

**Aplicação de C4.5 para a predição de espécies de anuros**

*Relatório Intercalar*

Inteligência Artificial

3º ano do Mestrado Integrado em Engenharia Informática e Computação

Elementos do Grupo:

Alessandro Antoniolli – up201710790 – ​up201710790@fe.up.pt

Luís Soares - up201406356 - ​up201406356@fe.up.pt

Nuno Pereira - up201506265 - ​up201506265@fe.up.pt

08 de Abril de 2018

# Conteúdo

1. **Objetivo 3**
2. **Descrição 3**
   1. Especificação . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . 3
      1. Descrição do tema . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3
      2. Análise do dataset .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3
      3. Pré Processamento de Dados . . . . . . . . . . . . . .. . . . . . . . . . . . . . 5
      4. Modelo de Aprendizagem a aplicar .. . . . . . . . . . . . . . . . . . . . . 6
      5. Variantes e opções dos algoritmos a utilizar . . . . . . . . . . . . .6

2.2 Representação do conhecimento . . . . . . . .. . . . . . . . . . . . . . . . . . . 8

1. **Desenvolvimento 10**
2. **Experiências 11**
3. **Conclusões 12**
4. **Melhoramentos 13**
5. **Recursos 13**

# Objetivo

Este trabalho tem como principal meta através do algoritmo C4.5, interpretar e tentar melhorar as regras de classificação para a identificação de espécies de anuros.

Os anuros possuem várias caraterísticas que os identificam mas uma que os permite distinguir de forma exata é a emissão de sons, pois fazem-no em frequências distintas.

Começaremos por utilizar uma amostra como base de informação e com base nesses dados, através do algoritmo de aprendizagem treinar a rede neuronal aplicando o algoritmo C4.5, definindo um processo de aprendizagem e como tal, melhorar a forma como são diferenciadas as distintas espécies.

# Descrição

## Especificação

**2.1.1 Descrição do tema**

O seguinte trabalho tem como tema espécies de anuros. Um anuro pode ser um sapo, uma rã ou uma rela. Ainda que se possam estabelecer algumas diferenças entre sapos e rãs, estas diferenças não são utilizadas pelos cientistas na sua classificação. Os anuros são capazes de efetuar sons bastante altos que podem ser ouvidos a mais de um quilómetro de distância. Cada espécie emite um som exclusivo para a sua espécie e como tal é possível a partir do som emitido por um anuro identificar a espécie a que pertence.

**2.1.2 Análise do Dataset**

A frequência da emissão de som de cada anuro é um dado suficiente para o associar a uma determinada espécie e como tal um dado bastante significativo a ter em conta. Este tipo de dados já foi utilizado anteriormente para efeitos de classificação de espécies.

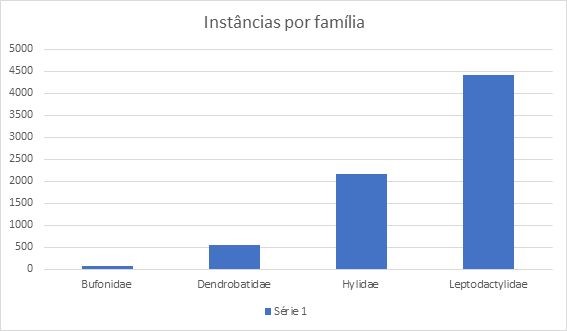
Este dataset é composto por 22 colunas com os respetivos MFCC’s, seguidos da família, género, espécie e um id da gravação. Foram analisadas gravações de 4 famílias, 8 gêneros e 10 espécies distintas de anuros. Existem quase 60 gravações para cada espécie analisada.

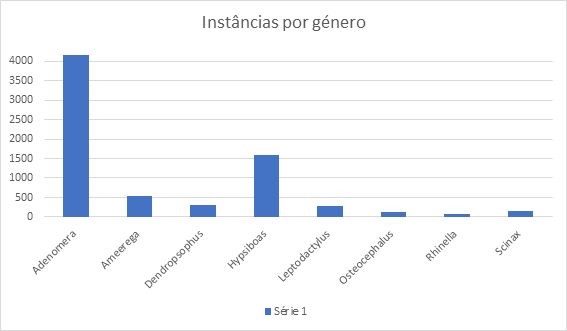
Para além disso, foram utilizados métodos de entropia espectral e clustering binário para encapsular os dados de cada sílaba. Depois da partição efetuada foram obtidas 7195 sílabas que serviram de base para teste e treino do classificador implementado.

É de realçar que estes dados foram recolhidos em ambientes reais, com ruído de fundo, tais como a mata Atlântica, Manaus e ainda de Córdoba na Argentina.

Os dados foram guardados em forma de onda com frequência de amostragem de 44 KHz e 32 bit de resolução, como tal, permitindo-nos analisar sinais até 22KHz. De cada sílaba obtida, foram extraídos 22 MFCC’s através da triangulação de 44 filtros.

O conjunto de sílabas que representa a base de dados têm as seguintes instâncias por cada classe:







**2.1.2 Pré-Processamento de Dados**

O processamento de dados é uma fase que depende maioritariamente da capacidade do analista de dados em identificar os problemas presentes nos dados e tentar sistematizar métodos apropriados para solucionar cada um deles.

Utilizamos alguns métodos de pré-processamento dos dados fracamente dependentes do conhecimento de domínio dado serem passíveis a sistematização, entre os quais:

Verificação da integridade dos dados: Nenhum dos dados apresenta discrepâncias de valores suficientemente grandes para justificar dúvidas quanto á sua origem.

Identificação de inconsistências: Eliminamos toda a redundância que pode ocorrer quando uma informação essencialmente idêntica é armazenada em diversos atributos, eliminando todas as ocorrências exatamente iguais.

Definir métodos de seleção de atributos: Seria inviável definirmos uma razão de decisão para a construção do nosso caminho de escolha que não fosse idealizada tendo em conta possíveis padrões e aspectos peculiares à mesma.

Neste trabalho não empregamos métodos de pré-processamento fortemente dependentes do conhecimento de domínio dado não termos encontrado uma norma específica que conseguíssemos automatizar.

**2.1.3 Modelo de Aprendizagem a aplicar**

Para garantir uma solução eficaz ao paradigma do nosso trabalho vamos utilizar o algoritmo de aprendizagem indutiva c4.5 para a predição das espécies de anuros.

O objetivo desta abordagem é a aquisição de conhecimento através da construção automática de novas regras de conhecimento. Identificados os atributos significativos dos objetos do domínio considerado, é possível construir árvores de decisão, dessas árvores obtêm-se regras.

Para a escolha da construção das árvores de decisão e para a escolha das árvores de decisão das quais vamos inferir regras existem critérios que advém de algumas escolhas, por exemplo, consoante as escolhas do atributo existem critérios de escolha que dependem de forma direta dum cálculo da entropia ganho de informação



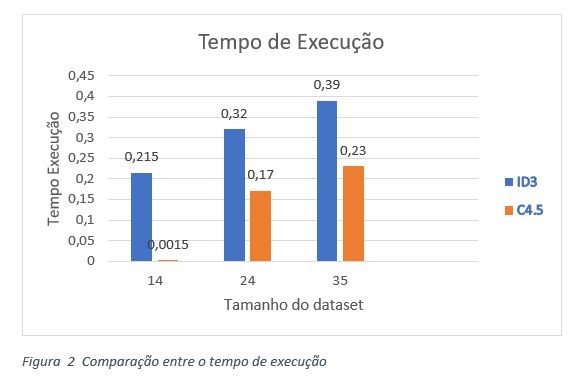
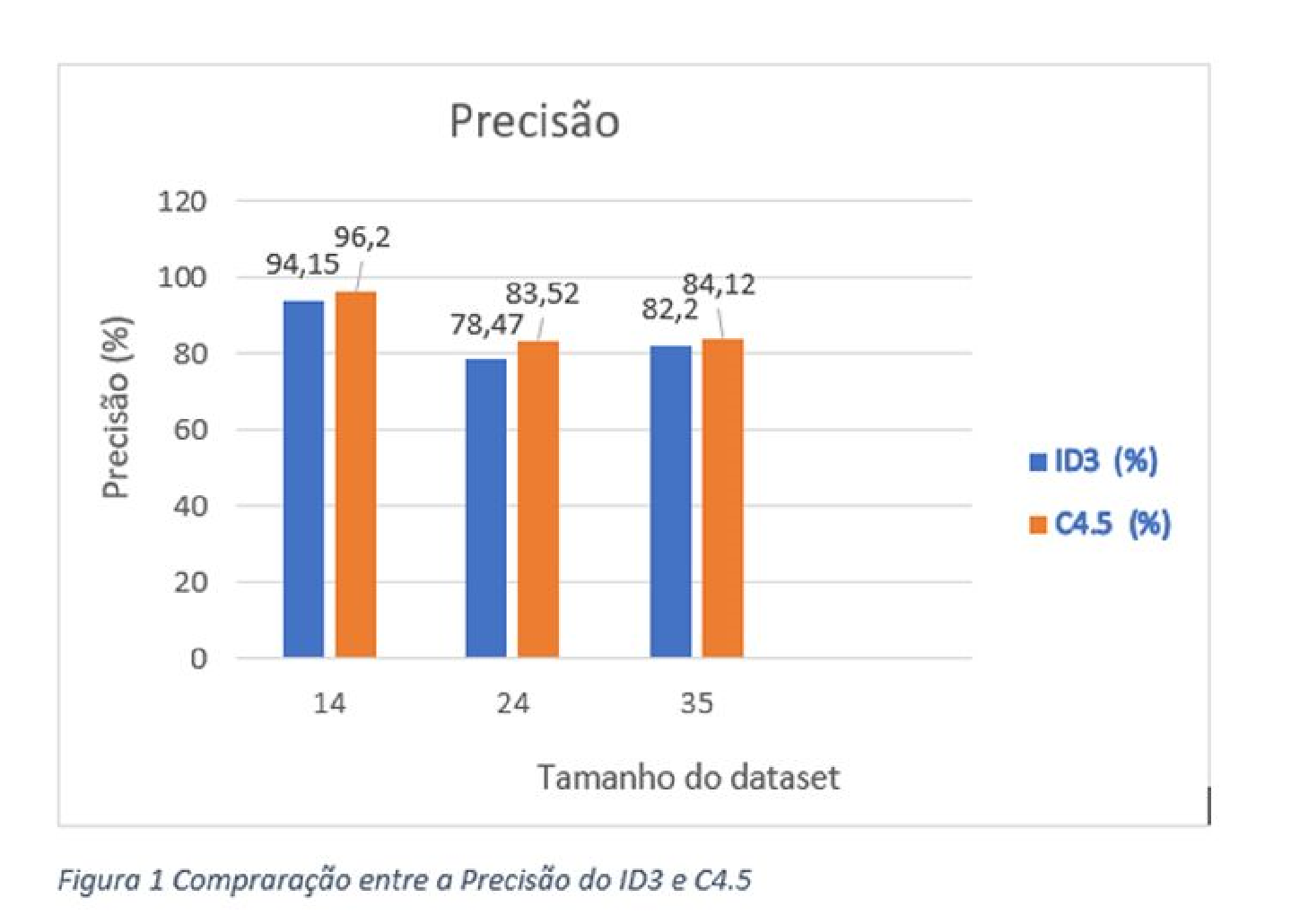
Pruning é uma técnica aplicada neste algoritmo que reduz o tamanho das árvores de decisão por remoção de seções da árvore que têm pouco poder para classificar instâncias, o duplo objetivo desta técnica é reduzir a complexidade final bem como aumentar a capacidade de previsão.

A razão de erro "pessimista" de uma subárvore é a média das razões de erro das suas folhas pesadas pela frequência relativa do número de casos do conjunto de treino que elas cobrem. Este processo de poda substitui então todas as subárvores que não aumentem significativamente o previsível erro da árvore.

A confiança na folha pode ser dada por: (e+1)/(n+2).

**2.1.4 Variantes e opções dos algoritmos a utilizar**

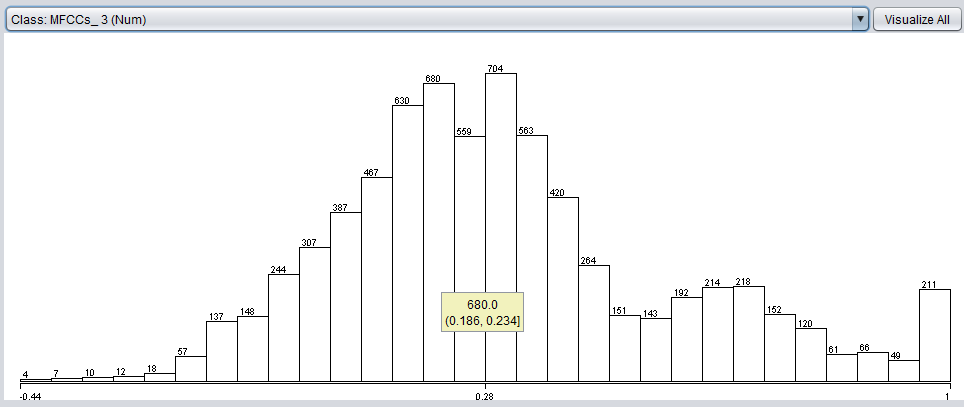
O algoritmo de classificação escolhido é antecedido do ID3 (inductive decision tree) e partilha várias características em comum, uma das principais razões deve-se a também ter como base a teoria da Informação de Shannon para a seleção dos nodos a serem utilizados na árvore de decisão. Apesar do algoritmo C4.5 agir de forma semelhante ao ID3 este apresenta algumas melhorias a nível de performance:



Não foram só as questões do potencial de “desempenho” dos algoritmos que nos baseamos para selecionar o algoritmo. Embora o ID3 continue a ser uma escolha bastante eficaz, este só lida com atributos discretos e conhecidos. O algoritmo C4.5 por sua vez permite-nos tratar conjuntos de treino com valores contínuos e desconhecidos, exatamente o que estamos á procura para a análise das frequências do som dos anuros. Não esquecer os métodos de pruning para amenizar a busca em árvores mais complexas, como anteriormente mencionado.

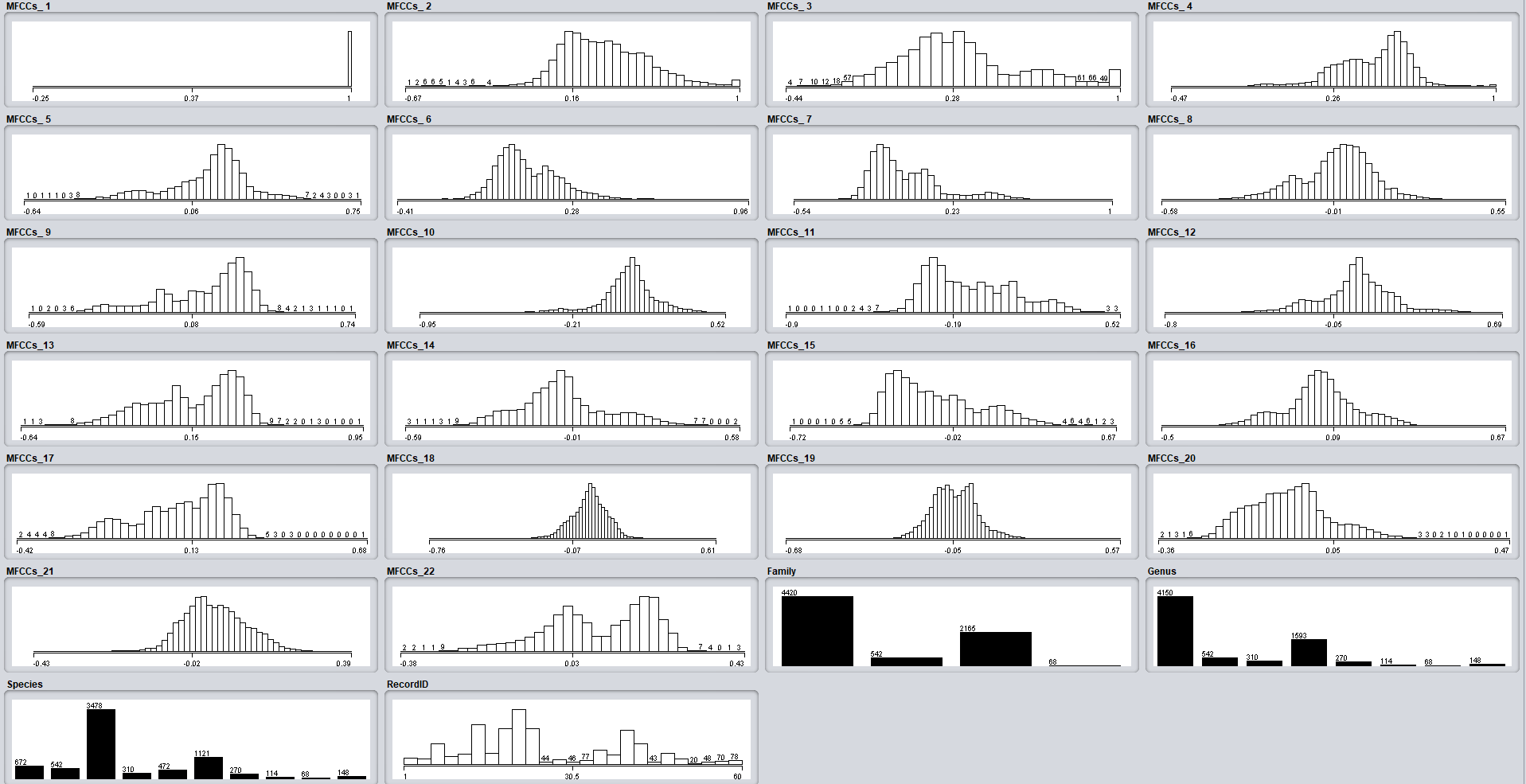
## Representação de conhecimento

Para ilustrar o conhecimento possuído até à data, optamos por escolher a framework Weka, visto que é uma ferramenta capaz de mostrar ao utilizador os dados fornecidos sob a forma de gráficos que permitem uma mais cómoda e rápida perceção e análise dos mesmos. É possível, por exemplo agrupar todos os dados numa representação ilustrativa dos dados, ou selecionar apenas uma das “variáveis” e visualizar apenas o seu gráfico com mais detalhe, como o exemplo seguinte que representa os dados que possuímos face ao MFCC 3 de cada sílaba na base de dados.

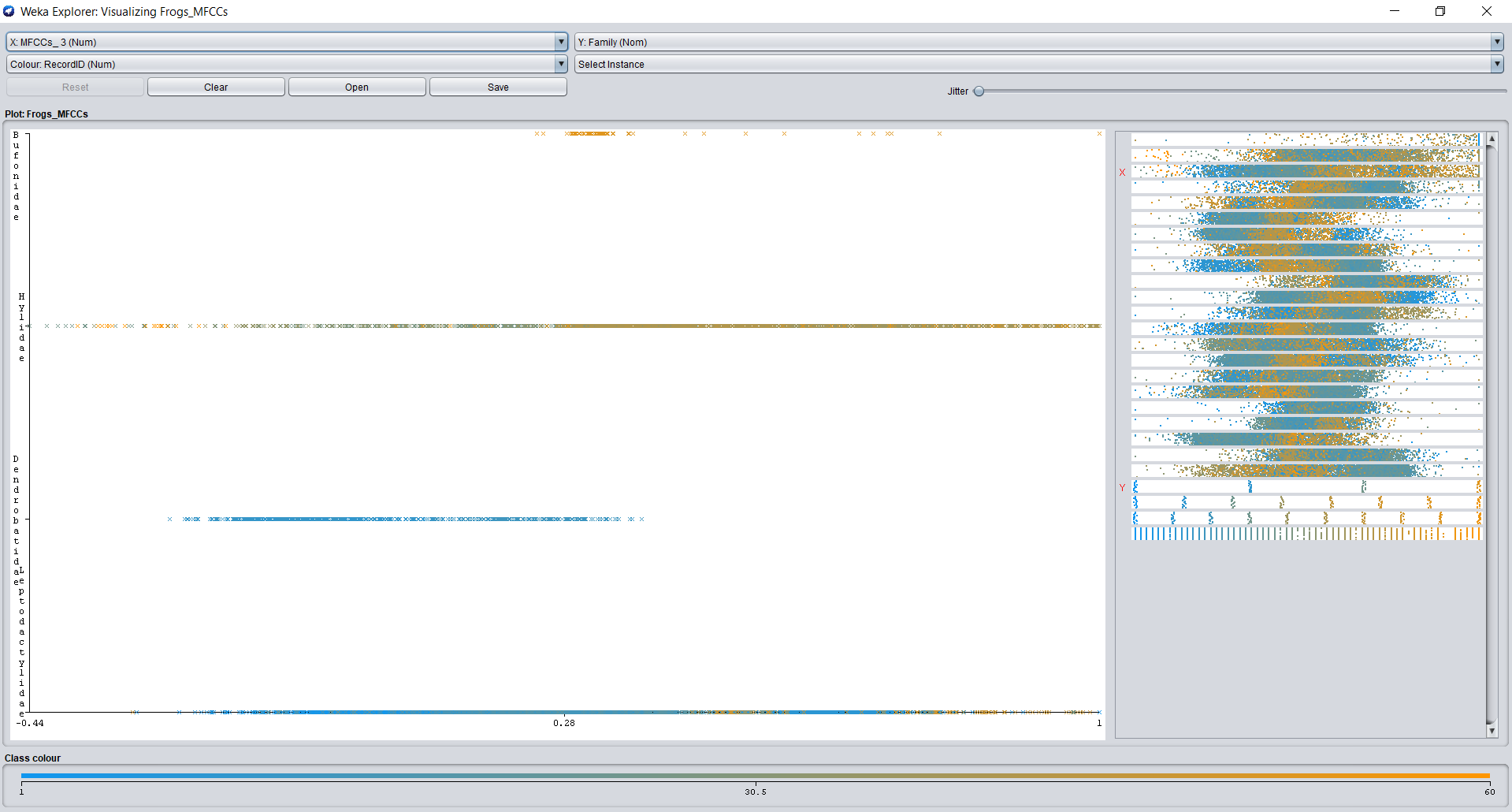


***Figura ilustrativa do MFCC 3.***

No entanto, no caso a seguir ilustrado, está representado um agrupamento de gráficos, um para cada um dos 22 MFCC’s, assim como a família, género e espécie dos anuros e ainda um gráfico que diferencia o número de sílabas encontradas em cada gravação.

 ***Figura ilustrativa de todos os dados agrupados.***

Esta ferramenta é também capaz de agrupar dois dados, sendo no caso abaixo a família e o MFCC 3 e representar em que zonas se situam os MFCC 3 de cada uma das espécies. No seguinte gráfico é ilustrado como é possível efetuar esse tipo de combinações.

****

Podemos ainda consultar os dados no log que a ferramenta disponibiliza.

# Desenvolvimento

Com vista ao desenvolvimento deste projeto foi utilizada a framework *Weka (Waikato Environment for Knowledge Anayisis)* capaz de suportar diferentes tipos de algoritmos de análise. Optamos por utilizar a mesma pois é capaz de representar os dados de uma forma user-friendly e dado que possui o algoritmo J48 que nos é bastante útil.

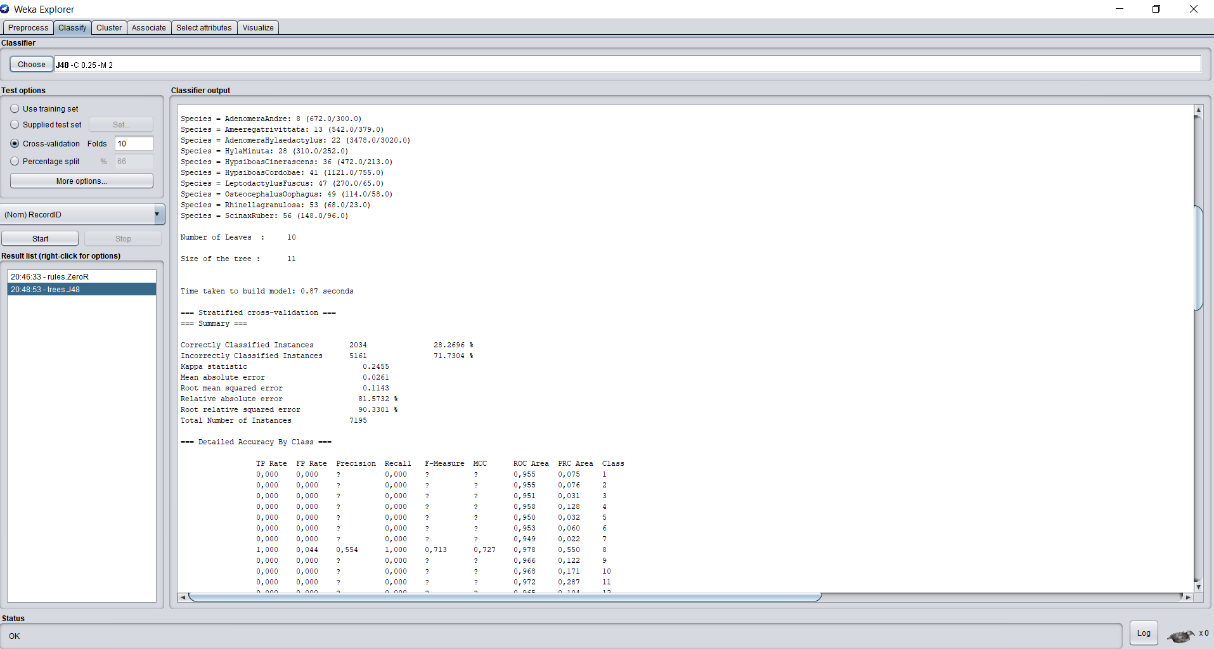
O projeto foi realizado em Java com auxílio ao programa Eclipse para posteriormente ser usado pelo WEKA.

O ambiente de desenvolvimento do projeto escolhido foi o Windows 10 por questões de comodidade.

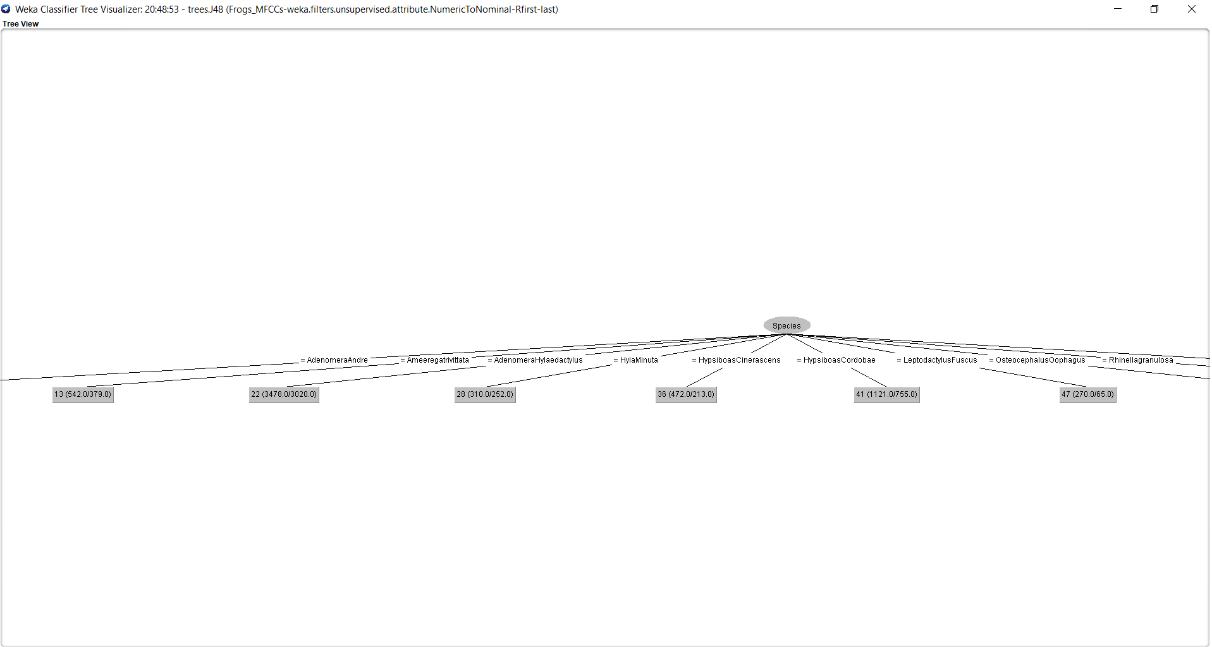
# Experiências

As experiências levadas a cabo tiveram como intenção perceber qual a melhor distribuição de dados para treino do algoritmo e para teste do correto funcionamento do mesmo, de forma a perceber a partir de que percentagem de dados fornecida como treino seria viável acreditar no correto funcionamento do algoritmo. Podemos também perceber a partir de que valores o aumento da percentagem dedicada a treino deixa de compensar o aumento da eficácia do algoritmo, tendo como principal objetivo encontrar uma distribuição capaz de satisfazer as necessidades básicas para um funcionamento ideal.

A seguir seguem algumas imagens após correr os dados com o algoritmo J48 de árvores de decisão:



É possível também visualizar a árvore através desta framework:



Apesar de não estruturada conforme planeado devido à falta de conclusão de implementação do código com as regras de decisão.

# Conclusões

Depois de uma cuidada análise, podemos concluir que ambos os algoritmos ID3 e C4.5 são capazes de resolver o problema apesar de o C4.5 possuir algumas vantagens em relação ao ID3, como por exemplo o facto de suportar tanto variáveis discretas e contínuas, sendo que consegue lidar melhor com dados de treino com atributos incompletos, lidar com atributos com diferentes pesos/custos, podendo também efetuar a “poda” da árvore de decisão após a sua criação através de um processo de retrocesso.

Apesar de não termos conseguido implementar alguns dos objetivos do trabalho, foi necessário um grande estudo sobre esta matéria que foi benéfico para melhorar o nosso conhecimento teórico. Infelizmente não o conseguimos transpor para o conhecimento prático como gostaríamos.

# Melhoramentos

Para melhorar o trabalho até aqui realizado, deveríamos optar por completar a alteração do algoritmo e implementar as regras de classificação.

Outra opção que poderia complementar o trabalho realizado seria implementar uma interface gráfica de forma a tornar a interação com o utilizador mais agradável. Não optamos por despender tempo para a realização desta tarefa porque considerámos mais importante o correto funcionamento dos algoritmo assim como da análise de resultados.

# Recursos

De forma a melhor compreender o problema que nos foi apresentado fomos investigar segundo as seguintes fontes:

* *“*​*Artificial Intelligence: A new synthesis" Nils Nilsson, Morgan Kaufmann, 1998*
* *“C4.5 programs for machine learning”* ​*J. Ross Quinlan, Morgan*

*Kaufmann, 1993*

http://saiconference.com/Downloads/SpecialIssueNo10/Paper\_3-A\_comparativ e\_study\_of\_decision\_tree\_ID3\_and\_C4.5.pdf https://paginas.fe.up.pt/~eol/IA/1718/APONTAMENTOS/6\_ASA.pdf

https://pt.slideshare.net/gladysCJ/evaluation-and-comparison-of-supervised-le arning-algorithms https://web.fe.up.pt/~arocha/MIA/0506/apdz2.pdf

# Apêndice (Manual)

De forma a melhor compreender o problema que nos foi apresentado fomos investigar segundo as seguintes fontes:

* *“*​*Artificial Intelligence: A new synthesis" Nils Nilsson, Morgan Kaufmann, 1998*
* *“C4.5 programs for machine learning”* ​*J. Ross Quinlan, Morgan*

*Kaufmann, 1993*

http://saiconference.com/Downloads/SpecialIssueNo10/Paper\_3-A\_comparativ e\_study\_of\_decision\_tree\_ID3\_and\_C4.5.pdf https://paginas.fe.up.pt/~eol/IA/1718/APONTAMENTOS/6\_ASA.pdf

https://pt.slideshare.net/gladysCJ/evaluation-and-comparison-of-supervised-le arning-algorithms https://web.fe.up.pt/~arocha/MIA/0506/apdz2.pdf