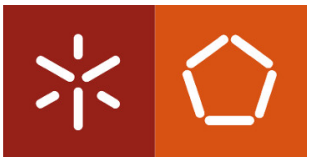
Universidade do Minho

Escola de Engenharia



# Aprendizagem e Decisão Inteligentes

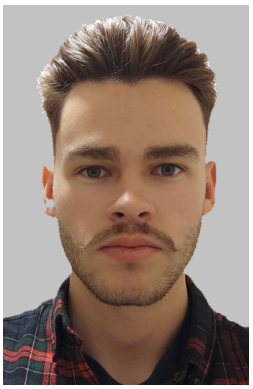
Licenciatura em Engenharia Informática

# Relatório do Trabalho Prático

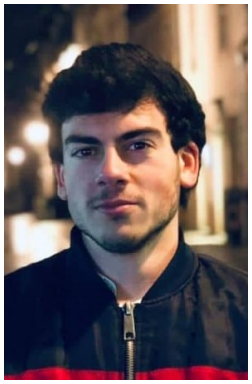
# Conceção de Modelos de Aprendizagem

# Grupo 30

José Magalhães **A93273**



Carlos Dias **A93185**



Francisco Izquierdo **a93241**

Duarte Lucas **A89526**

Maio, 2022

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# 1. Introdução

No âmbito da disciplina de Apredizagem e Decisões Inteligentes, foi proposto ao grupo de trabalho a elaboração de um plano que permitisse a extração e análise de conhecimento a partir de *data sets*. Este plano foi modelado e estruturado de forma a alcançar os objetivos proposto, bem como realizar retificações em relação ao mesmo. Com o intuito de auxiliar o processo de extração e análise de conhecimento, foi usado a ferramenta *KNIME*, no qual o grupo de trabalho usou exaustivamente a mesma. Além disso, o grupo de trabalho adotou uma metodologia com base no propósito de puder organizar e estruturar todo o trabalho desnevolvido, sendo a metodologia adotada, detalhada à frente. Com isto, o grupo de trabalho conseguiu também retificar algumas decisões tomadas, por forma a melhorar os resultados obtidos. Por fim, foi feito um levantamento dos vários modelos construidos e as correspondentes interpretações dos resultados obtidos.

# 2. Contextualização

## 2.1 Domínios e Objetivos

O ramo de Inteligência Artificial permite abordar resolver uma panóplia de problemas que não seriam de outra forma resolvidos através de técnicas de programação mais convencionais.

No contexto deste trabalho foi-nos atribuido um dataset de classificação de salários o qual contém atributos a cerca de trabalhadores assim como uma classificação de salário, podendo esta ser mais ou menos de 50K. Teremos assim como objetivo tentar criar um modelo de classificação capaz de prever com o maior grau de precisão e confiança possível qual o grupo de salário de cada trabalhador.

Foi-nos ainda pedido que escolhece-mos um dataset de uma das fontes mencionadas no enunciado. O grupo decidiu escolher o dataset MNIST, um dataset de classificação de números manuscritos, ou seja, tomámos como objetivo conseguir criar pelo menos um modelo capaz de classificar com o maior grau de precisão e confiança as imagens contidas neste dataset.

# 3. Metodologia

No ambito deste trabalho prático utilizámos um dataset escolhido pelo grupo e outro proposto pela equipa docente de forma a tentar analisar e consequentemente extrair conhecimento destes. Por forma a organizar e estruturar o processo de análise e extração de conhecimento a partir dos datasets supramencionados, o grupo de trabalho decidiu seguir a seguinte metodologia.

Primeiramente, foi feita a ingestão de dados através do fornecimento dos datasets, de forma a inferir sobre os mesmo e posteriormente analisar e extrair conhecimento.

Seguidamente foi feita uma análise e exploração dos dados de modo a tentarmos perceber a que tipo de tratamentos os datasets seriam sujeitos, usando diversos métodos na vertente estatística, de forma a perceber a informação contida nos datasets e as suas características, a qualidade dos dados e encontrar possíveis padrões e informação relevante, que nos ajudassem na preparação dos dados.

Tendo já uma ideia da estrutura e possíveis problemas nos dados, seguiu-se o tratamento dos dados e consequente preparação dos mesmos. Com isto, o grupo realizou a preparação dos dados tendo em conta algumas técnicas em duas vertentes: básica, das quais se destacam a filtragem de colunas; avançada, das quais se destacam normalização dos dados e SMOTE.

De seguida foram desenvolvidos modelos de classificação supervisionados aos quais foram passados os dados de treino de modo a retirarem informação útil e os mesmos puderem treinar, sendo feitas várias tentativas com valores diferentes de modo a procurar obter os melhores resultados possíveis, após serem feitas as devidas retificações ao serem fornecidos dados de teste, por forma a comparar e avaliar os modelos.

Com os modelos já na sua melhor versão foi feita uma análise para verificar qual deles é o mais apropriado para o problema em questão.

# 4. Datasets

## 4.1 Descrição e Exploração dos Datasets

### 4.1.1 MNIST

O dataset escolhido pelo grupo foi o MNIST, um dataset que contem cerca de 42000 imagens (28px x 28px) de numeros escritos à mão etiquetados com o valor do digito nelas contidos. Este dataset é bastante conhecido pela comunidade de Inteligência Artificial e normalmente visto como uma referência para comparar diferentes métodos de aprendizagem. O ficheiro utilizado contém 785 colunas, a primeira é a etiqueta com o valor escrito na imagem, isto é, um numero entre 0 e 9, as restantes 784 são as cores de cada um dos pixeis da imagem, valores estes que variam entre 0 e 255 e representam um nivel de luminosidade entre preto (0) e branco (1).

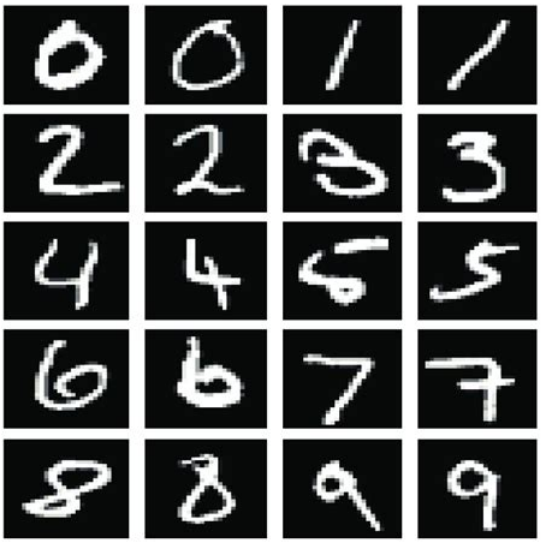


Figura 4.1: Dataset → *Mnist*

### 4.1.2 Salary_Classification

Além do dataset escolhido, o grupo teve de trabalhar com um segundo dataset proposto pela equipa docente, o dataset salary_classification. Este dataset tem em vista prever o salário de um trabalhador que resulta de diversos factores. Inconscientemente é influenciado pelo nível de escolaridade do trabalhador, idade, sexo, ocupação, etc.

Este contém um conjunto de dados composto por 15 colunas e etiquetado por duas classes, sendo menor ou igual a 50k e superior a 50k. Contém 14 atribuitos baseados em dados demográficos, entre outros recursos como descrição do individuo, todos estes se encontram detalhadamente descritos no pdf disponibilizado pelos docentes.

## 4.2 Tratamento dos Datasets

### 4.2.1 MNIST

O dataset MNIST, embora seja um dataset relativamente limpo e organizado contém colunas que contêm sempre o mesmo valor, ou seja, há certos pixeis que apresentam a mesma cor em todas as imagens, isto não adiciona qualquer tipo de informação relevante que ajude no processo de aprendizagem, como tal, e de modo a tentar otimizar o processo de aprendizagem, decidimos remover todas as colunas que contenham sempre o mesmo valor. Outra caracteristica deste dataset é que os valores das colunas variam entre 0 e 255, seria normal pensar que uma normalização dos valores seria desnecessária uma vez que se encontram todos na mesma escala, no entanto a normalização ajuda bastante alguns algoritmos de aprendizagem uma vez que permite uma convergência muito mais rápida do gradiente em algoritmos tais como propagação inversa (backpropagation), algoritmo este utilizado pelas redes neurais que implementamos. Importa também realçar que uma vez que este problema se trata de um problema de classificação as etiquetas devem ser do tipo categoria (string) e não do tipo inteiro ou double, para tal aplicámos também o nodo number to string que passa as etiquetas (valores entre 0 e 9) para strings de modo a serem aceites pelos métodos de aprendizagem.

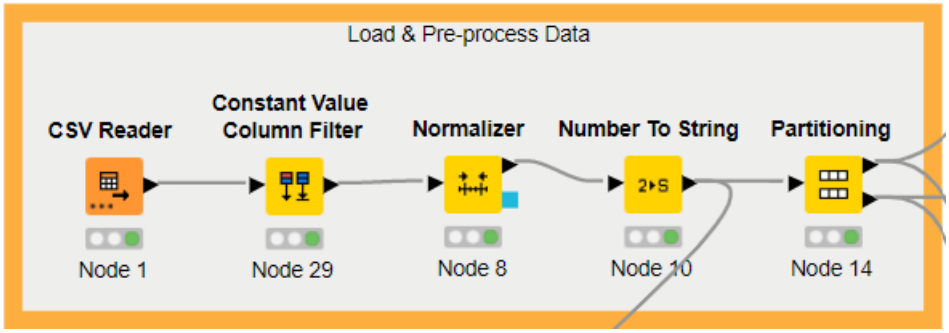



Figura 4.2: Tratamento de Dados: MNIST

### 4.2.2 Salary Classification

Tal como no dataset discutido anteriormente, também o salary_classification contém apenas informação pretinente e útil para o problema em questão, no entante este contém alguns missing values. Algumas tentativas de lidar com estes missing values (tais como eliminar as linhas ou substituir pela média) revelaram descidas de precisão na previsão de resultados o que nos mostrou que a falta de valores contém informação relevante para o processo de decisão do resultado, como tal, o grupo optou por não remover ou substituir os valores em falta.

Reparámos também durante a análise do dataset através do nodo statistics que o dataset se encontrava desbalanceado, ou seja, continha mais do dobro de salários menores oou iguais a 50K do que maiores. Como forma de resolver este problema o grupo optou por usar o nodo SMOTE, isto é, adicionar linhas artificiais com salários maiores que 50K criadas através dos valores dos 5 vizinhos mais próximos.

Uma vez que este dataset também seria utilizado por uma rede neural tivemos de normalizar os valores contidos pelas mesmas razões descritas no dataset anterior. Passámos ainda as

categorias sob a forma de string para valores numéricos de forma a poderem ser passados a uma rede neural. Este processo em nada afeta a performance dos outros métodos tais como decision tree, sendo por isso os métodos de tratamento do dataset partilhados entre os diferentes métodos.

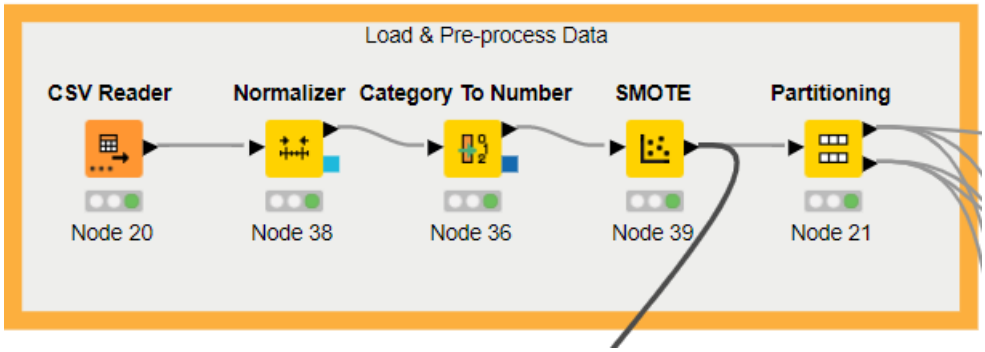



Figura 4.3: Tratamento de Dados: Salary Classification

# 5. Modelos

## 5.1 Modelos Desenvolvidos

De modo a antigirmos o maior nivel de precisão possível na previsão da etiqueta deste dataset tivemos de experimentar várias abordagens no que toca aos modelos utilizados. Uma vez que o problema abordado se trata de um problema de classificação e não de regressão tivemos logo de excluir todos os modelos de regressão.

### 5.1.1 Decision Tree

Começámos por utilizar o modelo decision tree learner em ambos os datasets uma vez que se trata de um dos modelos mais comuns e permite ter uma ideia do nivél de complexidade assim como perceber se se trata de um problema cuja aprendizagem possa ser difícil e demorada.

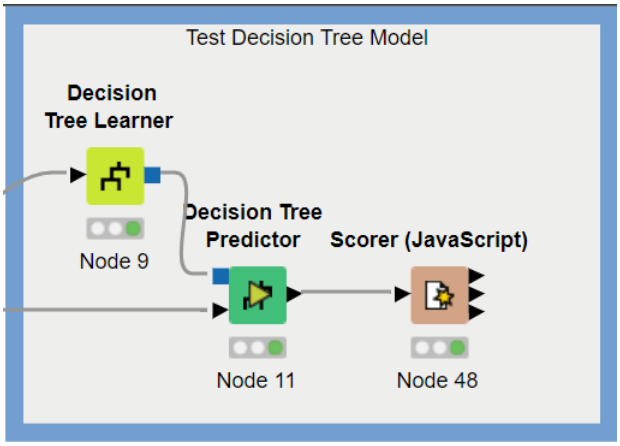



Figura 5.1: MNIST

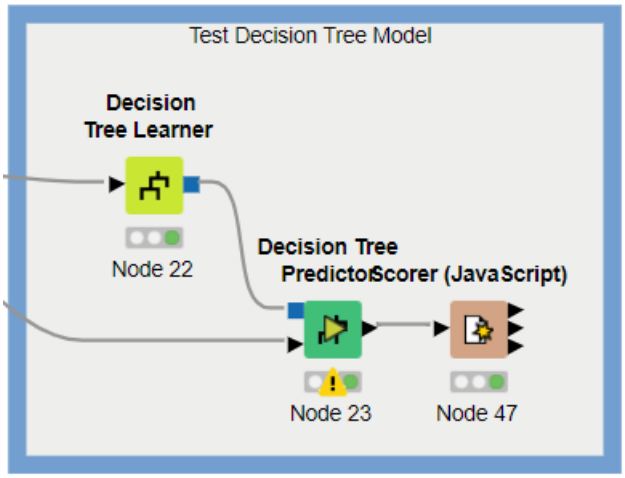



Figura 5.2: Salary Classification

### 5.1.2 Random Forest

Uma vez que as random forests são modelos baseados em conjuntos de decision trees o grupo optou por apenas usar este modelo nos datasets que apresentem bons resultados nos modelos do tipo decision tree uma vez que geralmente as random forests apresentam melhor performance do que as decision trees e têm a capacidade de aprender dados mais complexos tal como será de esperar no dataset salary classification.

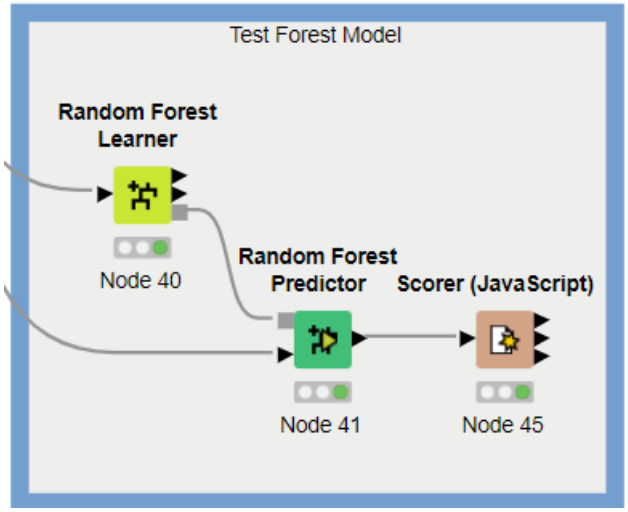



Figura 5.3: Salary Classification

### 5.1.3 Neural Network

As redes neurais acabaram por se revelar um dos métodos de aprendizagem mais poderosos, em grande parte pela capacidade de relacionar os inputs de forma muito mais profunda do as árvores de decisão em problemas como o MNIST, em que as colunas estão profundamente relacionadas e apenas podem ser tiradas conclusões analisando várias colunas ao mesmo tempo em vez de analisa-las separadamente e sequencialmente.

O dataset MNIST é um problema de reconhecimento visual e como tal, de modo a uma rede neural o poder aprender de forma precisa, a recolha de um elevado número de características dos dados iniciais é mais importante do que o processamento da complexidade desses dados, como tal, decidimos favorecer o número de neurónios por camada em relação ao número de camadas da rede neural por forma a captar o máximo de informação relevante dos dados iniciais sem a comprimir demasiado. A mesma linha de pensamento foi seguida e levada ao extremo no que toca ao dataset salary_classification, que aliada de um grande número de iterações (cerca de 1000) se revelou a forma mais eficaz de aprender e prever este dataset.

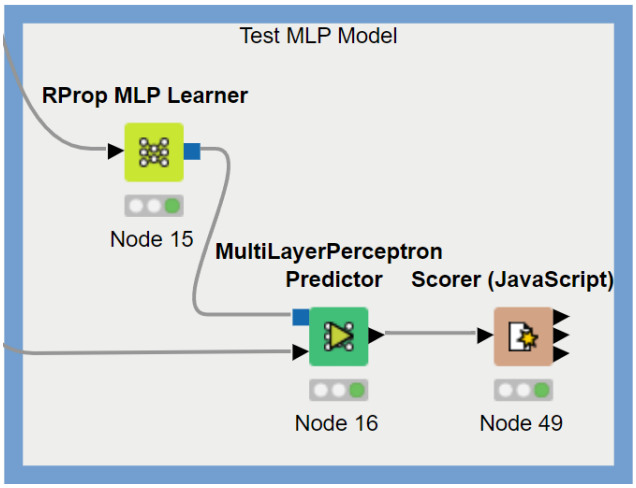



Figura 5.4: MNIST

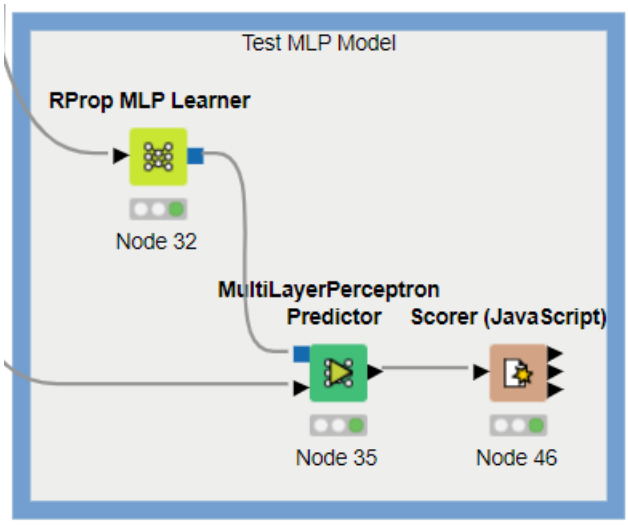



Figura 5.5: Salary Classification

# 6. Parâmetros de Treino e Resultados

| Tentativas | Partition Size | Quality | Pruning | Min Records/Node | Accuracy |
|---|---|---|---|---|---|
| 1 | 80 | Gain Ratio | MDL | 4 | 84.43% |
| 2 | 80 | Gain Ratio | No Prunnig | 4 | 84.81% |
| 3 | 80 | Gini Index | No prunnig | 2 | 85.00% |
| 4 | 80 | Gini Index | No Prunnig | 5 | 85.76% |
| 5 | 80 | Gini Index | No Prunnig | 3 | 85.83% |
| 6 | 80 | Gini Index | No Prunnig | 4 | 85.96% |
| 7 | 80 | Gini Index | MDL | 4 | 86.26% |

Tabela 6.1: Resultados: MNIST (Decision Tree)

| Tentativas | Partition Size | Iterations | Hidden Layers | Neurons per layer | Accuracy |
|---|---|---|---|---|---|
| 1 | 80 | 100 | 1 | 55 | 94.381% |
| 2 | 80 | 120 | 1 | 65 | 94.286% |
| 3 | 80 | 100 | 1 | 60 | 94.619% |
| 4 | 80 | 100 | 1 | 65 | 95.048% |
| 5 | 80 | 90 | 1 | 65 | 95.2% |

Tabela 6.2: Resultados: MNIST (Neural Network)

| Tentativas | Partition Size | Quality | Pruning | Min Records/Node | Accuracy |
|---|---|---|---|---|---|
| 1 | 75 | Gini Index | No Prunnig | 4 | 84.54% |
| 2 | 75 | Gain Ratio | No Prunnig | 4 | 84.92% |
| 3 | 75 | Gini Index | MDL | 4 | 85.21% |
| 4 | 75 | Gain Ratio | MDL | 3 | 85.32% |
| 5 | 75 | Gain Ratio | MDL | 5 | 85.67% |
| 6 | 75 | Gain Ratio | MDL | 4 | 85.72% |

Tabela 6.3: Resultados: Salary Classification (Decision Tree)

| Tentativas | Partition Size | Iterations | Hidden Layers | Neurons per layer | Accuracy |
|---|---|---|---|---|---|
| 1 | 75 | 100 | 2 | 20 | 82.721% |
| 2 | 75 | 100 | 2 | 70 | 82.732% |
| 3 | 75 | 100 | 1 | 30 | 82.759% |
| 4 | 75 | 100 | 1 | 25 | 82.883% |
| 5 | 75 | 100 | 1 | 20 | 83.034% |

Tabela 6.4: Resultados: Salary Classification (Neural Network)

| Tentativas | Partition Size | Split Criterion | Number of models | Accuracy |
|---|---|---|---|---|
| 1 | 75 | Information Gain Ratio | 100 | 88.325% |
| 2 | 75 | Information Gain | 100 | 88.664% |
| 3 | 75 | Gini Index | 100 | 88.831% |
| 4 | 75 | Gini Index | 110 | 88.874% |
| 5 | 75 | Gini Index | 120 | 88.885% |
| 6 | 75 | Gini Index | 200 | 88.912% |

Tabela 6.5: Resultados: Salary Classification (Random Forest)

## 6.1 Analíse Crítica

### 6.1.1 MNIST

Tal como fora proposto na secção de modelos desenvolvidos não só os melhores resultados na classificação das imagens do MNIST foram melhores usando o modelo de rede neural mas também fomos capazes de obter melhores resultados através do favorecimento do número de neurónios por camada em vez do número de camadas, provando também assim que a deteção de características é mais importante do que o processamento dessas características em problemas de reconhecimento de imagens.

### 6.1.2 Salary Classification

Os resultados obtidos em relação a este dataset mostram que métodos de aprendizagem tais como decision trees e random forests são capazes de gerar melhores resultados do que redes neurais. Tendo em conta os bons resultados obtidos na decision tree, a implementação da random forest revelou-se uma decisão acertada gerando estas os melhores resultados neste dataset.

# 7. Sugestões/Recomendações

Acreditamos ter cumprido com aqueles que eram os nossos objetivos em relação aos resultados que pretendiamos obter com o datatset MNIST, no entanto, gostariamos de ter obtido melhores resultados no que toca à classificação de salários. Uma das possíveis sugestões seria tentar utilizar outros tipos de modelos, talvez até de carácter não supervisionado de modo a aumentar não só o nível de precisão do modelo mas também o valor de Cohen's kappa (k).

# 8. Conclusão

Em suma, o grupo de trabalho cumpriu com os requisitios propostos, uma vez que conseguiu planificar de forma eficiente e completa o plano que serviu de base e que implementou a metodologia proposta e abordada ao longo deste relatório. Com a metodologia enunciada, o grupo de trabalho conseguiu organizar e tornar todo o processo coerente, na medida em que a mesma metodologia, permitiu subdividir de forma concisa o trabalho em tarefas. Cada tarefa mostrou ser importante, começando na tarefa de ingestão de dados, isto nos dois *datasets*, passando à análise dos dados adotando diversas estratégias, seguindo para o pré-processamento dos dados. Esta tarefa, mostrou ser bastante taxativa, mas bastante importante uma vez que se traduziu na melhoria da construção dos modelos bem como dos resultados obtidos. Posto isto, seguiu-se a já mencionada construção dos modelos, na vertente em que após a mesma, os modelos seriam treinados por forma a extrairem conhecimento dos dados fornecidos e tratados. Por fim, a tarefa que mais pesou na medida que permitiu guiar o grupo em busca de melhores soluções, foi o teste e retificação dos modelos anteriormente construidos e treinados. Com isto, o grupo de trabalho conseguiu tirar e retificar conclusões, pelo que o âmbito deste trabalho prático mostrou ser bastante fundamental de forma a consolidar conceitos.