



Universidade do Minho

Relatório
Aprendizagem e Decisão Inteligentes
2023/2024

Grupo 35

Rui Cerqueira a100537

Diogo Cunha a100481

Guilherme Rio a100898

Tomás Valente 100540

Índice

1. Introdução	4
2. Metodologia Utilizada	4
3. Tarefa DatasetAtribuido	5
3.1 Estudo do negócio	5
3.2 Estudo dos dados	5
3.2.1 Category	6
3.2.2 Age	6
3.2.3 Gender	7
3.2.4 ALB (Albumina)	7
3.2.5 ALP (Fosfatase alcalina)	8
3.2.6 ALT (Alanina aminotransferase)	8
3.2.7 AST (Aspartato aminotransferase)	9
3.2.8 BIL (Bilirrubina)	9
3.2.9 CHE (Colinesterase)	9
3.2.10 CHOL (Colesterol)	10
3.2.11 CREA (Creatinina)	10
3.2.12 GGT (Gamaglutamiltransferase)	10
3.2.13 PROT (Proteína)	10
3.2.14 Outliers	11
3.2.15. Correlações	11
3.3 Preparação dos dados	12
3.4 Modelação	12
3.4.1 Modelação sem tratamentos adicionais	12
3.4.2 Modelação com tratamentos adicionais	13
3.4.3 Modelos de regressão	14
3.4.4 Clustering	15
3.4.5 Redes neuronais	15
3.5 Avaliação	16
4. Tarefa Dataset Grupo	17
4.1. Estudo do negócio	17
4.2. Estudo dos Dados	18
4.2.1. Cnt (número de alugueres)	18
4.2.2 Casual	19
4.2.3 Registered	19
4.2.4 Season	19
4.2.5 Holiday	20

4.2.6 Weekday	20
4.2.7 Working day	20
4.2.8 Weathersit	21
4.2.9 Temp	21
4.2.10 Atemp	22
4.2.11 Hum.....	22
4.2.12 Windspeed.....	23
4.3 Preparação dos dados	23
4.4. Modelação.....	24
4.4.1. Modelação inicial	24
4.4.2. Modelação com tratamento de outliers	25
4.4.3. Modelação com feature selection	25
4.4.4. Modelação com redes neuronais.....	26
4.5. Avaliação	28
5. Conclusão	28

1. Introdução

No âmbito da disciplina de Aprendizagem e Decisão Inteligentes, foi proposto como projeto a conceção de modelos de aprendizagem. O projeto foi dividido em duas tarefas, a Tarefa DatasetGrupo, que consiste na consulta, análise e exploração de um dataset escolhido pelo grupo e a Tarefa DatasetAtribuído que consiste na exploração e análise de um dataset atribuído pelos docentes.

2. Metodologia Utilizada

A metodologia que escolhemos utilizar para ambas as tarefas deste projeto é o CRISP-DM, este é um modelo de processos com vista a definir um “guião” para o desenvolvimento de projetos de análise de dados, que se desenrola em 6 etapas:

- **Estudo do negócio**
- **Estudo dos dados**
- **Preparação dos dados**
- **Modelação**
- **Avaliação**
- **Desenvolvimento**

Vamos abordar estas etapas nas duas tarefas do projeto e o que foi desenvolvido para cada uma delas.

3. Tarefa DatasetAtribuido

Para esta tarefa, foi-nos atribuído o dataset para grupos com número ímpar, sobre pacientes, mais especificamente doadores de sangue e pacientes com hepatite c, com dados sobre as suas amostras de sangue, idade, sexo, entre outros.

3.1 Estudo do negócio

O propósito desta tarefa consiste em desenvolver um modelo capaz de prever a categoria a que pertence cada paciente, com base nos dados disponíveis de idade, valores das amostras de sangue, etc.

Objetivos do problema:

- Analisar o dataset
- Corrigir as inconsistências nos dados
- Utilizar gráficos para visualizar os dados e as suas relações
- Aplicar vários modelos de aprendizagem

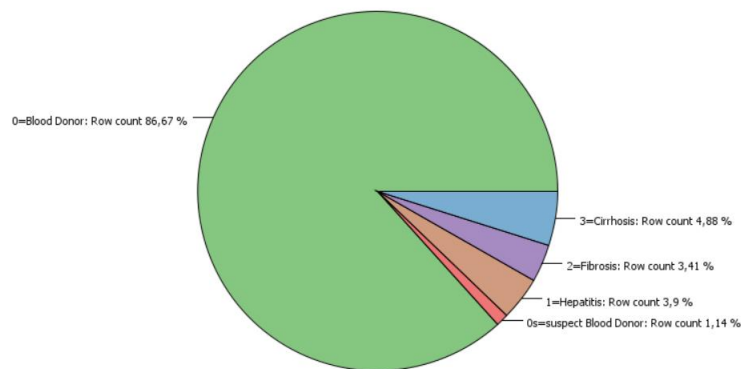
3.2 Estudo dos dados

O dataset contém os seguintes atributos:

1. **ID:** ID do paciente
2. **Age:** Idade
3. **Birth year:** Ano de nascimento
4. **Birth month:** Mês de nascimento
5. **Birth day:** Dia de nascimento
6. **Sex:** Sexo
7. **Birth location:** Local de Nascimento
8. **ALB:** Albumina
9. **ALP:** Fosfatase alcalina
10. **ALT:** Alanina aminotransferase
11. **AST:** Aspartato aminotransferase
12. **BIL:** Bilirrubina
13. **CHE:** Colinesterase
14. **CHOL:** Colesterol
15. **CREA:** Creatinina
16. **GGT:** Gamaglutamiltransferase
17. **PROT:** Proteína
18. **Category:** Categoria do paciente (0=Blood Donor, 0s=suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, 3=Cirrhosis)

3.2.1 Category

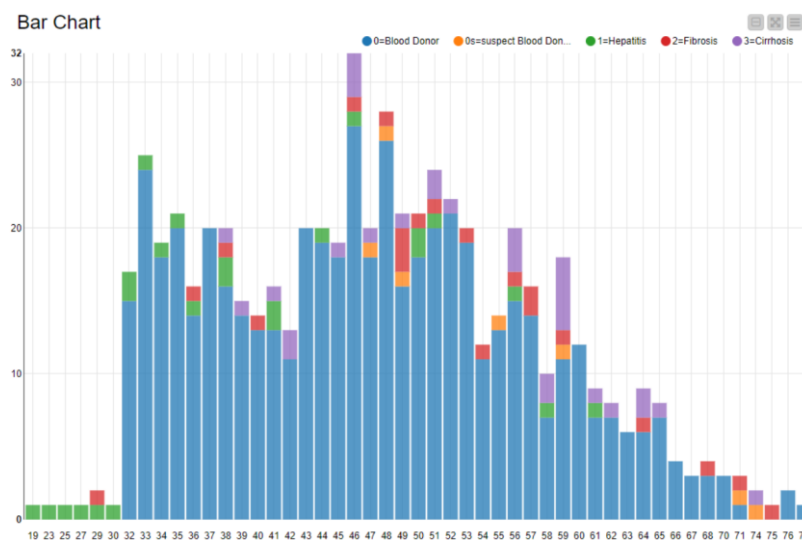
Começamos por analisar a categoria alvo a qual está dividida em 5 valores, sendo estes **Blood Donor**, **Suspect Blood Donor**, **Hepatitis**, **Fibrosis** e **Cirrhosis**. De modo a visualizar a distribuição das categorias utilizamos um pie chart e visualizamos que as amostras de doadores de sangue representam 87% das categorias, ou seja, que este atributo é bastante desequilibrado.



1. Distribuição Categorias

3.2.2 Age

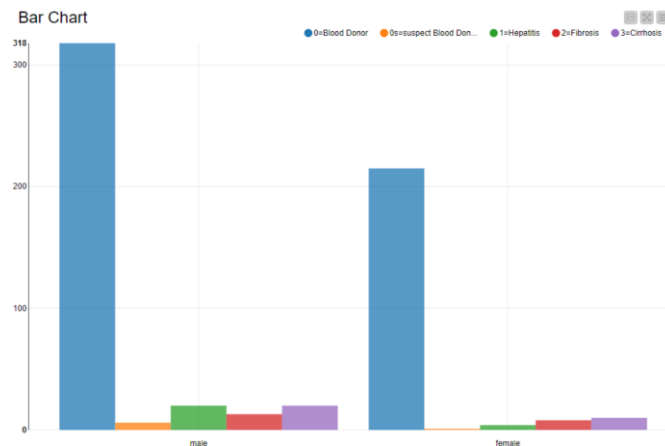
Depois de analisar o gráfico, verificamos que as idades no dataset variam entre 19 e 77 anos. Observa-se também que a hepatite pode afetar pessoas de todas as idades, enquanto a cirrose é mais comum entre os indivíduos mais idosos, indicando que esta última é uma doença que tende a se desenvolver com o envelhecimento.



2. Distribuição das idades/categorias

3.2.3 Gender

Começamos por verificar o número de homens e mulheres no dataset e verificamos que existem 377 homens e 238 mulheres, ou seja este atributo não está balanceado. Além disso observamos que o número de homens no dataset tem mais tendência a desenvolver hepatite.

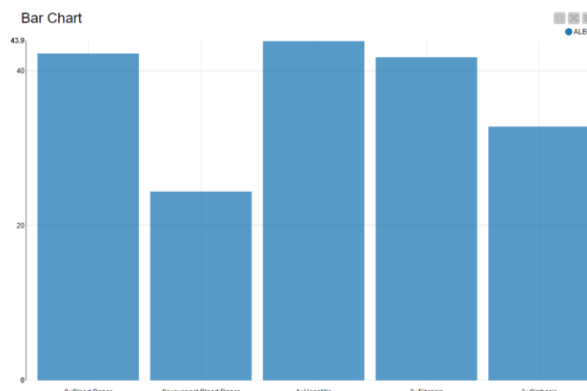


3. Distribuição Gênero/Categorias

Vamos agora analisar as amostras de sangue, de modo a fazer esta análise decidimos utilizar a média dos valores das amostras.

3.2.4 ALB (Albumina)

Após analisar o gráfico verificamos que os valores de ALB são constantes em pacientes doadores de sangue assim como pacientes de hepatite, começando a diminuir nos casos de pacientes que desenvolveram cirrose. O nível desta amostra nos pacientes suspeitos de portarem a doença pode ser devido à escassez de dados.

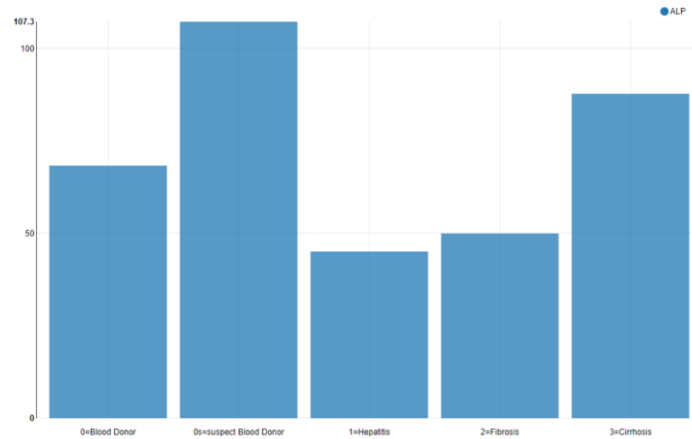


4. Média de valores de ALB/Categoria

Blood Donor: 42.2 | SBlood Donor: 24.4 | Hepatitis: 43.8 | Fibrosis: 41.7 | Cirrhosis: 41.8

3.2.5 ALP (Fosfatase alcalina)

Nesta amostra conseguimos observar que os valores são mais baixos nos casos de Hepatite e Fibrose, e mais altos quando se desenvolve para Cirrose. Os valores nos casos dos doadores de sangue suspeitos são também inesperadamente elevados, possivelmente devido à falta de dados.

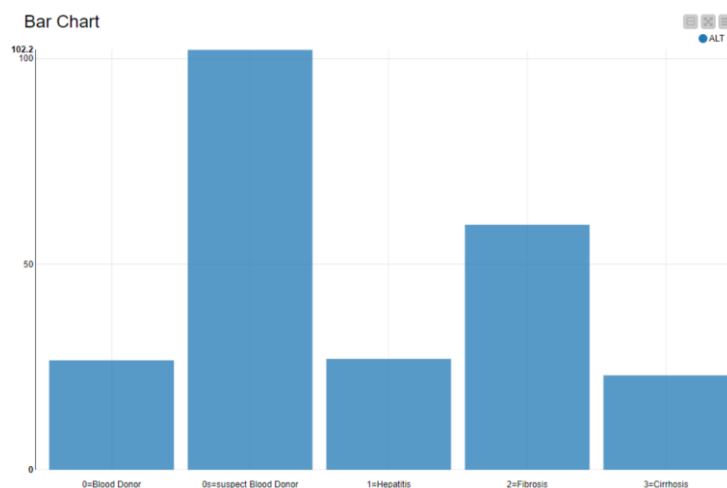


5. Média de valores de ALP/Categoria

Blood Donor: 68.3 | SBlood Donor: 107.3 | Hepatitis: 45 | Fibrosis: 50 | Cirrhosis: 87.8

3.2.6 ALT (Alanina aminotransferase)

Esta amostra demonstra um pico no caso de o paciente ser suspeito portador de doença, assim como no caso da Fibrose, e mantêm níveis similares nos restantes casos.

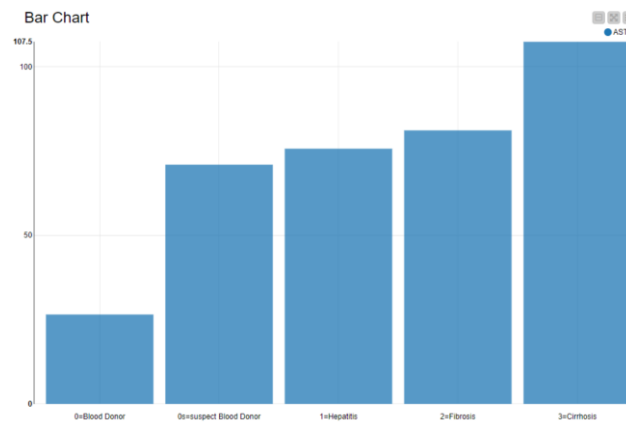


6. Média de valores de ALT/Categoria

Blood Donor: 26.6 | SBlood Donor: 102.1 | Hepatitis: 26.7 | Fibrosis: 59.6 | Cirrhosis: 23

3.2.7 AST (Aspartato aminotransferase)

Podemos observar nesta amostra, que os valores de AST tendem a aumentar ao se desenvolver a doença, sendo bastante baixo quando comparado com amostras de doadores de sangue.

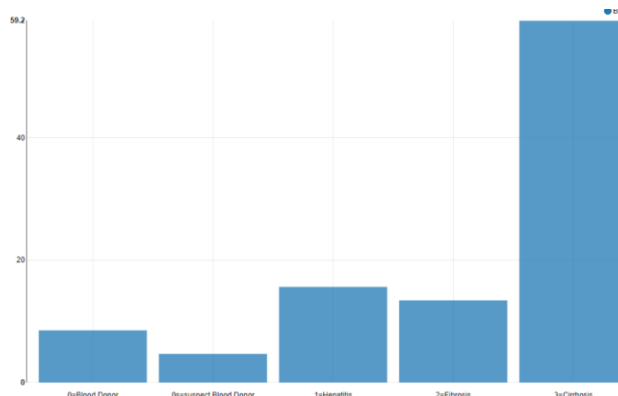


7. Média de valores de AST/Categoria

Blood Donor: 26.5 | SBlood Donor: 71 | Hepatitis: 75.7 | Fibrosis: 81.1 | Cirrhosis: 107.5

3.2.8 BIL (Bilirrubina)

Esta é outra amostra onde os valores são maiores quando se tem hepatite, havendo um pico nos casos em que se desenvolve para cirrose. Os suspect blood donors apresentam um valor baixo relativamente aos restantes.



8. Média de valores de BIL/Categoria

Blood Donor: 8.5 | SBlood Donor: 4.7 | Hepatitis: 15 | Fibrosis: 13.4 | Cirrhosis: 3.8

3.2.9 CHE (Colinesterase)

Os valores desta amostra no sangue parecem ser bastante similares em todas as categorias, excluindo nos pacientes com cirrose, onde os valores parecem ser bastante mais baixos.

Blood Donor: 8.4 | SBlood Donor: 7.5 | Hepatitis: 9.3 | Fibrosis: 8.3 | Cirrhosis: 3.8

3.2.10 CHOL (Cholesterol)

Neste caso parece que os valores das amostras são maiores no caso de pacientes doadores de sangue, e que os mesmos diminuem ao longo que a doença se vai desenvolvendo.

Blood Donor: 5.5 | **SBlood Donor:** 4.4 | **Hepatitis:** 5.1 | **Fibrosis:** 4.6 | **Cirrhosis:** 4.1

3.2.11 CREA (Creatinina)

Os valores desta amostra são similares em todas as categorias exceto no caso de pacientes com cirrose, onde o valor desta amostra é cerca de duas vezes maior que nos restantes pacientes.

Blood Donor: 79 | **SBlood Donor:** 61.7 | **Hepatitis:** 74 | **Fibrosis:** 73.4 | **Cirrhosis:** 138.2

3.2.12 GGT (Gamma-glutamyltransferase)

No dataset podemos observar que os pacientes portadores da doença têm um pico aquando comparado com os doadores de sangue assim como os suspect blood donors.

Blood Donor: 29 | **SBlood Donor:** 151.5 | **Hepatitis:** 92.6 | **Fibrosis:** 79.5 | **Cirrhosis:** 129

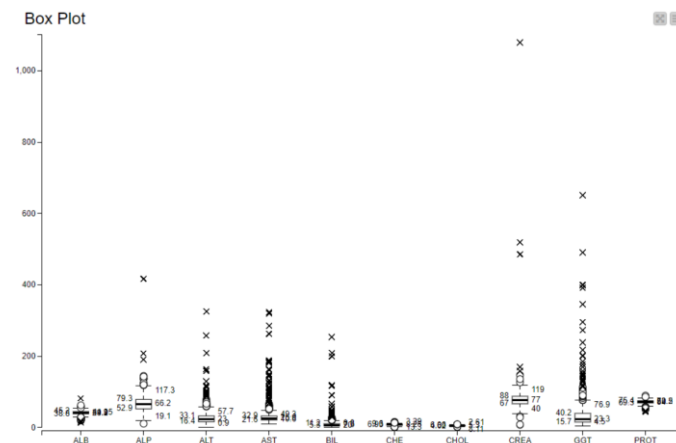
3.2.13 PROT (Proteína)

Nesta amostra os valores são bastante equilibrados em todos os casos exceto nos suspect blood donors, em que o valor da amostra é mais baixo quando comparado com os restantes.

Blood Donor: 72.1 | **SBlood Donor:** 53.9 | **Hepatitis:** 74.7 | **Fibrosis:** 76.1 | **Cirrhosis:** 70

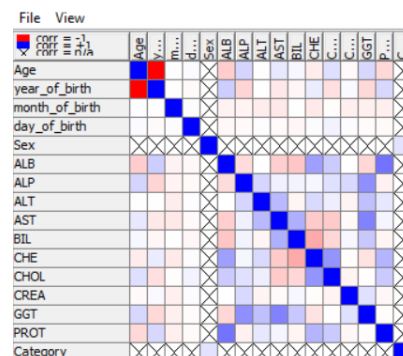
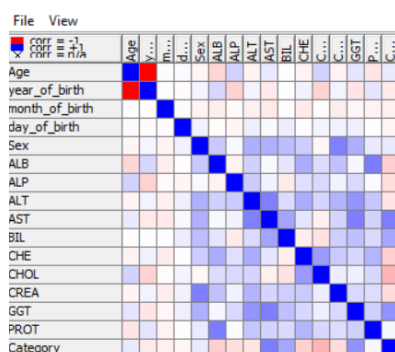
3.2.14 Outliers

De modo a verificar se existem outliers nos dados utilizamos o nodo de Box Plot, no qual podemos ver que existem bastantes outliers em vários dos dados das amostras que terão de ser tratados posteriormente. Destacam-se os outliers existentes na coluna da Creatinina por serem muito acima da norma, e os de GGT por serem também elevados.



9. Outliers nas amostras de sangue

3.2.15. Correlações



10. Rank/Liner correlation

Através de uma análise as matrizes de correlação podemos observar uma correlação negativa entre o ano de nascimento e a idade, que era esperada devido à sua óbvia relação. No que toca à correlação existente com o atributo objetivo, existem três atributos que apresentam uma correlação positiva mais destacada. A maior acontece com o AST (0,511), seguindo-se do CGT (0,427) e do BIL (0,349). Isto indica-nos que estes têm maior influência na categoria a que os pacientes pertencem. Para a correlação negativa destaca-se o CHO (-0,289).

3.3 Preparação dos dados

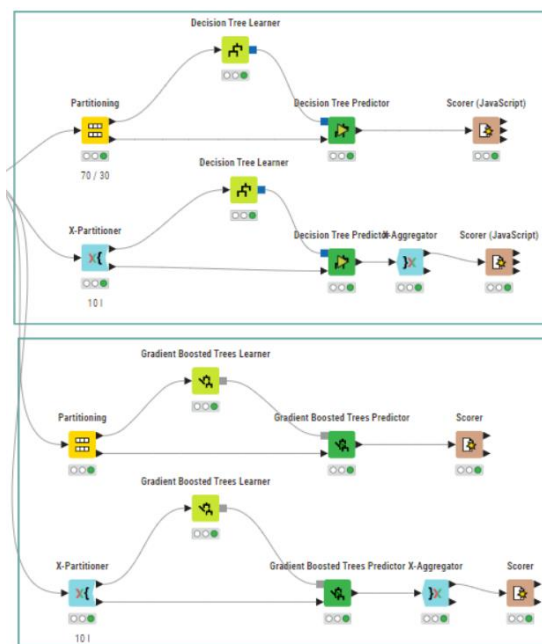
Nesta fase começamos por analisar os atributos para efetuar as correções necessárias para estes estarem prontos a ser utilizados.

- **Sex** - encontramos no dataset linhas com valores como “mm” em vez de “m” para representar homens. Em vez de remover estas linhas decidimos fazer o tratamento do atributo e transformar todas as linhas (tanto com “mm” ou “m”) no valor “male”.
- **Location**: ao analisar este atributo, observamos que apenas existiam 2 valores, “New Delhi” ou nulo. Tomamos então a decisão de remover este atributo para análise, pois não teria impacto ao fazer os modelos para os dados.
- **CHOL, ALP** – nestes atributos encontramos missing values identificados por NA. Tivemos então de utilizar um nodo para transformar estes em valores nulos e fizemos a transformação das colunas para doubles.
- **Missing values**: para tratar os valores nulos decidimos testar três configurações diferentes: remover os valores, utilizar a média dos valores e utilizar a mediana. Após a realização desses testes observamos que ao utilizar a mediana obtivamos os melhores resultados e utilizamos esse método para o tratamento dos mesmos.

3.4 Modelação

3.4.1 Modelação sem tratamentos adicionais

Na modelação de Machine Learning começamos por um modelo simples, sem qualquer tratamento adicional.



10. Modelação sem tratamentos adicionais

Tendo em conta que isto se trata de um problema de classificação, utilizamos nodos de árvores de decisão com os algoritmos de aprendizagem Decision Tree e Gradient Boosting. Escolhemos também utilizar cross-validation para correr os algoritmos várias vezes, assim como hold-out validation.

Confusion Matrix - 4:22:48:126:91 - Scorer

Category \...	0=Blood D...	0s=suspec...	1=Hepatitis	2=Fibrosis	3=Cirrhosis
0=Blood Donor	159	0	2	1	1
0s=suspect ...	0	1	0	0	0
1=Hepatitis	1	0	4	0	0
2=Fibrosis	1	0	4	3	0
3=Cirrhosis	1	0	0	1	6

Correct classified: 173
Accuracy: 93,514%
Cohen's kappa (κ): 0,711%

Wrong classified: 12
Error: 6,486%

11. Scorer: Modelação STA Gradient Boosted

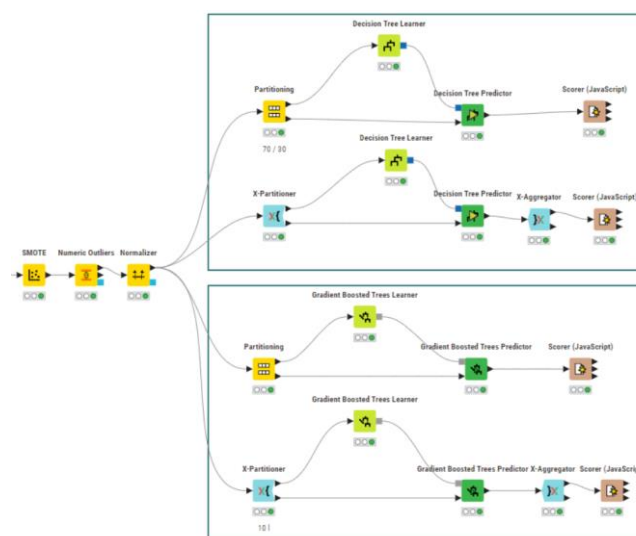
Este foi o resultado obtido pelo algoritmo Gradient Boosting com partitioning básico. Este foi o que obteve o melhor resultado, com uma accuracy de 93,514%.

3.4.2 Modelação com tratamentos adicionais

Testamos os modelos com hold-out validation, assim como cross validation, utilizando vários tipos de tratamento. Para isso foram usados os seguintes nodos:

- Normalizer: para normalizar os dados das amostras de sangue.
- SMOTE: utilizado para equilibrar o dataset.
- Numeric Outliers: para tratamento dos outliers que se encontram nas amostras de sangue. Utilizamos group selection sendo o atributo alvo as categorias.

Após realizar vários testes com os diferentes modelos, entre os quais tratamento com normalização, tratamento de outliers e injeção de dados artificiais, o modelo que utilizou todos estes tratamentos listados acima foi o que obteve os melhores resultados.



13. Modelação com tratamentos adicionais

Entre os dois algoritmos usados, o Gradient Boosted foi o que obteve os melhores resultados, com uma precisão quase perfeita de 99,47%.

Scorer View

Confusion Matrix

	0=Blood Donor (...)	0s=suspect Blo...	1=Hepatitis (Pre...	2=Fibrosis (Pred...	3=Cirrhosis (Pre...	
0=Blood Donor (...)	529	0	3	1	0	99.25%
0s=suspect Blo...	0	533	0	0	0	100.00%
1=Hepatitis (Act...	4	0	527	2	0	98.87%
2=Fibrosis (Actu...	0	0	1	532	0	99.81%
3=Cirrhosis (Act...	0	0	0	3	530	99.44%
	99.25%	100.00%	99.25%	98.88%	100.00%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
99.47%	0.53%	0.993	2651	14

14. Resultados modelo Gradient Boosted

3.4.3 Modelos de regressão

Decidimos também realizar modelos de regressão. Para isso transformamos o atributo alvo num valor numérico, onde aplicamos os algoritmos de **Simple Regression**, **Linear Regression** e **Gradient Boosted**.

Os resultados dos algoritmos utilizando cross validation e os tratamentos que obtiveram melhor resultados nos modelos anteriores foram os seguintes:

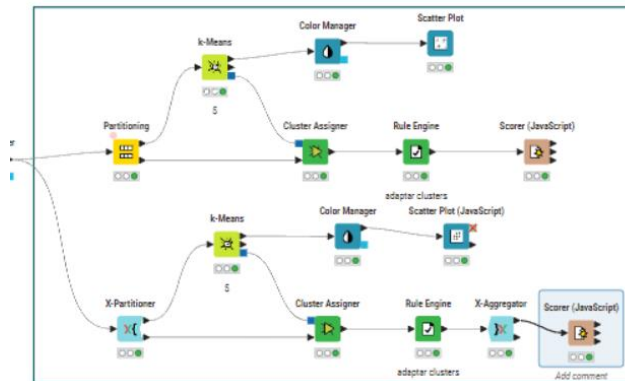
Statist... — □ × Statist... — □ × Statist... — □ ×

File	File	File
R²:	0,944	R²: 0,701 R²: 0,946
Mean absolute error:	0,064	Mean absolute error: 0,667 Mean absolute error: 0,101
Mean squared error:	0,249	Mean squared error: 1,32 Mean squared error: 0,238
Root mean squared error:	0,499	Root mean squared error: 1,149 Root mean squared error: 0,487
Mean signed difference:	-0,003	Mean signed difference: -0,001 Mean signed difference: -0,042
Mean absolute percentage error:	0,021	Mean absolute percentage error: 0,391 Mean absolute percentage error: 0,025
Adjusted R²:	0,944	Adjusted R²: 0,701 Adjusted R²: 0,946

15. Numeric Scorer: Modelos de regressão

3.4.4 Clustering

Ao realizar modelos com estratégia de clustering com os tratamentos dos dados efetuados, utilizamos o número de clusters igual a 5 sendo esse o número de categorias contidas no dataset, contudo não obtivemos bons resultados com este modelo.



16. Modelos de clustering

Scorer View

Confusion Matrix:

	0=Blood Donor (...)	0s=suspect Blo...	1=Hepatitis (Pre...	2=Fibrosis (Pred...	3=Cirrhosis (Pre...	
0=Blood Donor (...)	449	475	446	229	0	28.08%
0s=suspect Blo...	3	3	2	6	7	14.29%
1=Hepatitis (Act...	22	11	15	17	7	20.83%
2=Fibrosis (Actu...	9	12	8	21	13	33.33%
3=Cirrhosis (Act...	9	7	2	14	58	64.44%
	91.26%	0.59%	3.17%	7.32%	68.24%	

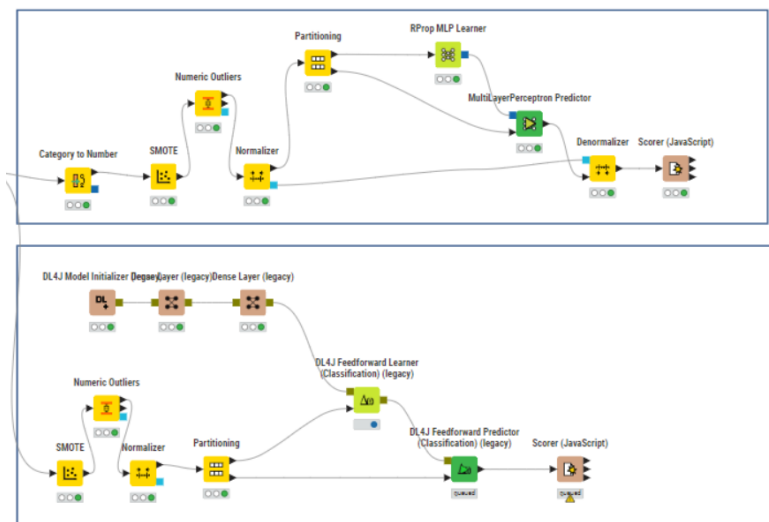
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
29.59%	70.41%	0.059	546	1299

17. Resultados clustering

3.4.5 Redes neuronais

Foi também feito modelo para rede neuronal, onde utilizamos o RProp MLP com o algoritmo MultiLayer Feedforward Networks que utiliza uma certa camada de neurónios, assim como o nodo DL4J para classificação.



18. Modelação com redes neuronais

De entre estes dois modelos, o que obteve melhores resultados foi o RProp MLP Learner onde utilizamos 250 iterações, 3 layers e 3 neurónios por layer, com os seguintes resultados:

Scorer View

Confusion Matrix

	0=Blood Donor (...)	0s=suspect Blo...	1=Hepatitis (Pre...	2=Fibrosis (Pred...	3=Cirrhosis (Pre...	
0=Blood Donor (...)	484	0	1	1	0	99.59%
0s=suspect Blo...	2	1	0	0	1	25.00%
1=Hepatitis (Act...	6	0	11	2	1	55.00%
2=Fibrosis (Actu...	3	0	8	10	2	43.48%
3=Cirrhosis (Act...	0	0	0	0	21	100.00%
	97.78%	100.00%	55.00%	76.92%	84.00%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
95.13%	4.87%	0.770	527	27

19. Scorer: redes neuronais

3.5 Avaliação

Neste dataset tivemos uma fase de tratamento de dados um pouco complexa, tendo sido necessário realizar vários tipos de tratamentos, onde observamos também que os dados deste dataset são bastante desequilibrados. Esse problema foi resolvido em vários modelos através do uso do nodo SMOTE.

Na modelação construímos vários modelos utilizando diferentes estratégias para cada um, de forma a descobrir o melhor modelo de previsão possível. Acabamos por concluir que os algoritmos com melhores resultados seriam os de árvores de decisão, aplicando-os a todos os atributos à exceção dos que representam a data de nascimento (dia, mês e ano de nascimento), que verificaram ser obsoletos uma vez que a idade era já apresentada como atributo.

De seguida resolvemos adaptar os dados de forma a transformar este problema num problema de regressão. Depois usamos algoritmos de regressão e conseguimos chegamos à conclusão de que esta é também uma forma adequada de resolver o problema.

A estratégia de clustering acabou por se revelar muito pior do que a original, e com isso podemos ver que esta estratégia não é indicada para esta tarefa.

E para finalizar a modelação deste problema resolvemos explorar os algoritmos de redes neuronais. Ambos os modelos criados obtiveram resultados razoáveis, mas inferiores aquele obtido com Gradient Boosted Trees, que entre todos os modelos criados, foi o algoritmo que conseguiu os melhores resultados, com uma precisão quase perfeita.

4. Tarefa Dataset Grupo

Na tarefa Dataset Grupo foi proposto que o grupo de trabalho escolhesse um dataset, de forma a analisá-lo e extrair o conhecimento relevante ao contexto do problema em questão. Nesta tarefa exploramos vários datasets de forma a encontrar aquele que mais correspondia às nossas expectativas. Acabamos por escolher o dataset Bike_Sharing, uma vez que apesar de na fase de tratamento de dados não necessitar de grande desenvolvimento, é aquele que nos oferece uma maior diversidade na exploração de outros tipos de modelação aqueles que a tarefa com o dataset atribuído nos dá.

O dataset escolhido contém a contagem por hora dos alugueres de bicicletas entre os anos de 2011 e 2012 no sistema da Capital Bikeshare com a correspondente informação meteorológica e sazonal.

4.1. Estudo do negócio

O objetivo desta tarefa consiste em desenvolver modelos capazes de prever o número de alugueres de bicicletas numa certa hora com os dados fornecidos pelo dataset. Estes dados contêm informação acerca das condições temporais e acerca dos dias, como se estes são feriados, dias úteis, etc.

Objetivos:

- Analisar o dataset
- Corrigir as inconsistências nos dados
- Utilizar gráficos para visualizar os dados e as suas relações
- Aplicar vários modelos de aprendizagem

4.2. Estudo dos Dados

O dataset Bike_Sharing é composto por 16 colunas e 17379 entradas. Cada entrada corresponde a uma hora de um dia para um total de aproximadamente dois anos.



De seguida encontra-se uma lista com todos os atributos e uma breve explicação do seu significado:

1. **dteday** : data
2. **season** : época do ano (1:Primavera, 2:Verão, 3:Outono, 4:Inverno)
3. **yr** : ano do estudo (0: 2011, 1:2012)
4. **mnth** : mês (1 to 12)
5. **hr** : hora (0 to 23)
6. **holiday** : se corresponde a um feriado ou não (0: não, 1: sim).
7. **weekday** : dia da semana (0 a 6)
8. **workingday** : se corresponde a um dia de trabalho (0: não, 1: sim).
9. **weathersit** : Situação do tempo. 1 corresponde a céu limpo, com algumas nuvens ou nublado ligeiro, 2 corresponde a névoa e nublado, 3 a chuva ligeira ou chuva com tempestade ligeira e 4 corresponde a chuva pesada, tempestade ou neve.
10. **temp** : temperatura em Celsius normalizada. A normalização destes valores foi feita seguindo a fórmula $(t-t_{\min})/(t_{\max}-t_{\min})$, com $t_{\min}=-8$ e $t_{\max}=+39$.
11. **atemp**: temperatura aparente em Celsius normalizada. A normalização destes valores foi feita seguindo a fórmula $(t-t_{\min})/(t_{\max}-t_{\min})$, com $t_{\min}=-16$ e $t_{\max}=+50$.
12. **hum**: humidade normalizada. A normalização foi feita dividindo os valores por 100.
13. **windspeed**: Velocidade do vento normalizada. A normalização foi feita dividindo os valores por 67.
14. **casual**: número de utilizadores casuais.
15. **registered**: número de utilizadores registados.
16. **cnt**: número total de utilizadores que inclui os casuais e os registados.

Este conjunto de dados tem como alvo o número total de alugueres nesse momento. É este atributo que iremos prever através de vários modelos de regressão.

4.2.1. Cnt (número de alugueres)

Começamos por realizar uma análise a categoria alvo, a qual indica o número de bicicletas alugadas numa certa hora e a sua respetiva data. Utilizamos o nodo Data Explorer para visualizar o número mínimo e máximo de bicicletas alugadas, sendo respetivamente 1 e 977. Vimos também que o número médio de bicicletas alugadas era 189

 cnt		1	977	189.463
-----------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------	---	-----	---------

4.2.2 Casual

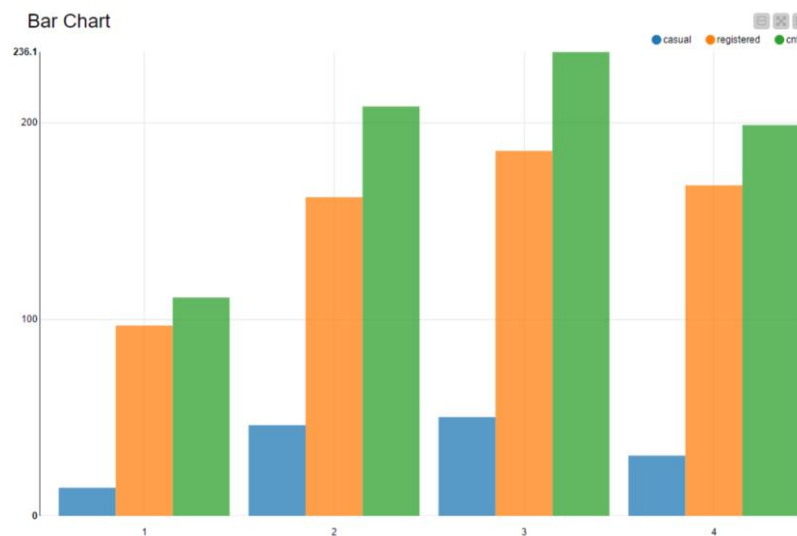
Análisamos também o atributo casual, que representa o número de alugueres feitos na respetiva hora por utilizadores “casuais” do sistema. Os valores mínimos e máximos para este atributo era respetivamente 0 e 367, sendo a média de alugueres por utilizadores casuais 36.

4.2.3 Registered

O atributo registered representa o número de alugueres na respetiva hora por utilizadores registados no sistema. Os valores mínimos e máximos para este atributo era respetivamente 0 e 886, sendo a média de alugueres por utilizadores registados 153.

4.2.4 Season

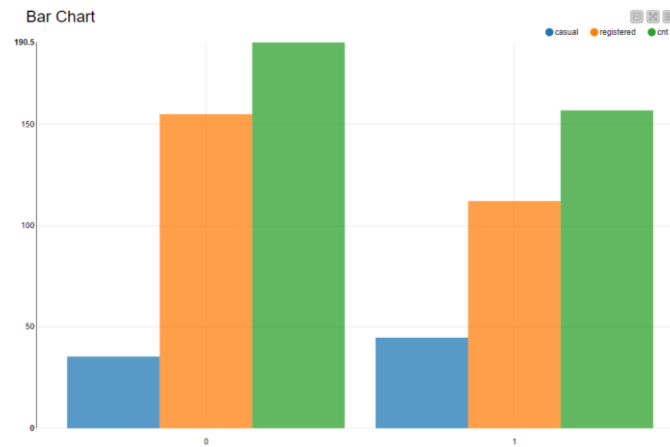
Através da análise deste atributo verificamos que se encontra bastante equilibrado, tendo iguais proporções das 4 épocas do ano. Podemos também visualizar que a época em que existem mais alugueres é o verão, e a que existem menos é o inverno.



21. Alugueres por época.

4.2.5 Holiday

Neste dataset existem 2.88% dos dados nos quais os dias são feriados. Verificamos também que em média são feitos mais alugueres em dias normais quando comparados com feriados.



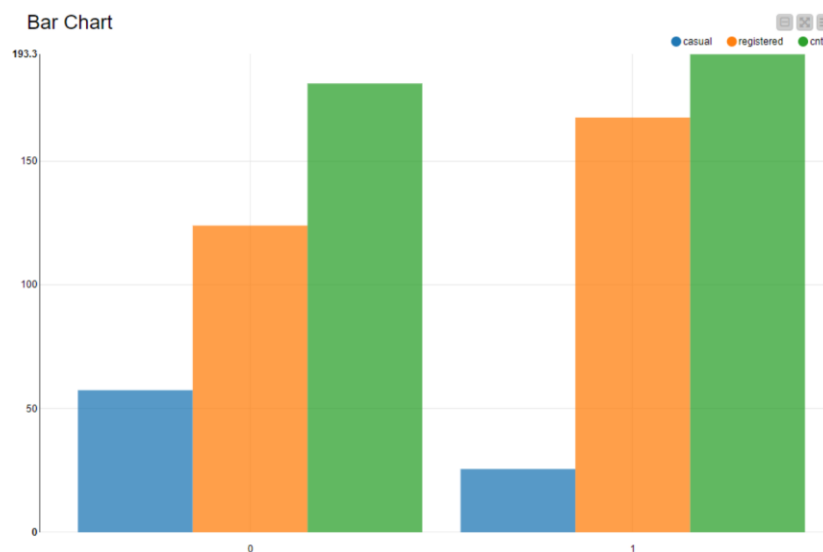
22. Alugueres em dias de férias/normais

4.2.6 Weekday

Este é o dado que nos indica o dia da semana, em que cada dia representa cerca de 14% dos dados. Todos os dias da semana tem números de alugueres bastante similares sendo o atributo equilibrado.

4.2.7 Working day

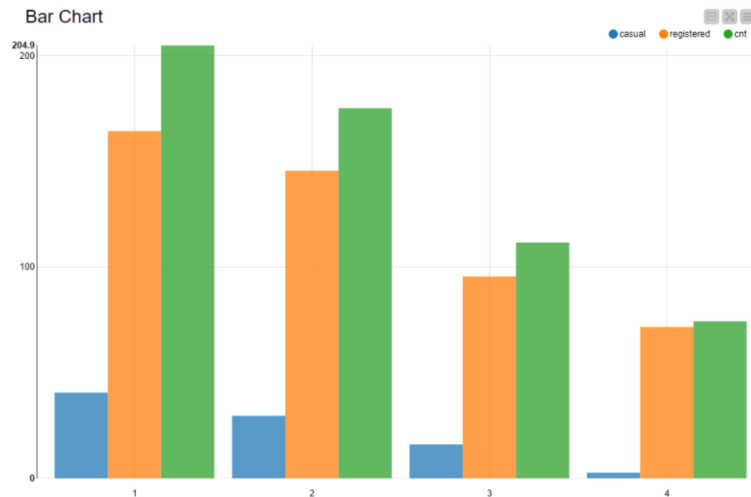
No dataset cerca de 68% dos dias representam dias de trabalhos, podemos visualizar que nesses dias existe um número de alugueres similar aos restantes.



23. Alugueres em dias de trabalho/restantes

4.2.8 Weathersit

Após uma análise a este atributo verificamos que cerca de 66% dos dados correspondiam a horas com o céu limpo, 26% com algumas nuvens, 8% nublado e apenas 0.017% a neve/chuva intensa. Podemos também ver que o número de alugueres diminui quando o tempo piora.

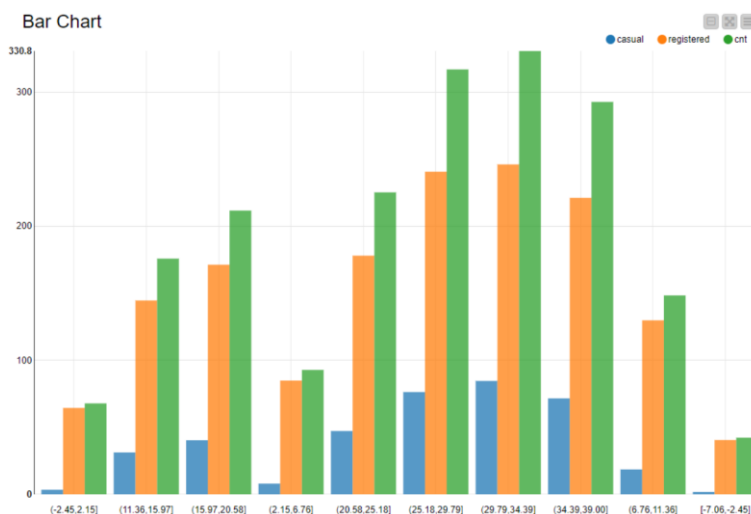


24. Alugueres em diversas condições meteorológicas

4.2.9 Temp

Ao analisar este dado, que estava normalizado no dataset, utilizamos a fórmula de normalização fornecida para desnormalizar os dados de modo a poder fazer uma melhor análise dos dados, e verificamos que o seu valor mínimo e máximo eram respetivamente -7 e 39, com uma média de 15.4.

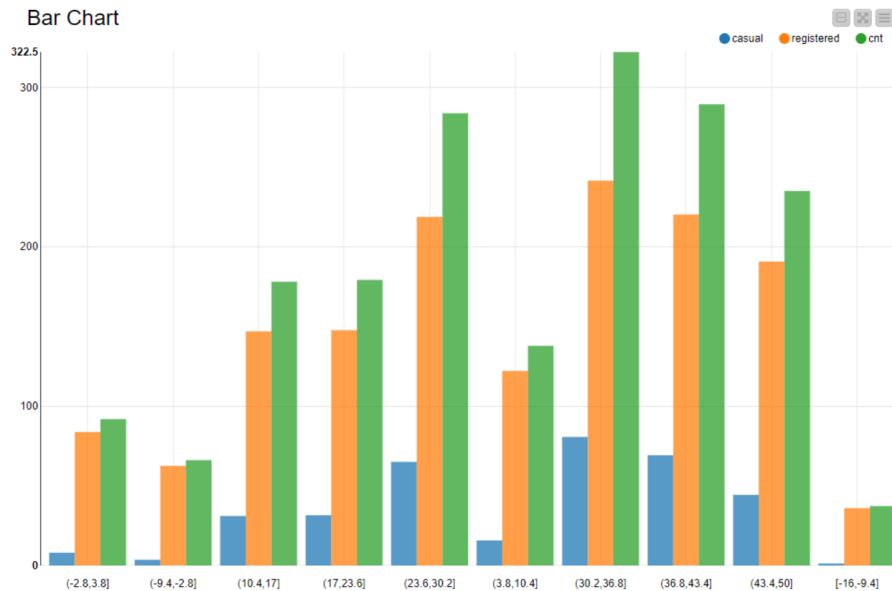
Utilizamos um auto-binner para dividir a temperatura em 10 grupos e visualizamos que o número de alugueres tende a ser menor com temperaturas muito baixas ou muito altas, havendo um pico de alugueres quando a temperatura se encontra entre os valores 25 e os 34.



25. Alugueres nas diferentes temperaturas

4.2.10 Atemp

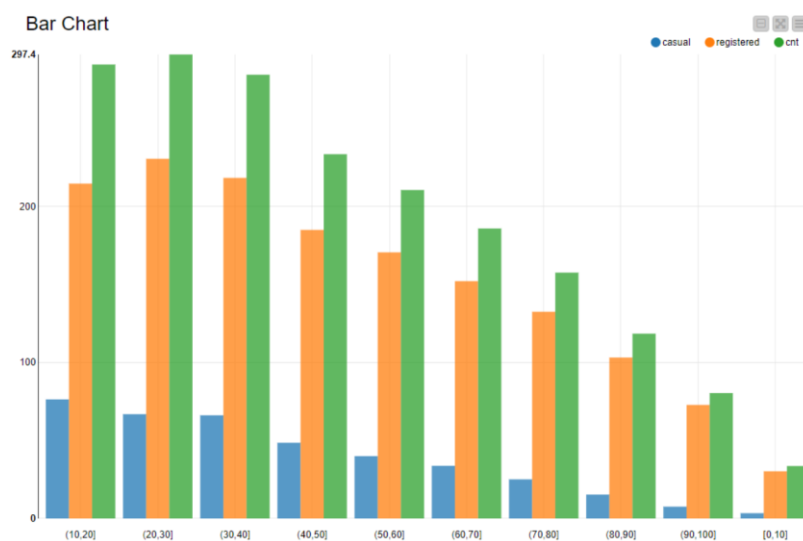
No atributo atemp, ou seja, a sensação térmica, verificamos uma tendência bastante similar à da temperatura, ou seja, um aumento nos alugueres nas temperaturas médias, e menor nas temperaturas baixas. Podemos visualizar que quando a sensação térmica se encontra entre os valores 30 e 36, é quando existe a maior média de bicicletas alugadas.



26. Alugueres nas diferentes sensações térmicas

4.2.11 Hum

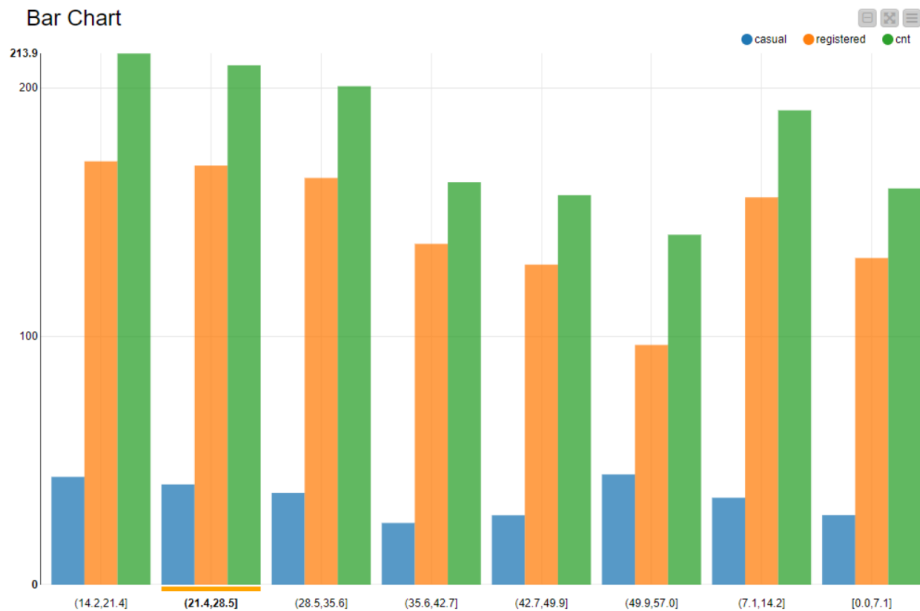
No atributo da humidade podemos visualizar também que quando a humidade aumenta a tendência é serem feitos menos alugueres, assim como quando o valor do atributo é muito baixo existem menos alugueres feitos.



27. Alugueres nas diferentes humidades

4.2.12 Windspeed

Neste atributo podemos observar que inicialmente entre os valores 0 e 21 existe um aumento, no entanto a partir desses valores, com o aumento da velocidade do vento, o aluguer de bicicletas diminui.



28. Alugueres nas diferentes velocidades do vento

4.3 Preparação dos dados

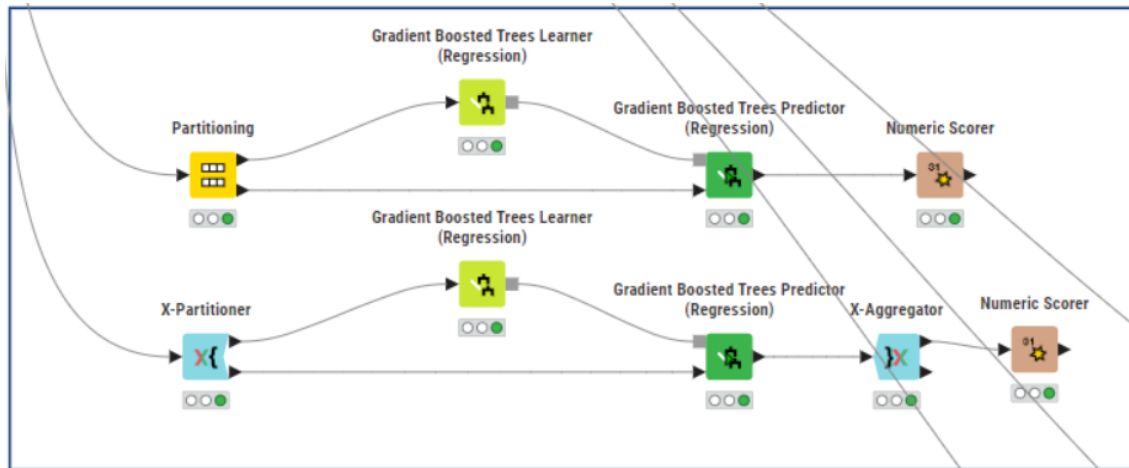
Após realizar uma análise dos dados presentes no dataset, o tratamento que decidimos efetuar foi a remoção do atributo dateday, devido a sua obvia correlação com os dados acerca de ano, mês e dia. Para os modelos também removemos os atributos casual e registered por causa da sua relação com cnt, sendo a soma destes atributos o resultado do atributo alvo.

Foi também feito posteriormente o tratamento de outliers encontrados no atributo cnt através de um Box Plot, isto para podermos realizar testes com diferentes tratamentos nos modelos.

4.4. Modelação

4.4.1. Modelação inicial

Na fase de modelação começamos por criar um modelo mais simples, que utiliza apenas os dados tratados na fase de preparação dos dados.



29. Modelação com Gradient Boosted Trees

Decidimos utilizar os nodos **Linear Regression**, **Simple Regression** e **Gradient Boosted Trees**, uma vez que este se trata de um problema de regressão. Outra escolha feita foi utilizar tanto nodos de **Cross Validation**, como de **Hold-Out Validation**, de forma a tentar obter o melhor método possível.

Statistics ...	Statistic...	Statistics ...
File	File	File
R ² : 0,677	R ² : 0,891	R ² : 0,919
Mean absolute error: 76,966	Mean absolute error: 35,339	Mean absolute error: 33,42
Mean squared error: 10 985,51	Mean squared error: 3 623,56	Mean squared error: 2 685,587
Root mean squared error: 104,812	Root mean squared error: 60,196	Root mean squared error: 51,823
Mean signed difference: -1,281	Mean signed difference: 1,476	Mean signed difference: -1,223
Mean absolute percentage error: 2,75	Mean absolute percentage error: 0,38	Mean absolute percentage error: 0,575
Adjusted R ² : 0,677	Adjusted R ² : 0,891	Adjusted R ² : 0,919

30. Numeric Scorer: Liner Regression vs Simple Regression vs Gradient Boosted

Estes são os resultados obtidos nos três algoritmos com Hold-Out Validation, que de um modo geral acabou por obter resultados ligeiramente melhores, embora que a diferença é quase insignificante.

Observa-se imediatamente que o algoritmo Gradient Boosted Trees é o que apresenta melhores resultados em quase todas as métricas. Podemos então considerá-lo o melhor algoritmo para este problema.

4.4.2. Modelação com tratamento de outliers

De seguida decidimos fazer uma modelação já com o tratamento dos outliers.

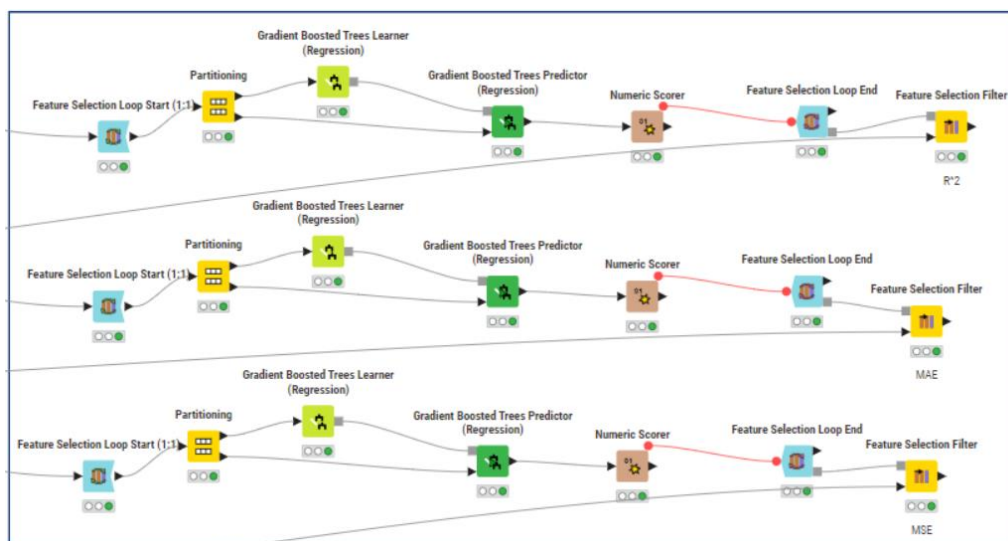
Statistics ...	Statistic...
File	File
R ² :	R ² :
Mean absolute error:	Mean absolute error:
Mean squared error:	Mean squared error:
Root mean squared error:	Root mean squared error:
Mean signed difference:	Mean signed difference:
Mean absolute percentage error:	Mean absolute percentage error:
Adjusted R ² :	Adjusted R ² :

31. Numeric Scorer: Gradient Boosted com/sem tratamento de outliers

Como é possível ver acima, os resultados melhoraram no que toca às principais métricas. O valor de R^2 está mais perto de 1, o que nos indica que os valores previstos por este modelo são uma melhor aproximação dos valores reais. O MAE e o MSE ambos baixaram, ou seja, o modelo em média erra menos que o anterior.

4.4.3. Modelação com feature selection

De forma a descobrir quais são os melhores atributos a ser utilizados, decidimos utilizar o processo de Feature Selection. Para isso usamos os nodos Feature Selection Loop Start, Feature Selection Loop End e Feature Selection Filter. Estes nodos utilizam um modelo de previsão e de seguida testam para descobrir quais são os melhores atributos a serem utilizados em modelos de previsão. Como modelo de aprendizagem usamos o Gradient Boosting, uma vez que este foi o que obteve os melhores resultados nos modelos anteriores.



32. Modelação com feature selection

Na figura acima pode-se ver que foram avaliadas três métricas diferentes. As três métricas escolhidas foram o R^2 , o MAE(Mean Absolute Error) e MSE(Mean Squared Error).

Nos processos para a avaliação da métrica R^2 os atributos escolhidos foram todos à exceção do atemp, com um valor de 0,926.

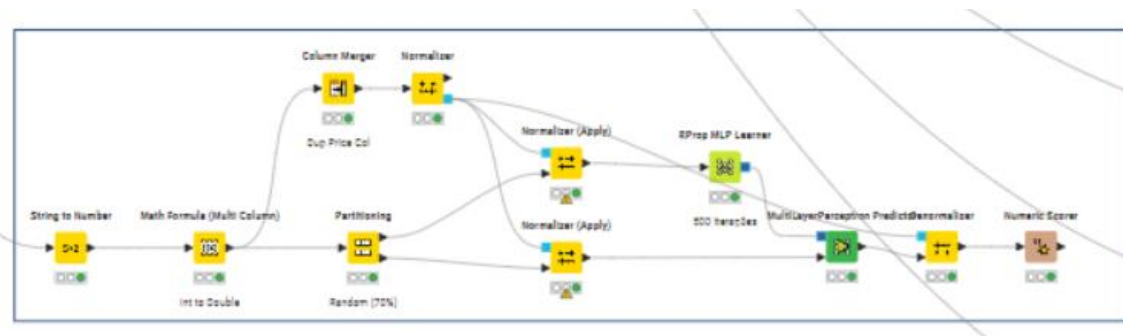
Para a métrica MAE os atributos são todos à exceção de weathersit e windspeed, com um valor de 31,546.

Por último, para a métrica MSE os atributos são season, yr, hr, weekday, workingday, weather sit, atemp e hum, com um valor de 2388,273.

É importante denotar que apesar de estes serem os melhores resultados para estes testes, pode haver uma combinação melhor de atributos que tenham um desempenho superior, comprovado pelo facto de haver outras combinações com valores muito semelhantes aos obtidos, e que em circunstâncias diferentes podem-se destacar outros atributos.

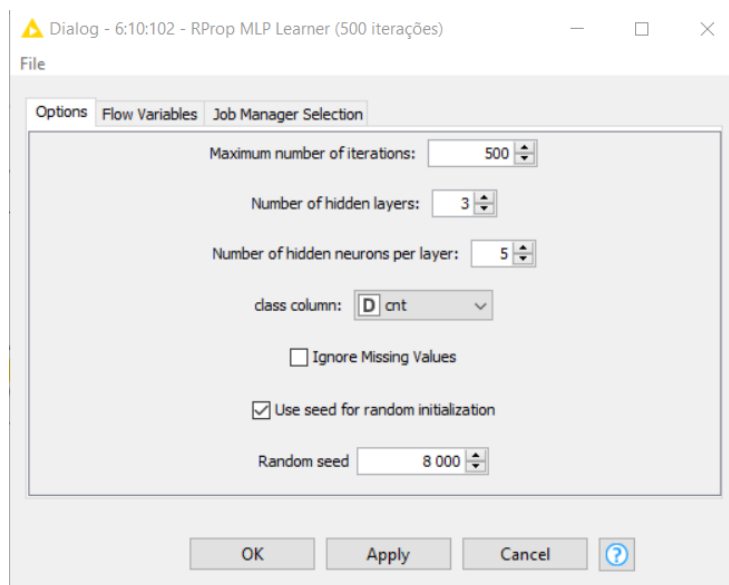
4.4.4. Modelação com redes neuronais

O último modelo realizado foi o das redes neuronais. Para este utilizamos os nodos Rprop MLP Learner e MultiLayerPerceptron Predictor como algoritmos de aprendizagem e previsão, respetivamente.



33. Modelação com Clustering

Como configuração utilizada na modelação do problema temos um número máximo de iterações de 500, 3 camadas escondidas e 3 neurónios contidos em cada camada. Esta foi a configuração que melhor resultado nos trouxe.



34. Configuração RProp MLP Learner

Statistics ...		Statistics ...	
File		File	
R ² :	0,847	R ² :	0,919
Mean absolute error:	47,582	Mean absolute error:	33,42
Mean squared error:	4 961,341	Mean squared error:	2 685,587
Root mean squared error:	70,437	Root mean squared error:	51,823
Mean signed difference:	1,686	Mean signed difference:	-1,223
Mean absolute percentage error:	0,976	Mean absolute percentage error:	0,575
Adjusted R ² :	0,847	Adjusted R ² :	0,919

35. Numeric scorer: Redes Neurais VS Gradient Boosted

Os valores obtidos pela rede neuronal são piores em todas as métricas que aqueles obtidos com Gradient Boosting, o que significa que este não é o modelo mais indicado para este problema.

4.5. Avaliação

Com base na nossa análise, podemos concluir que embora não tenha sido preciso muita preparação dos dados, deu para observar que nem todos os atributos neste dataset têm uma influência significativa no número de alugueres de bicicletas.

Com a modelação foi possível concluir que o algoritmo mais bem adaptado para este problema é o Gradient Boosted Trees, pois foi aquele com melhores valores em todas as métricas.

Com o Selection Filter não se conseguiu descobrir a combinação ideal de atributos para gerar o melhor modelo possível, uma vez que para cada métrica estudada houve um conjunto de atributos diferentes.

Quanto às redes neurais, o modelo foi incapaz de superar os resultados com Gradient Boosting, o que se deve possivelmente ao facto que em alguns casos, modelos mais simples como gradient boosting, podem ser mais eficazes em capturar as relações nos dados.

5. Conclusão

Com este trabalho prático, fomos capazes de consolidar a matéria lecionada e pôr em prática os conceitos de modelos de aprendizagem aprendidos ao longo do semestre.

Em relação aos datasets utilizados, escolhemos o CRISP-DM como metodologia de análise, realizamos então para cada uma exploração dos dados de forma a perceber melhor o problema e construir os melhores modelos para esse problema.

Ambas as tarefas apresentaram as suas complicações, incluindo a escolha do próprio dataset para a tarefa de Grupo, assim como a construção dos modelos e tratamento de dados. No entanto estamos satisfeitos com o trabalho realizado e consideramos ter alcançado os objetivos, tendo utilizado diversas estratégias para a criação de modelos e feito bastantes testes com diferentes tratamentos dos dados.