



Universidade do Minho  
Departamento de Informática

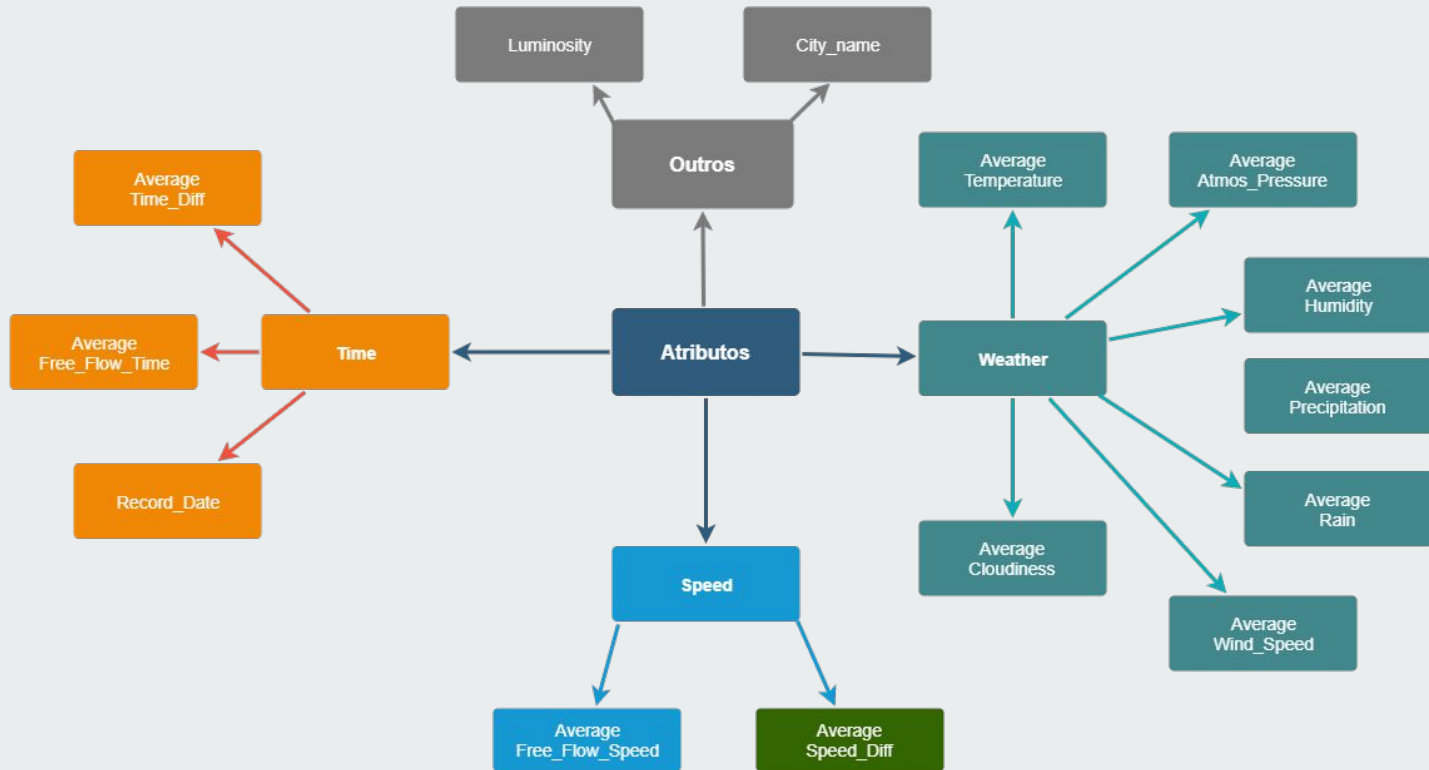
# Árvores de Decisão

## Grupo 3

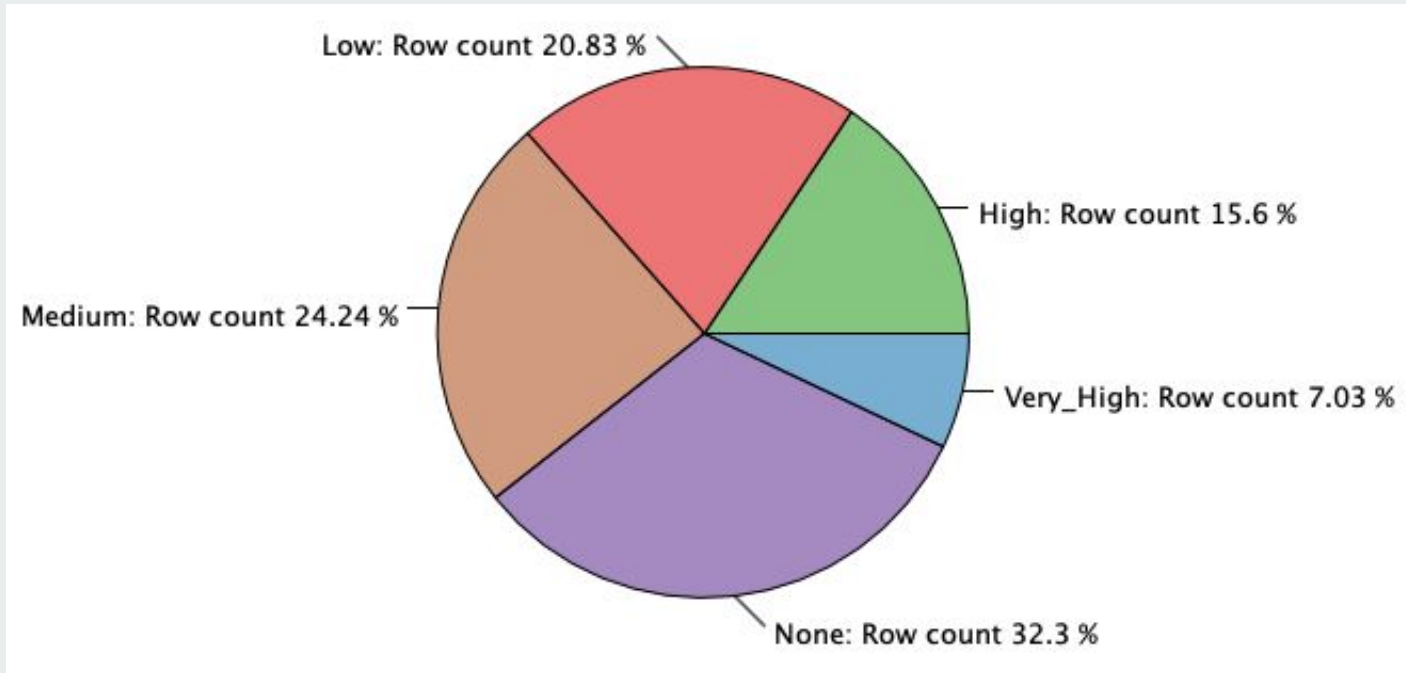
Sistemas Baseados em Similaridade

Universidade do Minho, Mestrado Integrado em Engenharia Informática,  
4º Ano, 1º Semestre, Novembro 2019

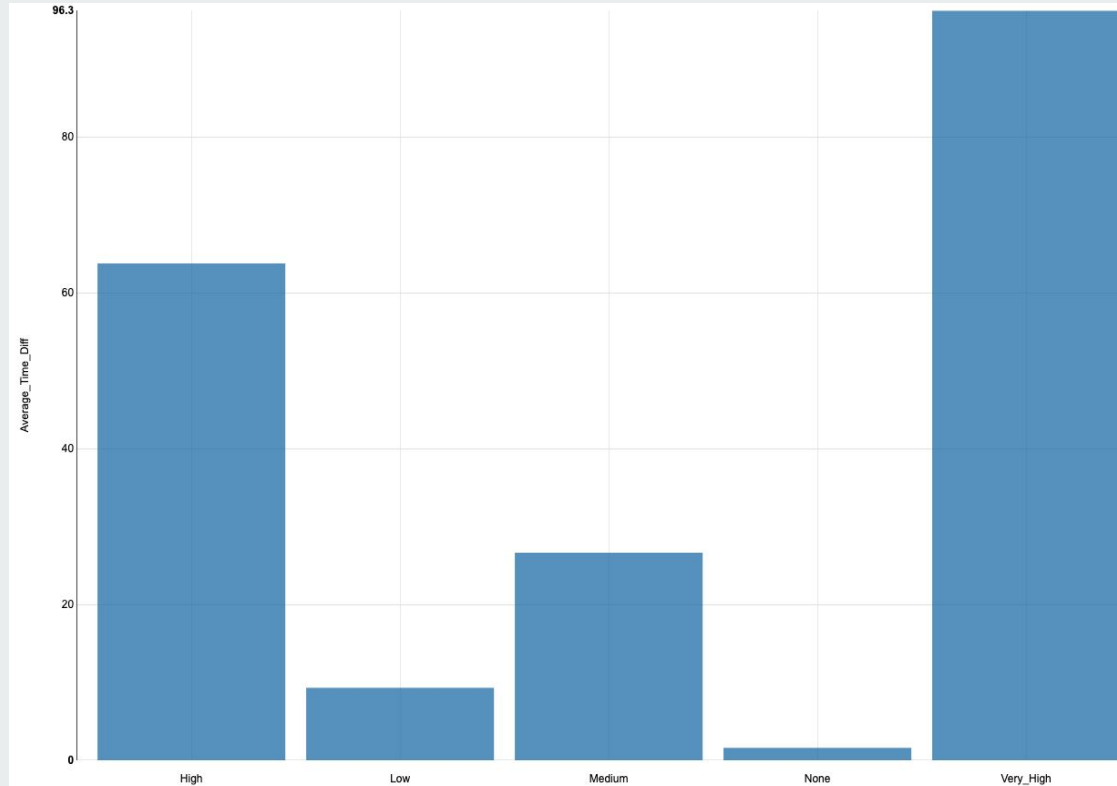
# Dataset de intensidade de trânsito



## AVERAGE\_SPEED\_DIFF: distribuição



# AVERAGE\_SPEED\_DIFF & AVERAGE\_TIME\_DIFF



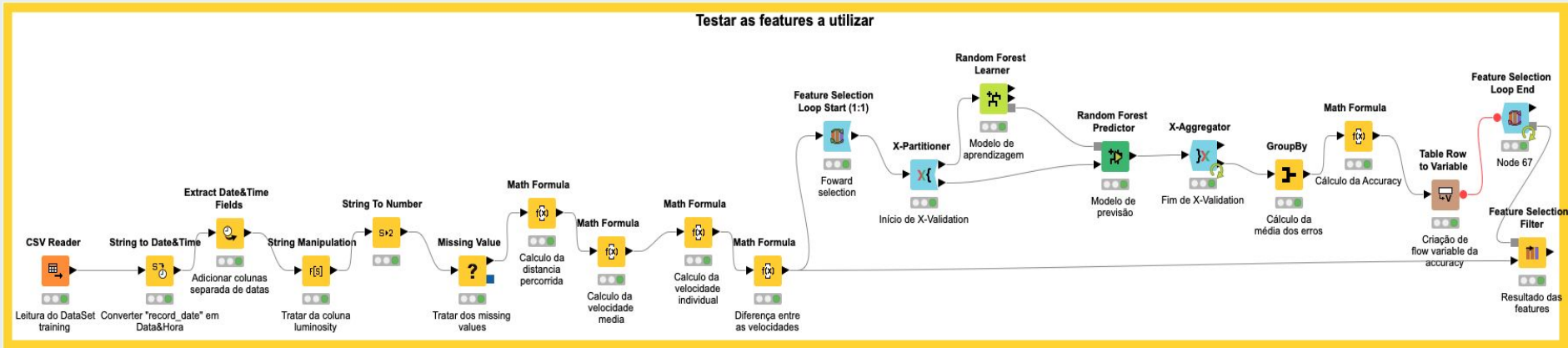
# Estatísticas do Dataset

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation
+ AVERAGE_FREE_FLOW_SPEED	<input type="checkbox"/>	30.500	55.900	40.661	4.119
+ AVERAGE_TIME_DIFF	<input type="checkbox"/>	0	296.500	25.637	33.511
+ AVERAGE_FREE_FLOW_TIME	<input type="checkbox"/>	46.400	112	81.144	8.294
+ LUMINOSITY	<input type="checkbox"/>	0	2	0.562	0.570
+ AVERAGE_TEMPERATURE	<input type="checkbox"/>	0	35	16.193	5.163
+ AVERAGE_ATMOSP_PRESSURE	<input type="checkbox"/>	985	1033	1017.388	5.751
+ AVERAGE_HUMIDITY	<input type="checkbox"/>	14	100	80.084	18.239
+ AVERAGE_WIND_SPEED	<input type="checkbox"/>	0	14	3.059	2.138
+ AVERAGE_PRECIPITATION	<input type="checkbox"/>	0	0	0	0
+ Year	<input type="checkbox"/>	2018	2019	2018.600	0.490
+ Month (number)	<input type="checkbox"/>	1	12	7.089	2.948
+ Day of year	<input type="checkbox"/>	15	346	200.643	88.943
+ Day of week (number)	<input type="checkbox"/>	1	7	4.023	2.010
+ Hour	<input type="checkbox"/>	0	23	11.534	6.940



# *Workflows* no KNIME

# Feature selection



# Exemplo Feature Selection

Column Selection

Flow Variables

Memory Policy

☐ Include static columns

☒ Select features manually

☐ Select features automatically by score threshold

Prediction score threshold

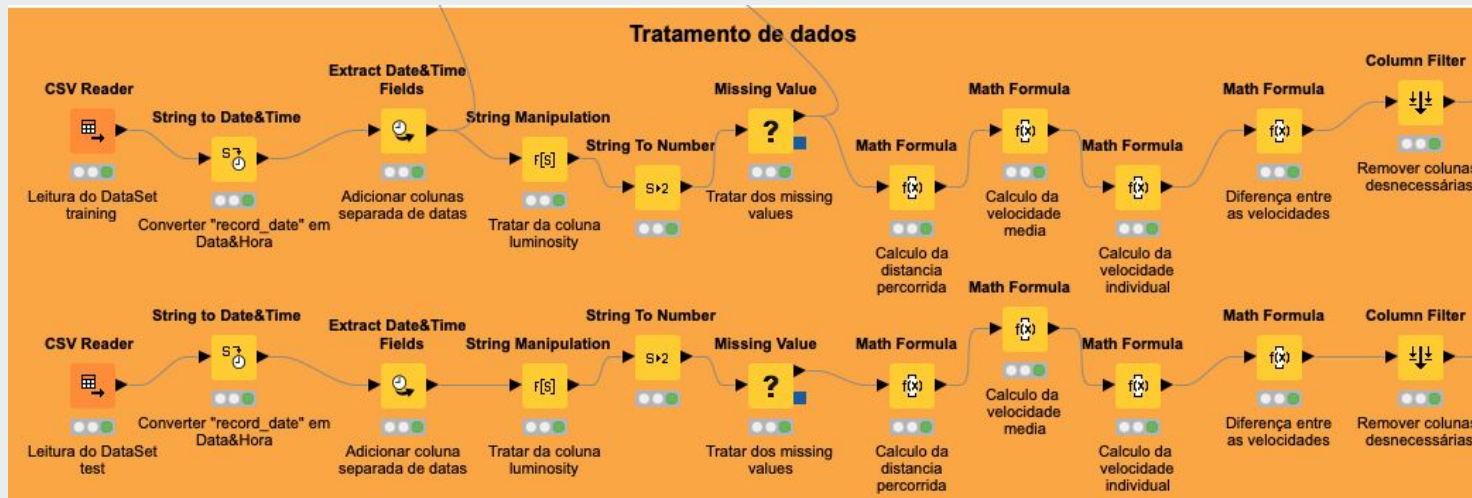
accuracy	Nr. of features	
80.623	16	S AVERAGE_SPEED_DIFF
80.226	17	D AVERAGE_FREE_FLOW_SPEED
80.137	17	D AVERAGE_TIME_DIFF
80.123	14	D AVERAGE_FREE_FLOW_TIME
78.92	11	I LUMINOSITY
78.641	10	D AVERAGE_TEMPERATURE
73.855	5	D AVERAGE_ATMOSP_PRESSURE
64.313	2	D AVERAGE_HUMIDITY
39.665	2	D AVERAGE_WIND_SPEED
32.296	1	S AVERAGE_CLOUDINESS
		D AVERAGE_PRECIPITATION
		S AVERAGE_RAIN
		I Year
		I Month (number)
		I Day of week (number)
		I Hour
		D DISTANCE
		D AVERAGE_SPEED_M/S
		D AVERAGE_INDIVIDUAL_SPEED_M/S
		D SPEED_DIFF_M/S



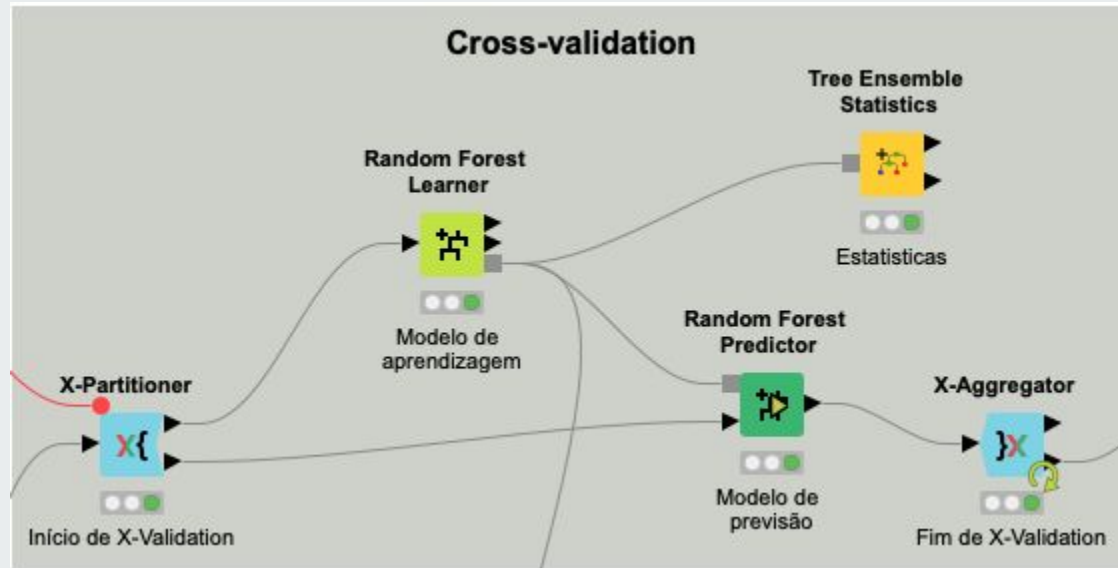
# Feature selection

Iteração	1	2	3	Intersecção
Accuracy	79,712%	80,241%	80.417%	
Average_Free_Flow_Speed		X	X	
Average_Free_Flow_Time	X	X	X	X
Luminosity	X	X	X	X
Average_Temperature	X	X	X	X
Average_Atmos_Pressure	X	X	X	X
Average_Humidity		X	X	
Average_Wind_Speed	X		X	
Average_Cloudiness	X	X	X	X
Average_Precipitation	X	X	X	X
Average_Rain	X		X	
Month(Number)	X	X	X	X
Day of Week(Number)		X	X	
Hour		X	X	
Distance		X	X	
Average_Speed_M/S	X		X	
Average_Individual_Speed_M/S	X	X	X	X
Speed_Diff_M/S	X	X	X	X
Average_Time_Diff		X		

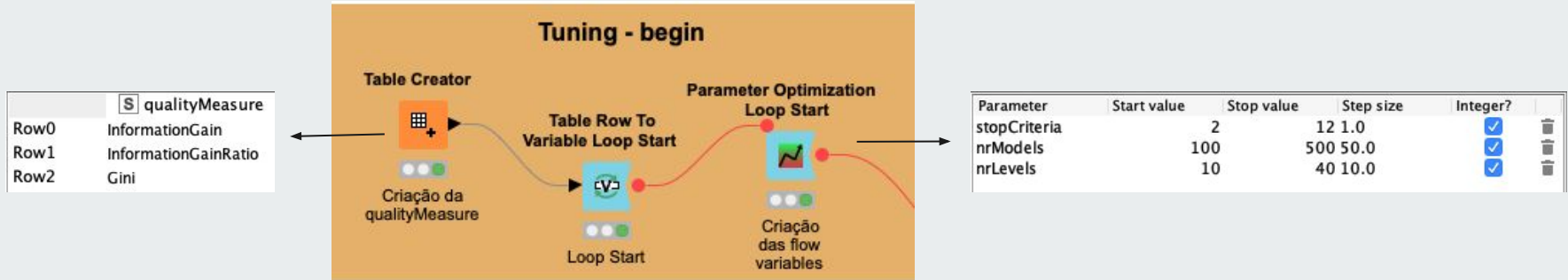
# Tratamento de dados



# Cross-Validation

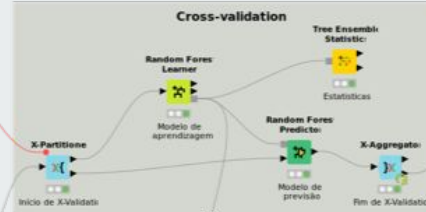
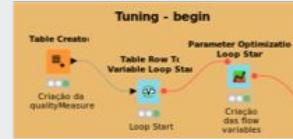
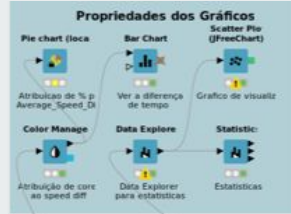


# Begin tuning



Row ID	I stopCriteria	I nrModels	I nrLevels	D Objec...	S RowID	I curren...	I maxIt...	S qualityMeasure
Row0	3	500	20	80.799	Best parameters	0	3	InformationGain
Row1	3	500	30	80.946	Best parameters	1	3	InformationGainRatio
Row2	5	300	20	80.608	Best parameters	2	3	Gini

# WorkFlow completo



# Kaggle

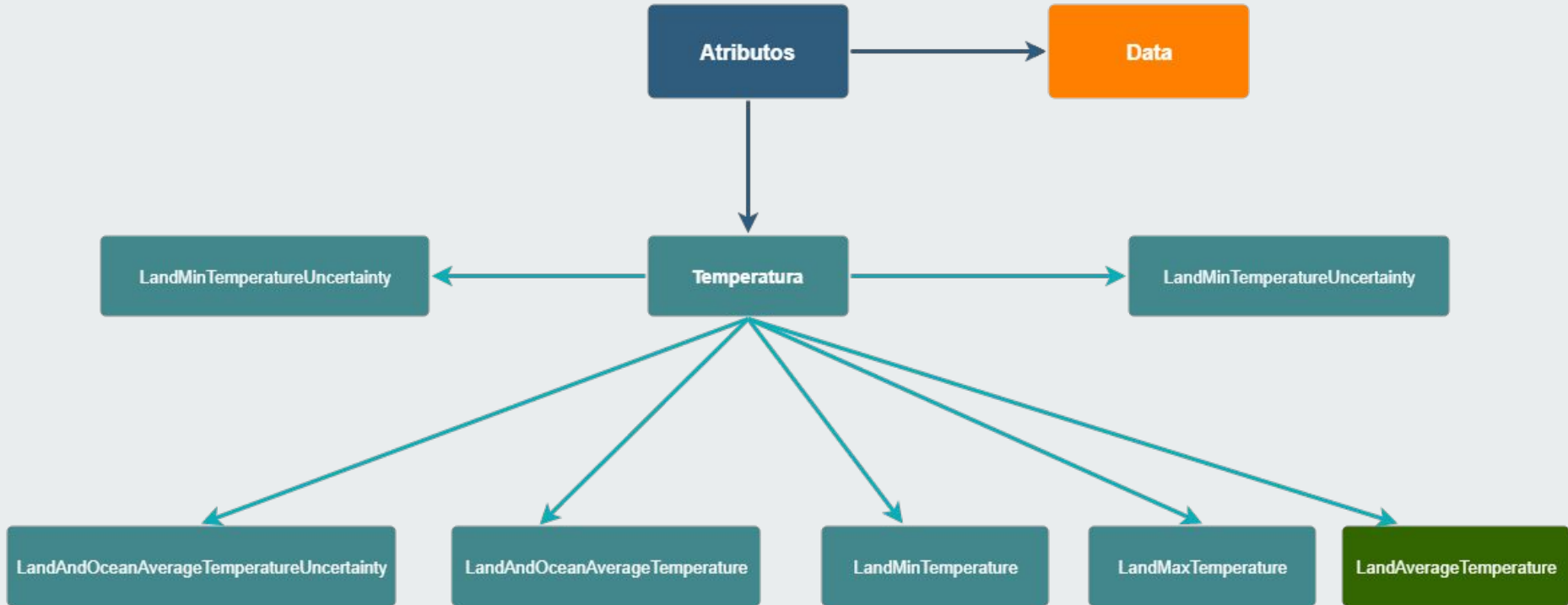


**Submissões Kaggle:**

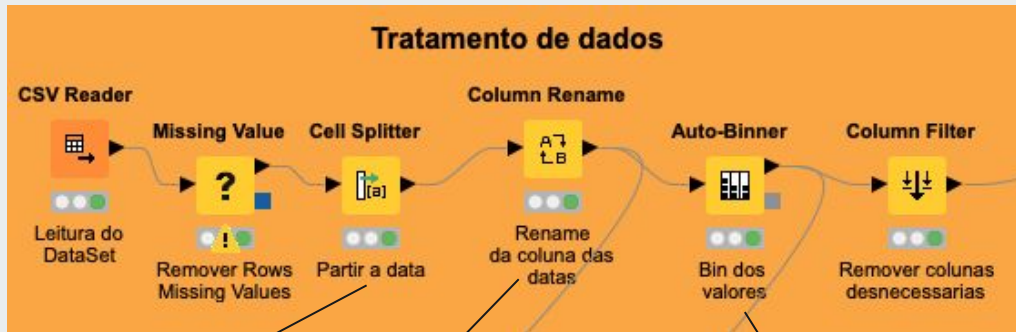
**Melhor Dataset: Público - 78,66%**  
**Privado - 82,19%**

**Dataset Submetido: Público - 81,11%**  
**Privado - 81,43%**

# Dataset da temperatura global



# Tratamento de dados



S	dt
	1850-01-01



Year	Month	Day
1850	1	1

**Binning Method**

☒ Fixed number of bins

Number of bins:

Equal:

☐ Sample quantiles

Quantiles (comma separated):

**Bin Naming**

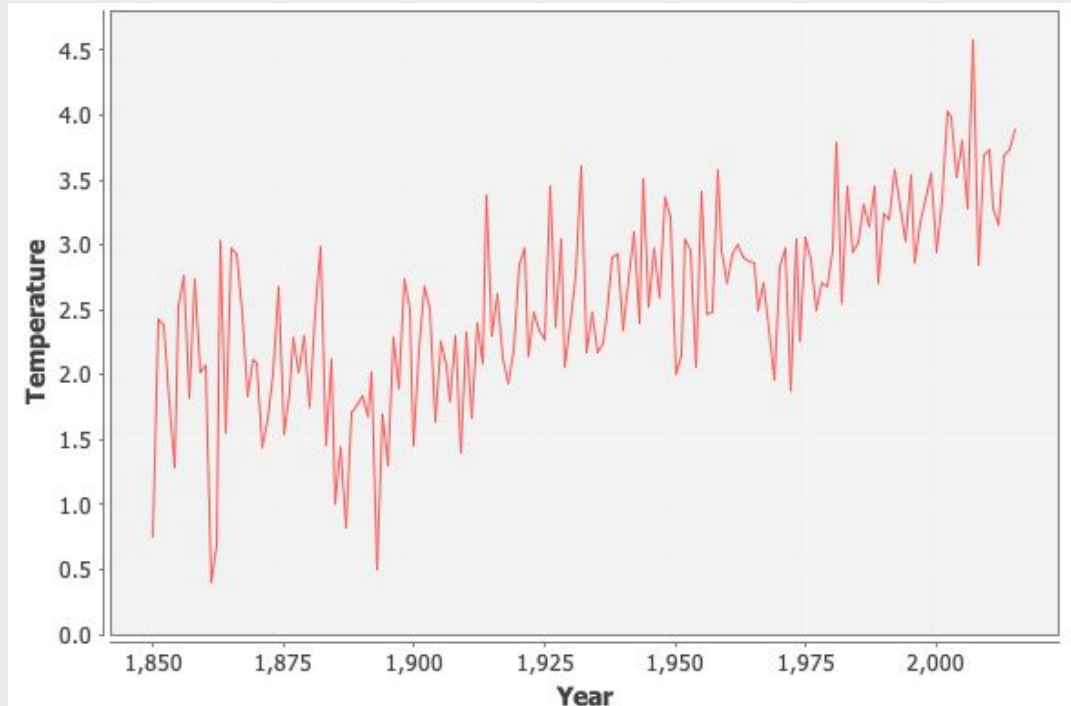
☐ Numbered e.g.: Bin 1, Bin 2, Bin 3

☒ Borders e.g.: [-10,0], (0,10], (10,20]

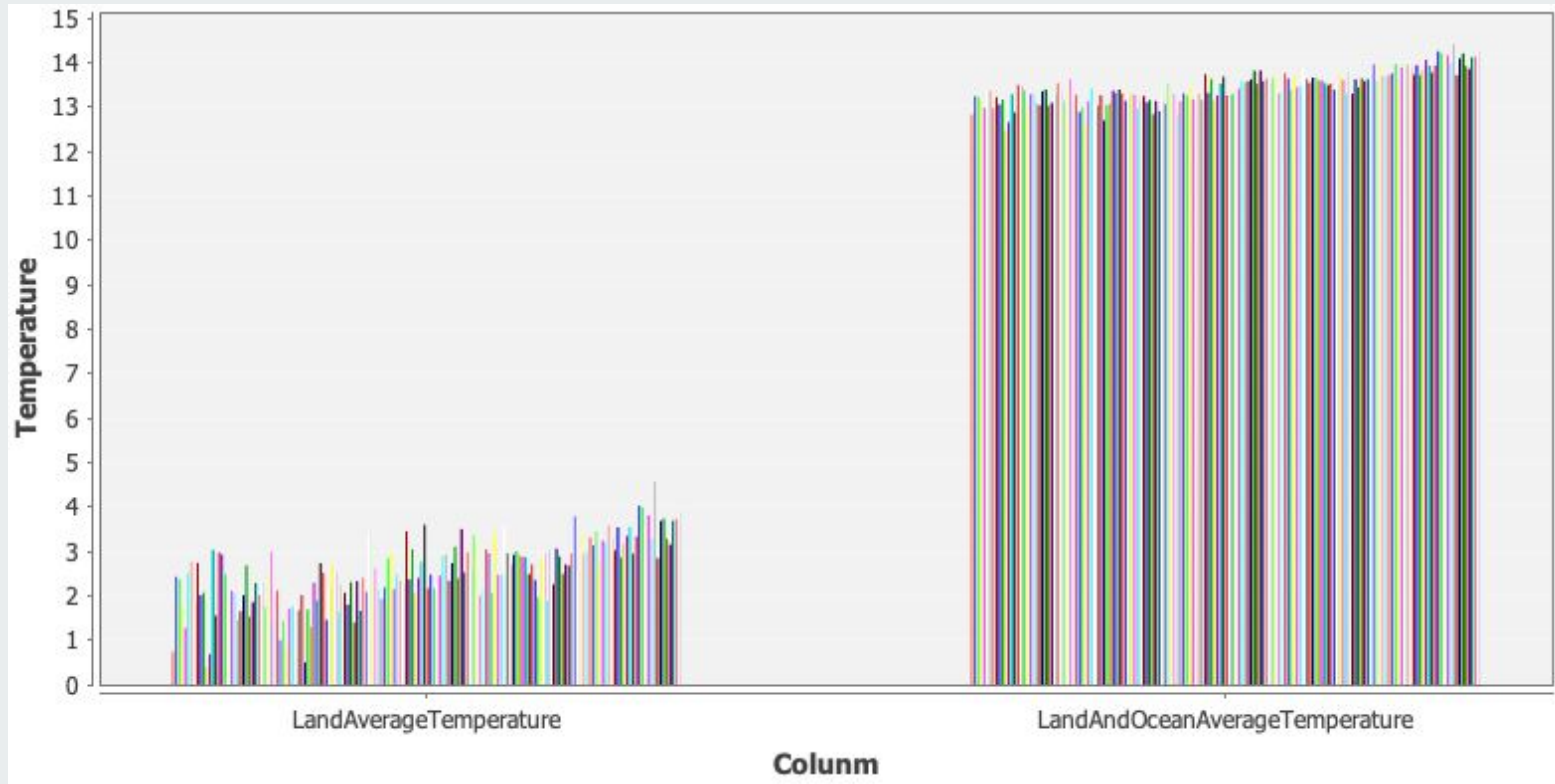
☐ Midpoints e.g.: -5, 5, 15




# AverageLandTemperature & Year



# Average LandTemperature & LandOceanTemperature

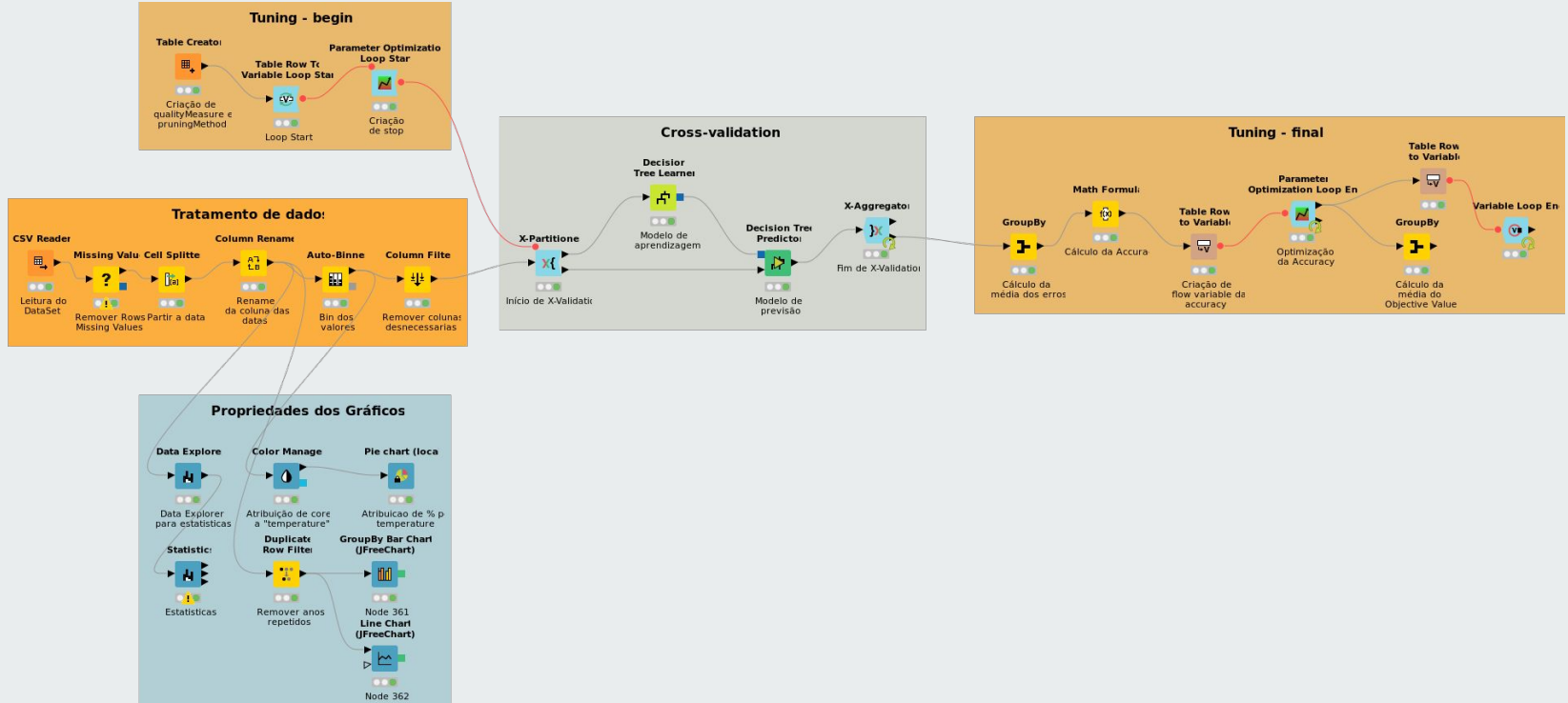


# Resultados Obtidos



Row ID	I stopCriteria	D Objective value	S RowID	I curren...	I maxlt...	S qualityMeasure	S pruningMethod
Row0	3	85.945	Best parameters	0	4	Gain ratio	No pruning
Row1	3	85.543	Best parameters	1	4	Gain ratio	MDL
Row2	4	87.399	Best parameters	2	4	Gini index	No pruning
Row3	2	86.095	Best parameters	3	4	Gini index	MDL

# WorkFlow completo





Universidade do Minho  
Departamento de Informática

# Árvores de Decisão

## Grupo 3

Sistemas Baseados em Similaridade

Universidade do Minho, Mestrado Integrado em Engenharia Informática,  
4º Ano, 1º Semestre, Novembro 2019