentos com diferentes algoritmos de aprendizado de máquina.

## 8 Resultados Obtidos

Nos **Modelos de Classificação** usando a técnica de **Hold-Out** existem dois valores cujo resultado é igual, tendo uma *accuracy* de **81,25%**, o Modelo **Random Forest** e **Tree Ensemble**, por outro lado, usando a técnica de **Cross Validation** o modelo que obteve melhor resultado foi novamente o Modelo **Tree Ensemble**, com uma *accuracy* de 76,023%. Mencionar também o uso de **Redes Neuronais, DL4J FeedForward**, para classificação que obteve uma *accuracy* de **64,583%**.

Nos **Modelos de Regressão** usando a técnica de **Hold-Out** sem correção de Outliers, o modelo que se destacou foi o **RProp MLP** obtendo  $r$  ao quadrado de 0,405 e 0,093 / 0,018 / 0,135 nos erros seguintes; usando técnica **Hold-Out** com correção de *Outliers*, o modelo que se destacou foi **Tree Ensemble**, obtendo  $r$  ao quadrado de 0,443 e erros seguintes de 0,084 / 0,017 / 0,13; por outro lado, usando a técnica de **Cross Validation** sem correção de *Outliers*, o modelo que obteve melhor resultado foi novamente o Modelo **Tree Ensemble**, com um  $r$  ao quadrado de 0,439 e erros seguintes de 0,083 / 0,017 / 0,131; por fim, **Gradiente Boosted Trees** foi o melhor modelo de **Cross Validation** com correção de *Outliers* obtendo  $r$  ao quadrado de 0,451 e erros de 0,08 / 0,016 / 0,125.

## 9 Conclusão

Durante a execução e desenvolvimento deste Trabalho Prático, visualizamos e percebemos todas as dificuldades que existem no tratamento de dados, considerando esta uma experiência enriquecedora.

Além de aprendermos a trabalhar com a plataforma **KNIME**, foi possível obter conhecimentos de filtragem e depuramento de dados (*datasets*), bem como vários algoritmos de *Machine Learning*, aprendendo a trabalhar com os mesmos, treinando a capacidade de decisão e previsão das máquinas face aos dados que elas analisam.

Em suma, estamos satisfeitos com o resultado obtido neste Trabalho Prático. Concluímos que conseguimos realizar um tratamento de dados versátil e eficiente e utilizamos diversos Algoritmos de Previsão e Treino de Dados para obtenção dos melhores resultados possíveis.