APS – 6º semestre – 2018/2

DESENVOLVIMENTO DE UM SISTEMA BIOMÉTRICO CAPAZ DE IDENTIFICAR ESPÉCIES DE PLANTAS A PARTIR DE SUAS FOLHAS

UNIP – Universidade Paulista

CC6P52

Integrantes:

* Carlos Eduardo de Castro – N902JC3
* Manoel de Freitas Gouvêa Junior – B999AI0
* Marcos Guilherme Afonso de Paula – N953BD7

Sumário

[Introdução, objetivo e motivação do trabalho 3](#_Toc530316804)

[Principais Técnicas Biométricas 5](#_Toc530316805)

[Biometria de plantas 6](#_Toc530316806)

[Nosso Projeto 7](#_Toc530316807)

[Classificadores 8](#_Toc530316808)

[kNN 8](#_Toc530316809)

[Sobre 8](#_Toc530316810)

[Fundamentação Matemática 8](#_Toc530316811)

[Funcionamento 8](#_Toc530316812)

[MLP (Multi Layer Perceptron) 8](#_Toc530316813)

[Sobre 8](#_Toc530316814)

[Fundamentação Matemática 8](#_Toc530316815)

[Funcionamento 8](#_Toc530316816)

[Regressão Logística 8](#_Toc530316817)

[Sobre 8](#_Toc530316818)

[Fundamentação Matemática 8](#_Toc530316819)

[Funcionamento 8](#_Toc530316820)

[Planejamento e experimentação 8](#_Toc530316821)

[Condução dos testes e Tabulação de Resultados 9](#_Toc530316822)

[kNN 9](#_Toc530316823)

[Desempenho e Avaliação de resultados 9](#_Toc530316824)

[MLP (Multi Layer Perceptron) 9](#_Toc530316825)

[Desempenho e Avaliação de resultados 9](#_Toc530316826)

[Regressão Logística 9](#_Toc530316827)

[Desempenho e Avaliação de resultados 9](#_Toc530316828)

[Conclusão e Discussão dos Resultados 10](#_Toc530316829)

[Referências Bibliográficas 11](#_Toc530316830)

[Apêndice A: Biometria na prática 11](#_Toc530316831)

[Apêndice B 13](#_Toc530316832)

Índice de Figuras

[Figura 1 - Formas de identificação biométrica para um ser humano 5](#_Toc530316797)

[Figura 2 - Exemplos de cada uma das espécies disponíveis no dataset original. 7](#_Toc530316798)

[Figura 3 - Esquema montado no software Orange®. 9](#_Toc530316799)

[Figura 4 - Exemplo de diferentes impressões digitais 11](#_Toc530316800)

[Figura 5 - Pontos que identificam uma digital 11](#_Toc530316801)

[Figura 6 - Processamento de uma digital: 12](#_Toc530316802)

[Figura 7 - Marcação dos pontos que serão comparados com o banco de dados 12](#_Toc530316803)

Índice de Tabelas

[Tabela 1 - Relação entre os modelos escolhidos para o projeto e seus respectivos nomes. 7](#_Toc530316793)

# Introdução, objetivo e motivação do trabalho

A história da Inteligência artificial teve início logo após a segunda guerra mundial com Alan Turing, conhecido por ser o pai da computação e um dos pioneiros do campo da IA, em 1956 em uma conferência no campus do Darthmouth College foi fundado o campo de pesquisa em inteligência artificial, definido como “A ciência e engenharia de produzir máquinas inteligentes”.

Porém como tudo que se é feito estritamente em campo teórico existiam problemas para implementação prática das ideias, com a inteligência artificial não foi diferente, na época de sua idealização não existia processamento suficiente para a aplicação dos conceitos, somente em 1997 a IBM® construiu um computador capaz de realizar tal nível de processamento, ele foi batizado como Deep Blue e foi o computador responsável por derrotar o enxadrista Garry Kasparov, tido como o melhor jogador de xadrez de toda a história, apesar de que o Deep Blue fosse capaz de realizar o processamento de uma IA ainda estava muito longe do que conhecemos nos dias atuais.

Indo para um momento mais recente na história, em 2011 a mesma empresa responsável pelo Deep Blue, a IBM®, pois a prova sua nova inteligência artificial, o Watson, em um jogo de perguntas e respostas chamado Jeopardy, mesmo competindo com os maiores vencedores desse jogo o supercomputador se mostrou capaz de dominar todas as rodadas do game.

Claro que a Inteligência Artificial não se limita a apenas a jogos e diversão é um campo que demonstra um grande avanço nas mais variadas áreas, podendo ser aplicada aos mais diferentes tipos de problemas, mas também ela não é capaz de resolver tudo ainda estamos em um momento de evolução onde certos problemas ainda estão fora do alcance dessa área tão promissora.

Conforme comentado anteriormente a área de inteligência artificial é muito ampla e possui tendências de crescer cada vez mais conforme os avanços científicos ocorrem, nas próximas linhas iremos comentar brevemente sobre algumas dessas áreas.

Uma das áreas mais antigas dentro de IA é o campo das buscas, que é focada em localizar soluções, respostas, em grafos de formas cada vez mais eficientes, outra área bastante famosa é a área de Machine Learning ou aprendizado de máquina, que consiste em criar algoritmos com a capacidade de aprender padrões para então realizar predições com base no conhecimento adquirido, de forma automática, outra área a citarmos é a área de PLN ou Processamento de Linguagem Natural, essa é uma área que consiste em analisar e produzir textos se baseando nas línguas faladas pelos seres humanos, como o português ou então o inglês, continuar nessas citações levaria muito tempo e seria o suficiente para uma monografia apenas para isso.

Para esse projeto utilizaremos algumas das ferramentas fornecidas pela área de Machine Learning, mais especificamente falando utilizaremos ferramentas para a solução de problemas de classificação, que são problemas cuja natureza é determinar a categoria de um certo individuo informado ao programa, um detalhe fundamental é que essas categorias já são conhecidas pelo algoritmo que irá avalia-las, e consiste em um aprendizado supervisionado, isto é, os programadores devem oferecer as respostas corretas para o algoritmo de treinamento para que então ele de forma automática extraia o conhecimento dos dados.

Indo especificamente para nosso projeto o problema consistem em classificar plantas com base em parâmetros retirados de imagens das folhas dessas plantas, nós não realizamos o tratamento dessas imagens tampouco coletamos os dados apenas utilizamos o dataset que já existe e está disponibilizado no site UCI, seu nome é [Leaf Data Set](https://archive.ics.uci.edu/ml/datasets/leaf), que é um repositório de detasets muito utilizado para a área de Machine Learning. Os dados originais foram coletados por Rubim Almeida da Silva e o dataset foi criado por Pedro F. B. Silva e André S. Marçal, todos da Universidade do Porto, Portugal.

Essa atividade consiste em realizar o treinamento e avaliação de resultados de três diferentes Inteligências Artificiais trabalhando como classificadores das informações retiradas do dataset. Os detalhes de implementação poderão ser dispensados, inclusive optamos pela utilização do software Orange®, que basicamente consiste em uma interface gráfica para realização de treinamentos, avaliação, montagem de gráficos de dispersão, geração de matrizes de confusão, entre outras opções contando com várias implementações de diferentes IA’s, como motor para tudo isso o Orange® utiliza como núcleo a linguagem python.

Neste projeto os integrantes também aproveitaram a oportunidade para utilizar ferramentas diferentes das até então conhecidas pelo grupo, além do software Orange® aproveitamos para aprimorar os conhecimentos dos integrantes na utilização da plataforma GitHub®, que é um repositório para divulgação de projetos, em sua maioria open-source. Link para o projeto: [Repositório Projeto](https://github.com/ManoelGouveaJunior/APS-SII).

# Principais Técnicas Biométricas

Por mais difícil que seja admitir um ser humano é uma criatura de hábitos e isso ocorre graças ao nosso subconsciente, pois esses pequenos hábitos nos tornam quem nós somos. Podemos perceber isso facilmente com situações normais do dia-a-dia, como por exemplo a posição de nossa cadeira favorita, caso alguém a mude de lugar, mesmo que poucos centímetros para o lado, quando a utilizamos algo nos diz que tem alguma coisa diferente e então começamos nossa busca incessante por essa diferença.

Mas esses hábitos não são apenas para nos rotularmos de chatos, maníacos entre outros adjetivos não muito agradáveis, essas manifestações do subconsciente podem ser usadas como uma das maneiras de nos identificarmos e de maneira única com relação a outras pessoas, isso é o objetivo principal quando o assunto é biometria, que em termos mais técnicos se refere ao estudo das características físicas e comportamentais de um ser.

Muitos quando ouvem essa palavra, biometria, logo associam com as digitais de nossas mãos, o que não está errado, porém essa não é a única forma de realizar a identificação única de uma pessoa apenas a mais conhecida.

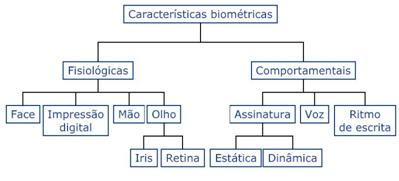


Figura 1 - Formas de identificação biométrica para um ser humano

Para classificar um meio de identificação bom o suficiente algumas características devem existir: Universidade, Singularidade, Permanência, Mensurabilidade que de forma resumida descrevem características que sejam válidas para qualquer outro ser de mesma espécie, únicas, permanentes e mensuráveis.

Mas uma coisa é certa por mais variados que sejam as formas de avaliar a biometria de um ser o processo entre qualquer um dos meios de avaliação é bem semelhante, de forma bem básica consiste no processamento de uma imagem captada por um scanner com um alto grau de resolução e que faça a eliminação de “ruídos” desnecessários para o método específico, apenas para ilustrar podemos pegar o exemplo de um scanner que registra o padrão de vasos e circulação de uma pessoa através do escaneamento de sua mão, é um sistema que deve avaliar as veias e artérias da mão da pessoa portanto é irrelevante a cor da pele dessa pessoa, por esse motivo o scanner que realiza essa atividade não basta ser uma simples câmera que capta a luz refletida pela pele dessa pessoa, isto é, sua cor de pele.

O segundo passo realizado pelos sistemas biométricos é compara a imagem registrada com uma imagem armazenada em um banco de dados, se as imagens forem iguais ou então semelhantes até certo grau de aceitação a operação é autorizada, em caso negativo a operação é rejeitada.

O método de avaliação biométrica está diretamente relacionado com o custo e também com o nível de segurança oferecido, por exemplo a autenticação através de digital é simples e de baixo custo, porém é relativamente simples de ser burlado, enquanto um método que verifique o padrão da retina é um dos meios mais confiáveis, porém também é um dos meios mais custosos em termos financeiros e de processamento computacional, ao final deste documento o apêndice A trata um pouco mais sobre o assunto de como é feito a avaliação de biometria na prática.

## Biometria de plantas

Até agora falamos muito sobre biometria de forma geral com um pouco mais de atenção para o modo como é feito para seres humanos, porém esse trabalho trata-se de comparar e classificar plantas, então nessa seção discutiremos um pouco sobre como avaliar a biometria de uma planta.

Na documentação fornecida junto ao Leaf Dataset no UCI podemos verificar vários pontos que podem ser comparados entre as diferentes espécies de plantas, mais especificamente são 14 pontos que foram comparados na pesquisa original, alguns desses pontos são:

* Eccentricity (Excentricidade): Diz respeito à excentricidade da elipse que compõe o formato da folha, quanto maior essa excentricidade, mais estreito é o formato da folha, esses valores são medidos no intervalo entre 0 e 1.
* Elongation (Prolongamento): O prolongamento é obtido através da fórmula **1-2dmax/D** onde a relação de 2dmax/D é a razão entre o maior círculo inscrito e o menor círculo circunscrito.
* Solidity (Solidez): Determina o quão bem o formato da folha se adequa à uma forma convexa.

Através do cumprimento de um processo idêntico para a captura das imagens de todos, e para todos os 340 exemplares que compões o dataset foi feito o processamento dessas imagens e extraído os valores informados na tabela que está junto com os dados no repositório do UCI.

# Nosso Projeto

Neste projeto não será necessário a aquisição dos dados, pois eles já estão disponíveis e foram adquiridos de maneira confiável. Nosso trabalho consiste em escolher alguns dos exemplares disponíveis e aplicar métodos de classificação para a aquisição de conhecimento, no caso com técnicas voltadas em aprendizado automático, isto é, Machine Learning.

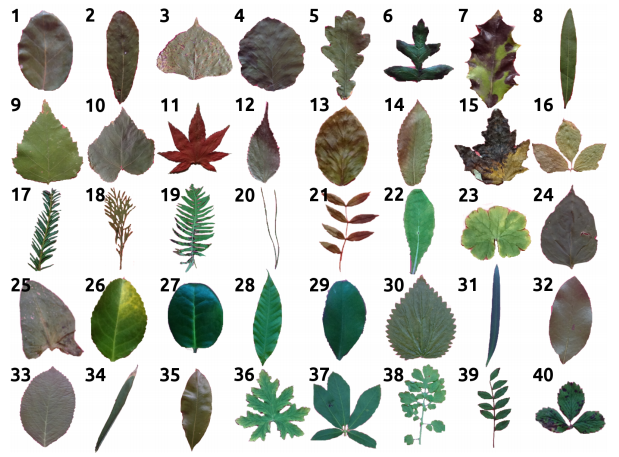


Figura 2 - Exemplos de cada uma das espécies disponíveis no dataset original.

Nós escolhemos alguns dos tipos de folhas de maneira aleatória e alguns foram escolhidos pela semelhança com algum outro modelo justamente para tentar provocar algum tipo de “confusão” nas inteligências artificiais.

|  |  |
| --- | --- |
| **Número** | **Nome Científico** |
| 2 | *Salix atrocinera* |
| 5 | *Quercus robur* |
| 6 | *Crataegus monogyna* |
| 10 | *Tilia tomentosa* |
| 22 | *Primula vulgaris* |
| 26 | *Euonymus japonicus* |
| 27 | *Ilex perado ssp. azorica* |
| 28 | *Magnolia soulangeana* |
| 30 | *Urtica dioica* |
| 34 | *Pseudosasa japonica* |

Tabela 1 - Relação entre os modelos escolhidos para o projeto e seus respectivos nomes.

A continuação de como foram realizados os testes estão no tópico “Planejamento e experimentação”.

# Classificadores

## kNN

### Sobre

### Fundamentação Matemática

### Funcionamento

## MLP (Multi Layer Perceptron)

### Sobre

### Fundamentação Matemática

### Funcionamento

## Regressão Logística

### Sobre

### Fundamentação Matemática

### Funcionamento

# Planejamento e experimentação

Primeiramente separamos os dados que escolhemos utilizar de todos os dados disponíveis, esse passo resultou na criação de uma planilha com 17 colunas e 113 linhas de dados, após isso foi colocado um cabeçalho que descreva os dados que cada coluna possui. Nos dados originais existiam 16 colunas, porém para fins de melhorar a visualização foi acrescentado uma coluna com os nomes de cada valor classe.

Após a separação dos dados a planilha foi formatada de tal maneira que possibilitasse que o software Orange pudesse realizar a importação e identificação automática das colunas de atributos e classe. Essa formatação foi o acréscimo de 2 linhas que contém quais são os tipos de dados e a informação das palavras “class” e “meta” para identificar os nomes das classes e quais os valores alvo, target.

Foi feita a importação dos dados através da opção F**ile,** para visualização dos dados foram utilizados os itens **Data Table** e **Scatter Plot,** que permitem ver os dados de uma forma tabular com a marcação do que é um atributo e uma classe e também os dados distribuídos no formato de gráficos de dispersão, respectivamente.

Dentro dos modelos disponíveis selecionamos três que são os utilizados para os testes, são eles: **kNN, Rede Neural e Regressão Logística,** uma observação a se fazer é que no caso das redes neurais o Orange trabalha com o modelo MLP (Multi Layer Perceptron) então todas as observações feitas nesse tópico serão referentes a esse método de classificação.

Por fim utilizamos meios de avaliação chamados **Test & Score** e **Confusion Matrix,** os dados fornecidos por essas opções serão utilizados na avaliação de desempenho. A figura 3 mostra o esquema montado no Orange.

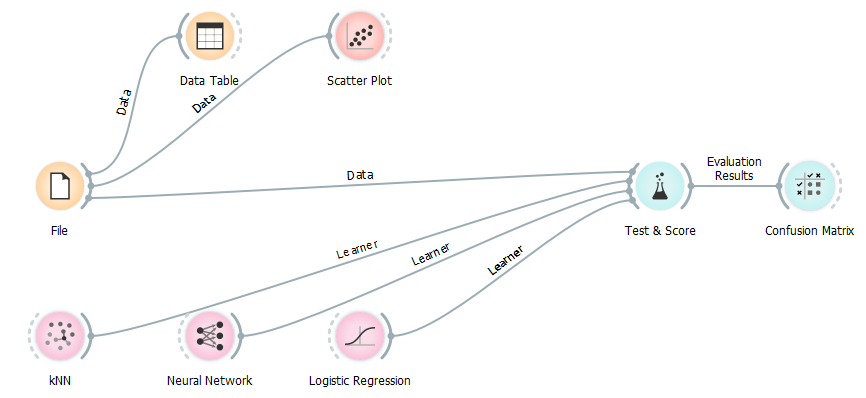


Figura 3 - Esquema montado no software Orange®.

# Condução dos testes e Tabulação de Resultados

## kNN

### Desempenho e Avaliação de resultados

## MLP (Multi Layer Perceptron)

### Desempenho e Avaliação de resultados

## Regressão Logística

### Desempenho e Avaliação de resultados

# Conclusão e Discussão dos Resultados

# Referências Bibliográficas

# Apêndice A: Biometria na prática

No site [linhadecódigo](http://www.linhadecodigo.com.br/artigo/1162/biometria-processamento-de-imagens-capturadas-em-leitores-de-impressao-digital.aspx) existe um tutorial de como realizar a implementação prática de um sistema com alto grau de confiabilidade de uma impressão digital, como este não é o foco de nosso projeto iremos abordar brevemente as etapas do processo de biometria.

Primeiramente é necessário a extração dos dados que serão avaliados, no caso imagens de uma impressão digital, isso pode ser obtido por sensores especializados nesse tipo de informação ou então com o escaneamento de uma folha que contenha a impressão digital carimbada.

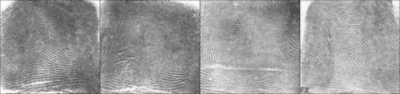


Figura 4 - Exemplo de diferentes impressões digitais

Existem alguns pontos importantes nas imagens extraídas que são justamente os pontos observados pelo sistema automatizado ou então por um papiloscopista, que é o profissional especialista na identificação de impressões digitais.

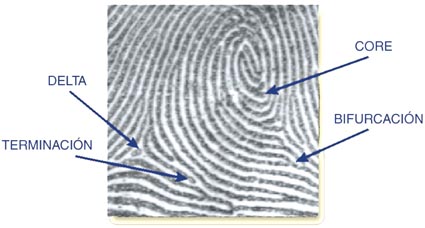


Figura 5 - Pontos que identificam uma digital

Então o processo se resume em obtenção da imagem da impressão digital, processamento dessa imagem através da aplicação do filtro de Gabor, binarização, afinamento das linhas, por fim é feita a avaliação da imagem.

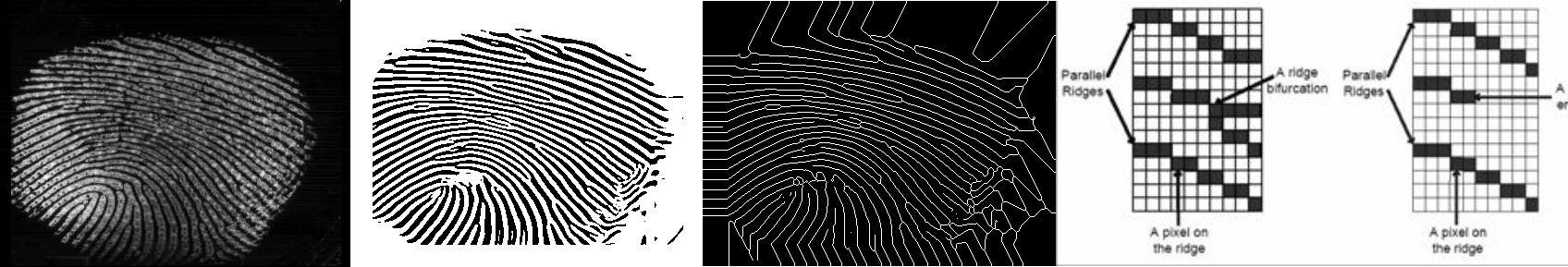


Figura 6 - Processamento de uma digital:

Passos descritos pela imagem acima:

1. Leitura da imagem pelo sensor;
2. Aplicação do filtro de Gabor;
3. Binarização e afinamento das linhas para 1 px de espessura;
4. Avaliação dos pontos importantes através da análise dos pixels;

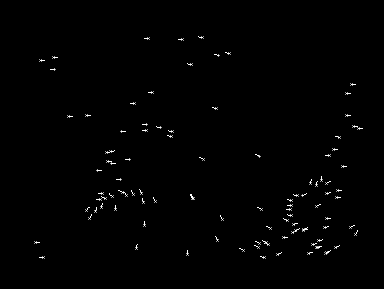


Figura 7 - Marcação dos pontos que serão comparados com o banco de dados

Após a comparação com o banco de dados é realizada a liberação ou bloqueio da pessoa de acordo com o comportamento descrito para o sistema em questão.

# Apêndice B