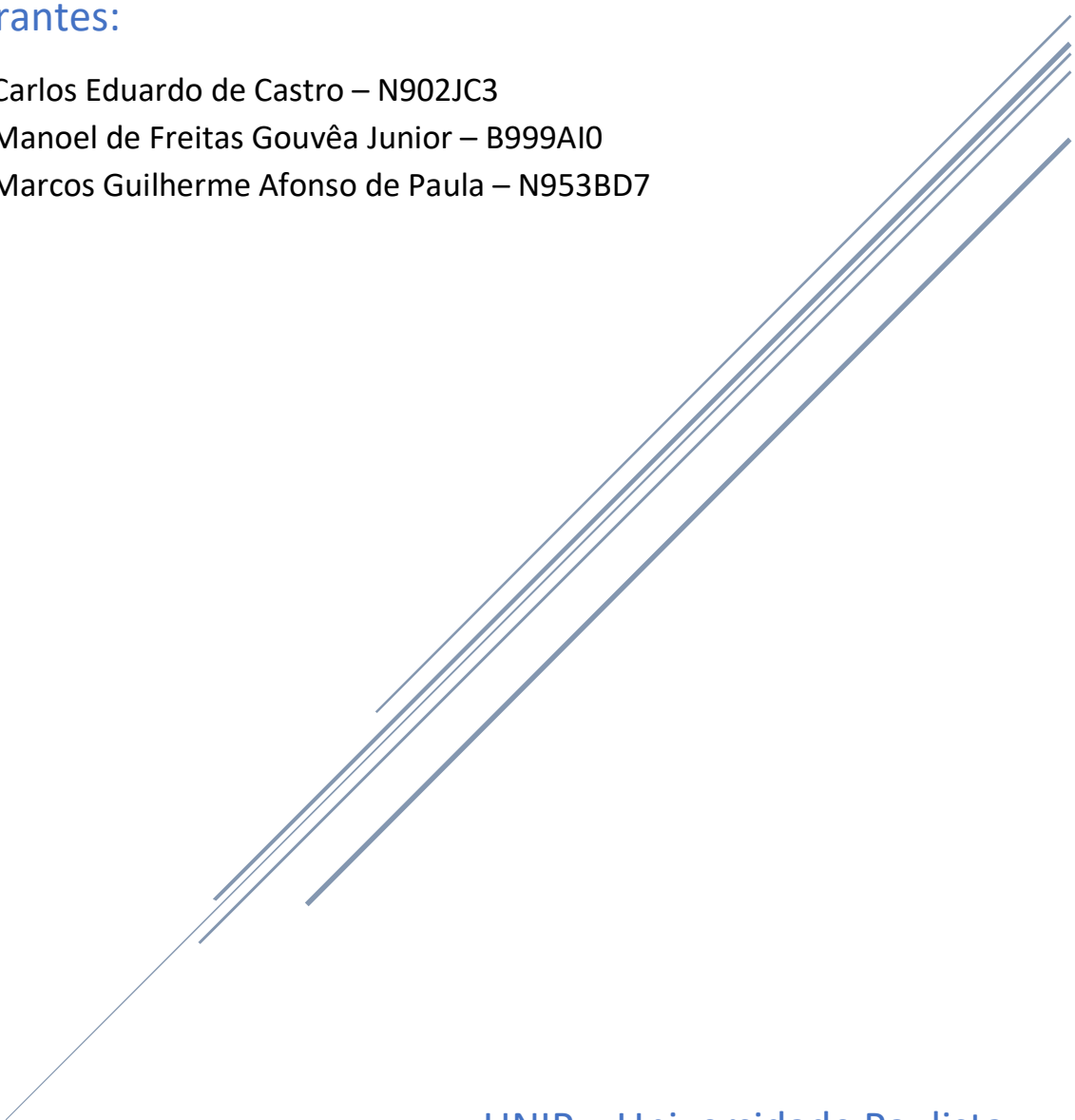
APS – 6º SEMESTRE – 2018/2

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

Integrantes:

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



UNIP – Universidade Paulista

CC6P52

Sumário

Índice de Figuras

Índice de Tabelas

Índice de Equações

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®, colocou 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, 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.

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.

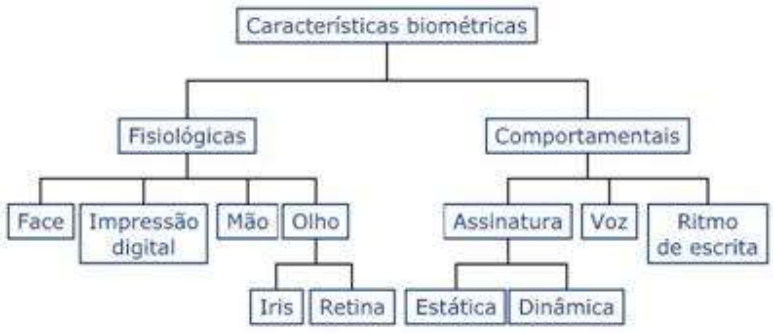


Figura 1 - Formas de identificação biométrica para um ser humano

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

Nesta seção iremos discutir brevemente sobre particularidades, princípios matemáticos e a forma de funcionamento dos classificadores trabalhados nesse projeto, que para conhecimento são os algoritmos kNN, MLP e Regressão Logística.

kNN

Sobre

O classificador KNN ou K-Nearest Neighbors, do português k-vizinhos mais próximos, é um dos mais conhecidos e fáceis de usar e seu uso deve sempre ser considerado na hora de se usar uma IA para classificação. O KNN determina a classificação de novas instancias de acordo com a "distância" de similaridade entre elas, isto é, para cada novo elemento, o kNN procura os k-vizinhos mais próximos em sua base de treino e verifica qual ocorre com mais frequência, então o novo elemento é classificado com essa classe. Ao se escolher o tipo de projeto alguns fatores devem ser considerados como os atributos que serão utilizados para a classificação, uma vez que se os dados forem muito dispersos o kNN pode gerar erros e causar uma distorção dos cálculos e assim gerando erros indesejáveis.

Imagine que queremos classificar uma raça de cães, sabemos que cada um tem sua peculiaridade, como tamanho, cor, comprimento dos pelos, forma dos pelos entre outras características. Esse são tipos de atributos bons para utilizar em nosso classificador. Valores comuns que podem gerar confusão em nosso classificador não deve ser levado em conta, um exemplo é se o cachorro tem pelo ou não ou se é macho ou fêmea, sabendo que a maioria desses animais possuem pelos, e são machos ou fêmeas.

Fundamentação Matemática

O KNN usa como métrica o cálculo da distância, a Euclidiana é a mais conhecida e usualmente utilizada, mas existem outras como por exemplo, a distância Manhattan.

$$d = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

Equação 1 - Cálculo da Distância Euclidiana

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Equação 2 - Cálculo da Distância Manhattan (City Block)

Funcionamento

A funcionalidade do KNN é dada classificando uma nova entrada baseada nos k-vizinhos mais próximos, ou seja, baseado na similaridade do que ocorre com mais frequência, em relação a nova entrada.

Um passo importante quando trabalhamos com o kNN é que se caso os dados não estejam normalizados o cálculo das distâncias irá apresentar falsos positivos pois por não estarem devidamente distribuídos um atributo pode gerar um peso maior para a avaliação de um indivíduo.

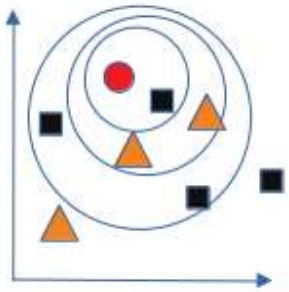


Figura 4 - Exemplo do funcionamento do kNN

Conforme ilustrado na Figura 5 o valor do k é de extrema relevância, pois dependendo desse valor a resposta será diferente em cada execução, seguindo o exemplo da figura 5 temos que se o valor de k for 1, o círculo vermelho, que é nosso indivíduo que está sendo classificado, será categorizado como um quadrado preto, entretanto se o valor de k for 3 ele passará a ser classificado como um triangulo laranja, já que para k=3 triângulos laranja é o que ocorre com maior frequência na vizinhança do novo indivíduo, o mesmo vale para k=5 onde o individuo seria classificado como um quadrado preto novamente.

Uma observação com relação à escolha de k é que para evitar impasses é interessante adotar um valor para k que seja um número primo ou então pelo menos um número par, claro que isso não é uma regra, mas é uma boa prática interessante de ser seguida.

MLP (Multi Layer Perceptron)

Sobre

Assim como outras técnicas a área de Machine Learning também possui meios inspirados na biologia dos seres humanos, e nesse caso para a classificação foi escolhido o método das Redes Neurais Artificias (RNA), mais precisamente dizendo foi utilizado a rede MLP (Multi Layer Perceptron) para classificação das espécies de plantas utilizadas nesse projeto. De forma geral uma RNA se baseia no funcionamento do cérebro humano onde cada célula é comparada e chamada de neurônio.

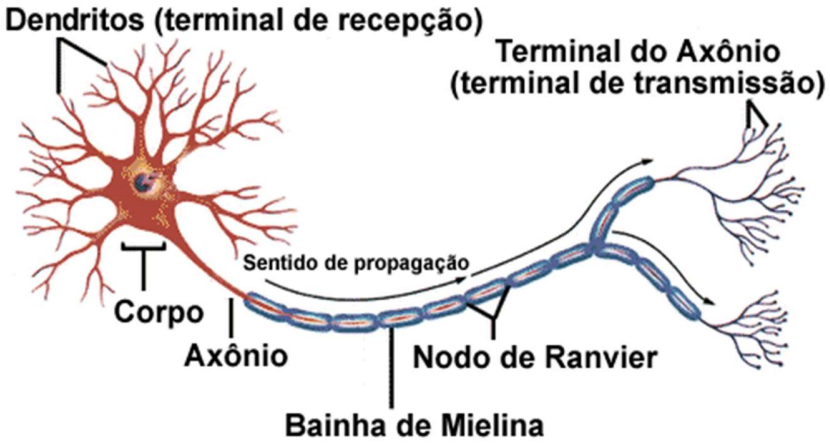


Figura 3 - Modelo de um Neurônio humano e suas partes

É claro que não é possível criar um neurônio computacionalmente, para isso existe uma modelagem matemática para descrever o funcionamento desse neurônio. Enquanto temos como um exemplo simples o Perceptron sendo representado por apenas um neurônio as MLP, como o nome sugere, são vários neurônios interligados em várias camadas, usualmente duas camadas para o processamento e um último neurônio para realizar a saída da rede.

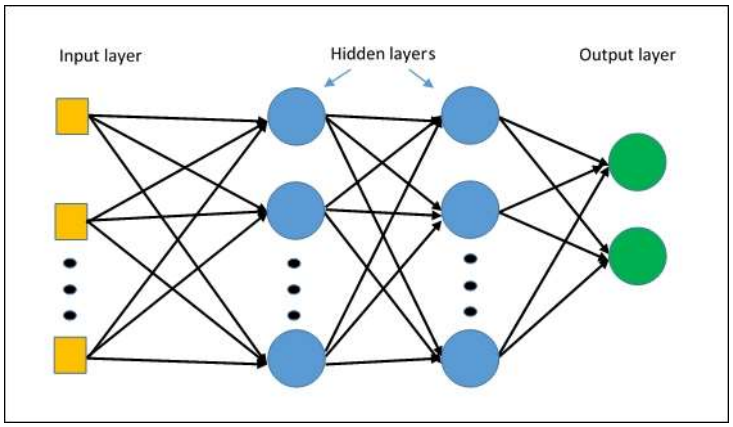


Figura 4 - Exemplo de uma MLP

Fundamentação Matemática

Enquanto Perceptrons são muito bons em classificar elementos linearmente separáveis, isto é, que se realizarmos a distribuição dos indivíduos de treinamento em um gráfico é possível dividir esses indivíduos através de uma reta, por outro lado em conjuntos mais complexos que não são linearmente separáveis as MLP's ganham seu espaço.

Habitualmente utiliza-se da função sigmoide para realizar o treinamento das redes, já que se utilizarmos uma função linear o funcionamento será o mesmo que um perceptron, também é possível utilizar outras funções além da sigmoide, como a Tanh ou Gaussiana.

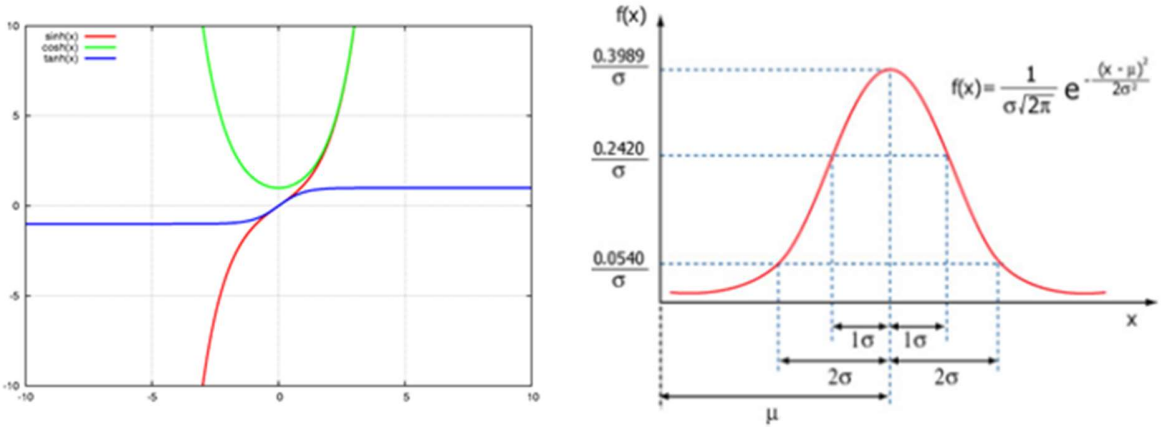


Figura 5 - Funções Tanh e Gaussiana

Funcionamento

Novamente interligando a ideia com o funcionamento de um Perceptron, que nada mais é que uma MLP de uma camada, o funcionamento se dá em duas partes, a primeira para realizar o treinamento da rede e a segunda de fato para o funcionamento e aplicação desse treinamento.

Para isso é utilizado um "Neurônio Artificial", onde temos como entradas (Xn) os valores dos atributos de cada indivíduo, então para cada atributo é atribuído um peso (Wn), também temos a inclusão de uma entrada chamada de bias utilizada para fins de correções (usualmente é atribuído o X0 para a entrada do bias e o peso W0), então é realizado o somatório da multiplicação de todas as entradas pelos seus respectivos pesos, então esse valor é passado por uma função de ativação ou transferência, e determinado a saída do neurônio. Apesar dessa ser a explicação para um neurônio, na MLP esse raciocínio é o mesmo, porém para vários neurônios onde a saída de um muitas vezes é a entrada de outros.

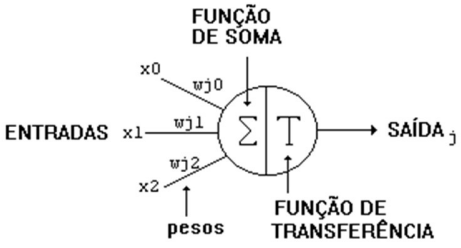
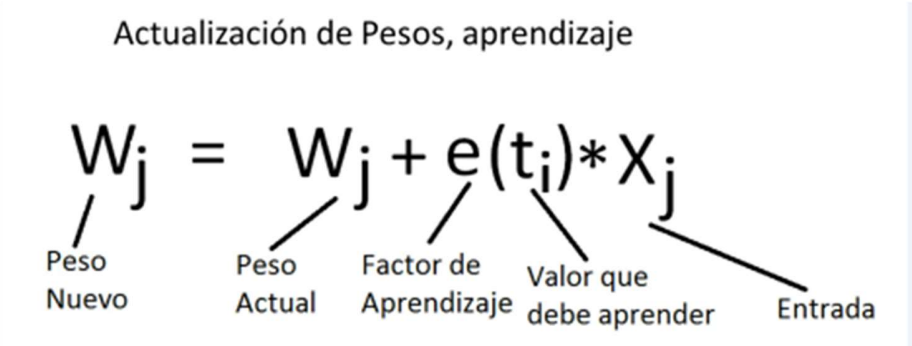


Figura 6 - Exemplo de um "Neurônio Artificial"

Ao final desse processo é determinado o erro da rede, se houver algum erro diferente de 0 será necessário o reajuste dos pesos, para isso é utilizado o processo de backpropagation, onde são determinados o quanto cada neurônio contribuiu para o erro total. Após isso é feito o reajuste dos pesos através da equação 1.



Equação 3 - Equação para reajuste de pesos.

Esse processo se repete até que não exista mais erro nenhum para todos os casos de teste ou então pelo número de iterações (épocas) definidos. Depois disso o processo de treino foi concluído e a rede está pronta para realizar a classificação de novos indivíduos.

Regressão Logística

Sobre

A regressão logística é uma técnica estatística também conhecida como classificador de máxima entropia, que a partir de análise de um conjunto de dados tenta prever valores de uma variável de saída que possui a natureza categórica, em outras palavras a partir de um conjunto de dados tenta classificar novos elementos apresentados, comumente é comparada com a regressão linear e de fato ambas possuem muito em comum, a grande diferença está na forma de interpretação, uma vez que na regressão linear o objetivo é encontrar uma relação linear entre duas variáveis, a partir de dois dados tenta-se estabelecer uma reta que seja capaz de dividir esses dados em grupos classificatórios. A regressão logística possui um núcleo idêntico ao da regressão linear, entretanto ao invés de tentar achar uma divisão através de uma reta ela tenta criar uma divisão através de uma curva sigmoide, isto é, uma curva com um formato de 'S'. A figura 10 mostra a comparação entre uma regressão linear e uma regressão logística.

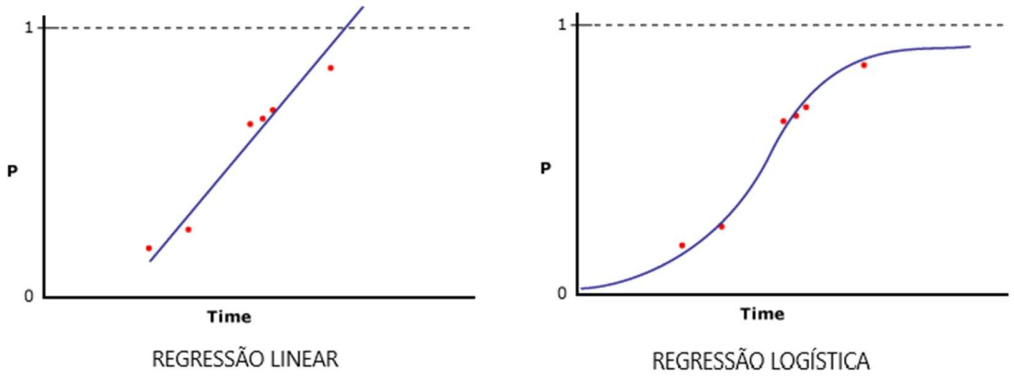


Figura 7 - Comparação entre as Regressões Linear e Logística

Fundamentação Matemática

Resumindo toda a explicação matemática a regressão logística tenta calcular probabilidades e dessa forma tenta realizar inferências, isso ocorre através do cálculo de Logit.

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Equação 4 - Função Logit

Como podemos perceber na função de Logit da Equação 4 não consideramos o erro a parte como nos cálculos de um Perceptron, por exemplo, que por sinal é um modelo que é baseado na regressão linear.

Realizando todas as transformações matemáticas podemos isolar a probabilidade, que é representada pela letra P na Equação 1, o resultado podemos ver na Equação 5.

$$p = \frac{1}{1 + e^{-logit(p)}}$$

Equação 5 - P isolado na função Logit

Funcionamento

Conforme mencionado no tópico anterior tudo se resume em um cálculo de probabilidade, ou seja, são feitos testes com os dados de atributos para determinar a probabilidade de um dos indivíduos ser ou não ser de uma certa classe, isso se dá por causa que a regressão logística é uma técnica binária, então os testes são repetidos para todas as classes até que se encontre a classe que apresenta a melhor classificação.

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 4 mostra o esquema montado no Orange.

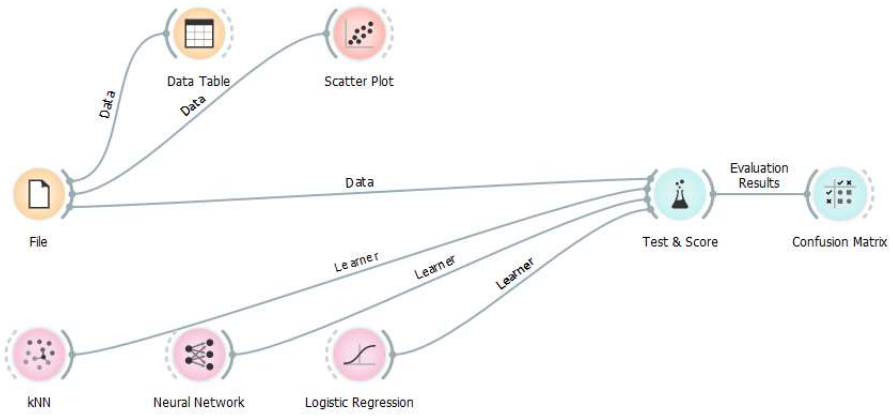


Figura 8 - Esquema montado no software Orange®.

Condução dos testes e Tabulação de Resultados

Cada integrante ficou responsável por avaliar o desempenho de um classificador o modelo utilizado apesar de ser o mesmo foi aplicado em computadores diferentes, com sistemas operacionais diferentes e com componentes de hardware diferentes, portanto apesar de que os valores propriamente ditos poderem sofrer alterações iremos utilizar essas divergências para também realizar a comparação se o ambiente de aplicação dos classificadores influência nos resultados finais, abaixo encontram-se os tópicos de cada classificador e as observações realizadas pelos integrantes.

kNN

Desempenho e Avaliação de resultados

Para realizar as mudanças de parâmetro do kNN o Orange® oferece como opção de parâmetro o valor de k e a métrica de distância, e para o teste foi utilizado um k com o valor de 5 e a distância euclidiana na parte de métrica.

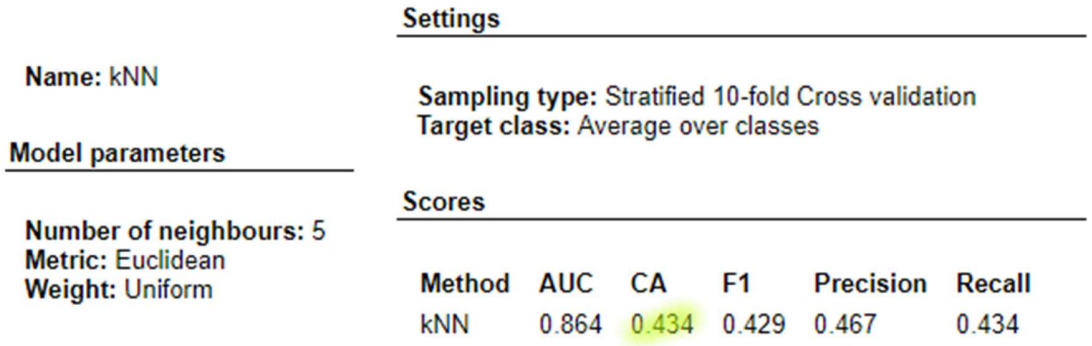


Figura 9 - Resultados obtidos para o kNN

Como podemos notar o kNN teve um desempenho não muito bom tendo como precisão de acerto apenas 43,4%. Agora vamos conferir a matriz de confusão para algumas considerações extras.

		Predição										
Real		**6**	**26**	**27**	**28**	**22**	**34**	**5**	**2**	**10**	**30**	***∑***
	6	8	0	0	0	0	0	0	0	0	0	***8***
	26	0	4	4	0	1	0	0	1	1	1	***12***
	27	0	7	3	0	0	0	0	1	0	0	***11***
	28	0	2	4	4	1	0	0	1	0	0	***12***
	22	2	1	1	1	3	0	2	2	0	0	***12***
	37	0	0	0	0	0	11	0	0	0	0	***11***
	5	0	1	3	0	0	0	5	2	1	0	***12***
	2	0	2	2	0	4	0	0	2	0	0	***10***
	10	0	2	1	0	0	0	0	0	9	1	***13***
	30	0	4	3	0	0	0	0	0	5	0	***12***
	∑	***10***	***23***	***21***	***5***	***9***	***11***	***7***	***9***	***16***	***2***	

Tabela 2 - Matriz de Confusão do kNN

Analisando a matriz de confusão podemos perceber que a classificação apesar de ter cometido muitos erros eles ficaram concentrados em algumas espécies, essas foram "*Ilex perado ssp. azorica* " (espécie de número 27) e "*Euonymus japonicus*" (Espécie de número 26), mas o algoritmo não apenas cometeu erros também houve um caso onde ele foi capaz de acertar todas as espécies sem confundi-la com nenhuma outra, esse foi o caso da "*Pseudosasa japonica*" (Espécie de número 34).

Os erros cometidos pelo kNN apesar de muitos são aceitáveis para o caso deste projeto, uma vez que os atributos fornecidos estão quase sempre gerando gráficos de dispersão muito misturados torna possível a confusão de espécies parecidas como é o caso das plantas 26 e 27, e o mesmo se aplica à planta 34, afinal ela possui uma aparência bastante própria com relação as demais utilizadas, portanto é natural que a classificação dela ocorra de uma forma mais certa.

Figura 10 - Comparação entre as plantas de número 26 (esquerda) e 27 (direita)

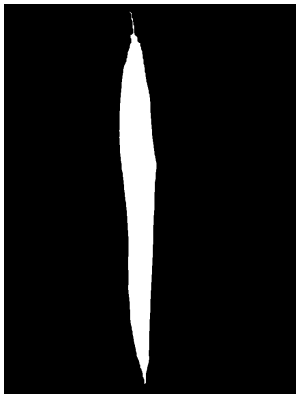


Figura 11 - Espécie 34 (Pseudosasa japonica)

MLP (Multi Layer Perceptron)

Dentre os métodos utilizados nesse projeto a MLP no Orange® é uma das que mais possuem parâmetros configuráveis para sua aplicação, são eles:

- Hidden Layers: Correspondem a quantidade de neurônios existentes na camada oculta da rede.
- Activation: Representa a função de ativação dos neurônios, as opções dessa opção são função identidade, logística, tanh, ReLu.
- Alpha: Correspondente à taxa de aprendizado, apesar de que em literaturas e dados teóricos essa variável seja representada, geralmente, pela letra grega η (eta) em aplicações práticas costuma-se nomeá-la de alpha (α).
- Max iterations: Número de épocas que serão usadas no treinamento.

Desempenho e Avaliação de resultados

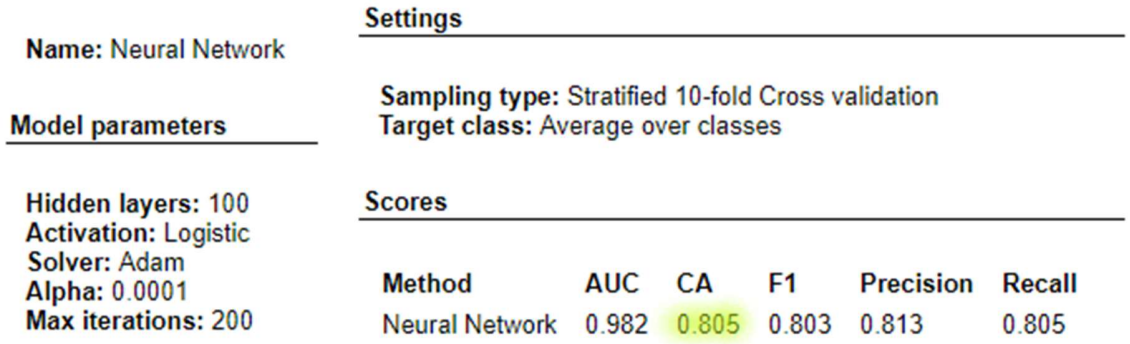


Figura 12 - Teste 1: Configuração padrão Orange®

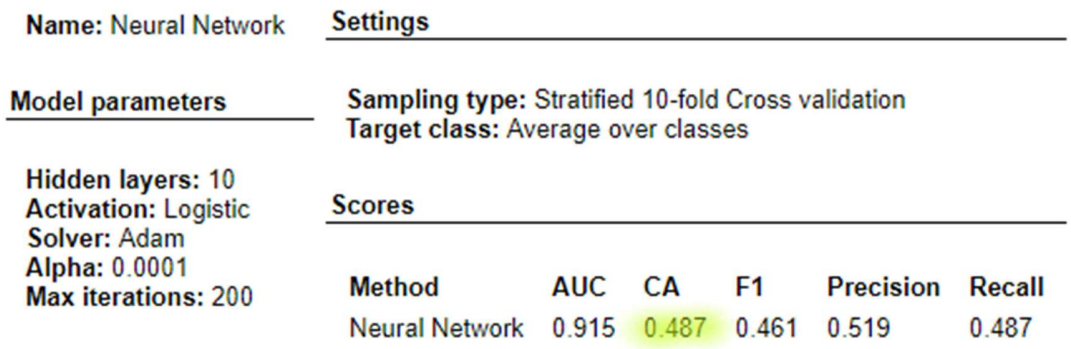


Figura 13 - Teste 2: Alteração do Hidden para 10

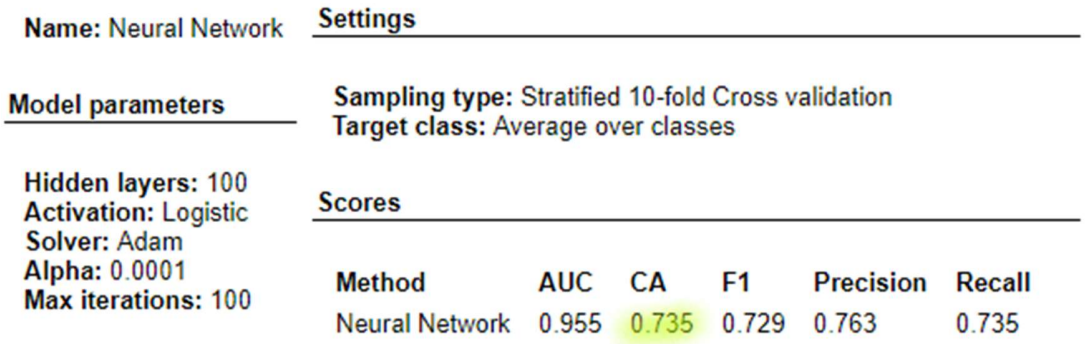


Figura 14 - Teste 3: Alteração do número de épocas para 100

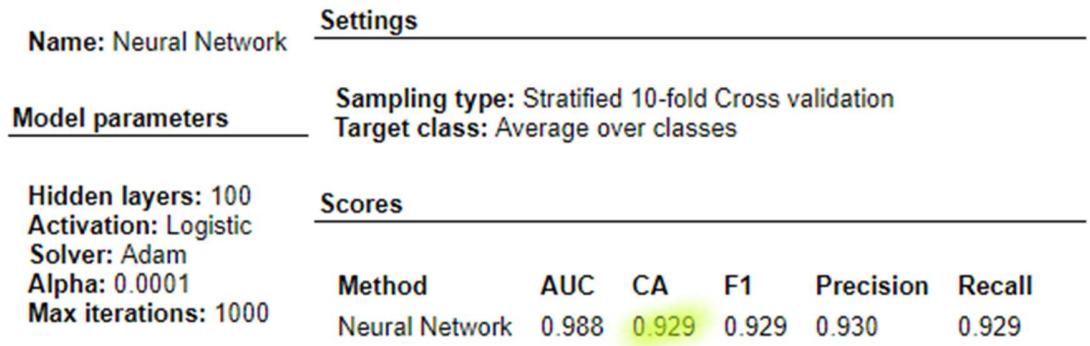


Figura 15 - Teste 4: Alteração do número de épocas para 1000

Acima foram apresentados os testes que apresentaram a maior variação dos resultados de acordo com a alteração de parâmetros, como podemos observar nas configurações padrão do Orange® (Teste 1) obtivemos uma acurácia de 80,5%, enquanto ao alterar apenas a quantidade de neurônios na camada interna da rede MLP tivemos uma queda brusca nesse valor de acurácia 48,7% (Teste 2).

De maneira semelhante com o aumento do número de épocas para 1000 tivemos um grande ganho de precisão 92,9% (Teste 5), um ganho de 7,4% de acurácia.

Apenas com a informação da acurácia não dá para confirmar se os erros cometidos pela RNA são erros aceitáveis para o ambiente dos nossos testes, para isso abaixo serão colocadas as matrizes de confusão respectivas a cada teste para uma melhor avaliação.

		Predição										
		6	26	27	28	22	34	5	2	10	30	∑
Real	6	8	0	0	0	0	0	0	0	0	0	8
	26	0	6	2	0	0	0	0	3	0	1	12
	27	0	1	9	0	0	0	0	1	0	0	11
	28	0	0	0	10	1	0	0	1	0	0	12
	22	1	0	0	1	7	0	1	2	0	0	12
	37	0	0	0	0	0	11	0	0	0	0	11
	5	0	0	0	0	1	0	11	0	0	0	12
	2	0	1	0	0	0	0	0	9	0	0	10
	10	0	0	0	0	0	0	0	0	12	1	13
	30	0	2	0	0	0	0	0	0	2	8	12
	∑	9	10	11	11	9	11	12	16	14	10	

Tabela 3 - Matriz de Confusão Teste 1

		Predição										
		6	26	27	28	22	34	5	2	10	30	∑
Real	6	3	0	0	0	0	0	5	0	0	0	8
	26	0	1	6	0	0	1	0	1	2	1	12
	27	0	4	5	0	0	0	0	2	0	0	11
	28	0	0	0	6	0	2	0	0	4	0	12
	22	1	1	0	1	2	3	2	1	1	0	12
	37	0	0	0	0	0	11	0	0	0	0	11
	5	0	0	0	0	1	1	7	0	3	0	12
	2	0	2	2	1	0	1	0	4	0	0	10
	10	0	0	0	1	0	0	0	0	11	1	13
	30	0	0	1	0	0	0	0	0	6	5	12
	∑	4	8	14	9	3	19	14	8	27	7	

Tabela 4 - Matriz de Confusão Teste 2

		Predição										
		6	26	27	28	22	34	5	2	10	30	∑
Real	6	8	0	0	0	0	0	0	0	0	0	8
	26	0	6	2	1	0	0	0	2	1	0	12
	27	0	6	5	0	0	0	0	0	0	0	11
	28	0	0	0	9	0	0	0	2	1	0	12
	22	2	0	0	2	6	1	0	1	0	0	12
	37	0	0	0	0	0	11	0	0	0	0	11
	5	0	0	0	0	0	0	12	0	0	0	12
	2	0	1	1	0	0	0	0	8	0	0	10
	10	0	0	0	0	0	0	0	0	12	1	13
	30	0	2	0	0	0	0	0	0	4	6	12
	∑	10	15	8	12	6	12	12	13	18	7	

Tabela 5 - Matriz de Confusão Teste 3

		Predição										
		6	**26**	**27**	**28**	**22**	**34**	**5**	**2**	**10**	**30**	***∑***
Real	**6**	8	0	0	0	0	0	0	0	0	0	***8***
	26	0	10	0	0	1	0	0	1	0	0	***12***
	27	0	1	9	0	0	0	0	1	0	0	***11***
	28	0	0	0	12	0	0	0	0	0	0	***12***
	22	0	0	0	0	12	0	0	0	0	0	***12***
	37	0	0	0	0	0	11	0	0	0	0	***11***
	5	0	0	0	0	0	0	12	0	0	0	***12***
	2	0	0	2	0	0	0	0	8	0	0	***10***
	10	0	0	0	0	0	0	0	0	12	1	***13***
	30	0	1	0	0	0	0	0	0	0	11	***12***
	∑	***8***	***12***	***11***	***12***	***13***	***11***	***12***	***10***	***12***	***12***	

Tabela 6 - Matriz de confusão Teste 4

De uma forma geral a rede MLP atendeu muito bem ao proposto por esse trabalho, cometendo alguns erros, mas de forma geral acertando a maior parte das situações, e também pudemos notar que o desempenho teve uma grande melhora a partir da alteração da variável responsável pela quantidade de iterações, enquanto uma diminuição do número de épocas pela metade, com relação as configurações padrão do Orange®, já obtivemos uma diminuição considerável na precisão dos acertos, e em contrapartida com o aumento para 1000 iterações causou um grande aumento no número de acertos, entretanto esse aumento apesar de considerável também teve um custo, esse custo foi o tempo de execução que apesar de não ter sido contabilizado apresentou um aumento considerável, mas de forma geral a rede MLP conseguiu lidar muito bem com os dados apesar de possuírem uma distribuição não muito simples.

Regressão Logística

No Orange® a regressão logística possui apenas um parâmetro de configuração chamado “Strenght” porém para nosso caso não houve mudanças nos resultados com a variação desse parâmetro, o que resultou em alguma mudança foi a forma de distribuição dos dados de entrada, isto é o método de divisão entre dados de teste e de treino, obtivemos dois resultados basicamente, um para uma divisão estratificada com 10-fold, 10 pastas, na divisão dos dados e outra para 5-fold, abaixo estão os resultados obtidos e uma breve discussão sobre esses dados.

Desempenho e Avaliação de resultados

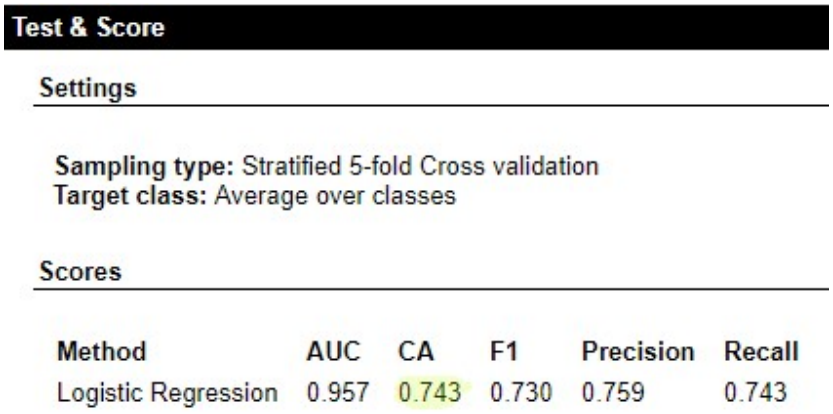


Figura 16 - Resultados 5-Fold

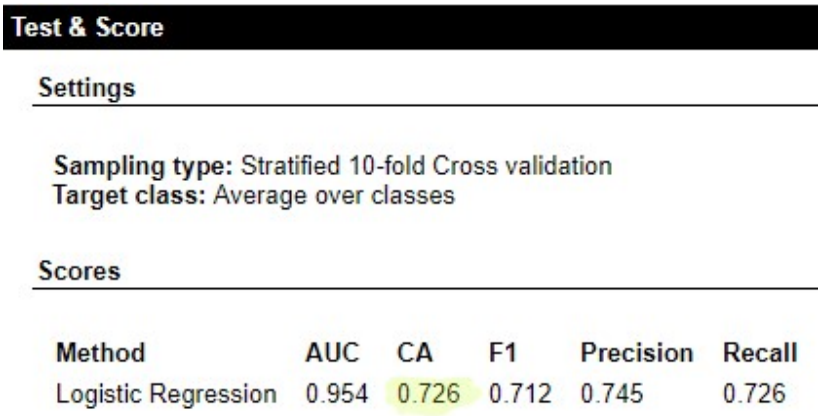


Figura 17 - Resultados 10-Fold

Como podemos observar tivemos um breve aumento na acurácia com a utilização da divisão da base em 5-Fold (5-Fold = 74,3% / 10-Fold = 72,6%) com validação cruzada e dados estratificados. Agora iremos olhar para as matrizes de confusão de cada um dos métodos acima.

		Predição										
		6	**26**	**27**	**28**	**22**	**34**	**5**	**2**	**10**	**30**	***∑***
Real	**6**	8	0	0	0	0	0	0	0	0	0	***8***
	26	0	7	1	0	0	0	1	2	0	1	***12***
	27	0	4	6	1	0	0	0	0	0	0	***11***
	28	0	0	0	10	0	0	2	0	0	0	***12***
	22	2	0	0	0	7	0	0	3	0	0	***12***
	37	0	0	0	0	0	11	0	0	0	0	***11***
	5	0	0	0	0	0	0	12	0	0	0	***12***
	2	0	1	0	1	2	0	0	6	0	0	***10***
	10	0	0	0	0	0	0	0	0	13	0	***13***
	30	0	2	0	0	0	0	0	0	6	4	***12***
	∑	***10***	***14***	***7***	***12***	***9***	***11***	***15***	***11***	***19***	***5***	

Tabela 7 - Matriz de Confusão para 5-Fold

		Predição										
		6	**26**	**27**	**28**	**22**	**37**	**5**	**2**	**10**	**30**	***∑***
Real	**6**	8	0	0	0	0	0	0	0	0	0	***8***
	26	0	8	1	0	0	0	1	1	0	1	***12***
	27	0	5	5	1	0	0	0	0	0	0	***11***
	28	0	0	0	9	0	1	2	0	0	0	***12***
	22	2	0	0	0	7	0	0	3	0	0	***12***
	34	0	0	0	0	0	11	0	0	0	0	***11***
	5	0	0	0	0	0	0	11	1	0	0	***12***
	2	0	1	0	1	2	0	0	6	0	0	***10***
	10	0	0	0	0	0	0	0	0	13	0	***13***
	30	0	2	0	0	0	0	0	0	6	4	***12***
	∑	***10***	***16***	***6***	***11***	***9***	***12***	***14***	***11***	***19***	***5***	

Tabela 8 - Matriz de Confusão para 10-Fold

Com as matrizes de confusão podemos verificar algo que já poderia ter sito previsto de acordo com as respectivas acurácias, que as diferenças entre os erros são bem pequenos, mas também podemos observar que para a planta da classe 34 (*Pseudosasa japonica*) o acerto em ambos casos foi de 100%, isto é, para esse caso a base de conhecimento oferecida foi suficiente para realizar uma identificação precisa da espécie, enquanto com outras espécies isso ocorreu de forma contrário, ou seja, o algoritmo de classificação teve muita dificuldade para classifica-las, como nos indivíduos 26 (Euonymus japonicus) e 10 (Tilia tomentosa) que foram confundidos muitas vezes com as espécies 27 (Ilex perado ssp. Azorica) e 30 (Urtica dioica) respectivamente, entretanto se analisarmos as imagens da Figura 2 podemos notar que essas confusões seriam possíveis até mesmo aos nossos olhos e que para o caso de acerto total as diferenças entre ela e os demais exemplos são bem grandes, por fim podemos concluir que apesar de alguns erros o comportamento desse classificador seguiu com o esperado.

Conclusão e Discussão dos Resultados

Quando o assunto é escolher algum tipo de técnica computacional para algum determinado tipo de tarefa a resposta em geral é a mesma, depende de cada caso, e com classificadores isso não é diferente cada um tem suas particularidades, cada qual possui situações favoráveis e também situações desfavoráveis, alguns exemplos das situações dos dados foram colocados em forma de gráfico de dispersão no Apêndice B deste documento.

Mencionado anteriormente os testes foram aplicados em ambientes diferentes mas mesmo assim os resultados obtidos foram idênticos entre os métodos nesses ambientes, portanto para essa questão podemos notar que o software Orange® foi capaz de apresentar resultados iguais mesmo em situações diferentes, o único ponto divergente é quanto ao tempo de execução que por existirem processamentos mais rápidos e lentos entre os computadores utilizados houve uma influência direta no tempo.

Algo sobre essas situações particulares ficou muito claro na escolha dos classificadores kNN, MLP e Regressão Logística aplicados no dataset Leaf, pudemos perceber que para esses dados o kNN teve um desempenho inferior aos demais pela natureza da dispersão dos mesmos. Em contrapartida a MLP e a Regressão Logística não apresentaram a mesma dificuldade que o kNN, novamente para os dados apresentados.

Como dito no começo desse documento as escolhas das plantas foram feitas de tal forma que algumas espécies fossem similares entre si para tentar gerar erros propositais nos desempenhos dos classificadores, entretanto pudemos observar que apenas o kNN foi engano por nossas técnicas, enquanto os demais não se limitaram apenas a isso, porém devido as naturezas de cada um também era esperado essa diferença no desempenho individual, algo importante é que nenhum dos atributos avaliados na pesquisa original diziam respeito às cores de cada folha, se caso isso ocorresse provavelmente o desempenho de todos seria melhor, mas especialmente do kNN, já que a cor é um fator muito relevante nesse tipo de classificação.

Por fim, com esse projeto os integrantes puderam expandir o próprio horizonte de conhecimento dentro da gigante área chamada ciência da computação, especialmente nos campos de Inteligência Artificial e Processamento de imagens, apesar que essa segunda tenha sido menos relevante durante a elaboração do trabalho em si, pois o processamento em si já havia sido feito através da pesquisa original, foi possível agregar informações novas sobre a área. Também não podemos deixar de mencionar que com as pesquisas realizadas durante o andamento do projeto foi possível ver um pouco de como as maneiras de classificação de espécies ocorre e também como está evoluindo.

Referências Bibliográficas

http://www.sinfic.pt/SinficWeb/displayconteudo.do2?numero=25032
Acessado em 17/11/2018

https://www.tecmundo.com.br/o-que-e/3121-o-que-e-biometria-.htm
Acessado em 17/11/2018

http://www.linhadecodigo.com.br/artigo/1162/biometria-processamento-de-imagens-capturadas-em-leitores-de-impressao-digital.aspx
Acessado em 17/11/2018

https://iascblog.wordpress.com/2015/10/23/subareas-da-inteligencia-artificial-do-ponto-de-vista-computacional/
Acessado em 20/11/2018

https://medium.com/@eliezerfb/intelig%C3%AAncia-artificial-499fc2c4aa79
Acessado em 20/11/2018

https://matheusfacure.github.io/2017/01/15/pre-req-ml/
Acessado em 20/11/2018

https://pt.wikipedia.org/wiki/Regress%C3%A3o_linear
Acessado em 20/11/2018

https://www.youtube.com/watch?v=Fs8LhzhEMwI
Acessado em 20/11/2018

https://www.youtube.com/watch?v=CVL5vj1N1U8
Acessado em 20/11/2018

https://www.analyticsinsight.net/introduction-to-logistic-regression/
Acessado 20/11/2018

http://conteudo.icmc.usp.br/pessoas/andre/research/neural/
Acessado em 23/11/2018

https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781786468574/4/ch04lvl1sec28/multi-layer-perceptron
Acessado em 24/11/2018

http://www.cbpf.br/cat/pdsi/gauss.html
Acessado em 20/11/2018

http://www.cerebromente.org.br/n05/tecnologia/rna.htm
Acessado em 20/11/2018

https://sites.google.com/site/profleandrocfernandes/disciplinas/sii/SII-kNN.pdf?attredirects=0
Acessado em 18/11/2018

https://www.maxwell.vrac.puc-rio.br/7587/7587_6.PDF
Acessado em 18/11/2018

Apêndice A: Biometria na prática

No site linhadecódigo 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.

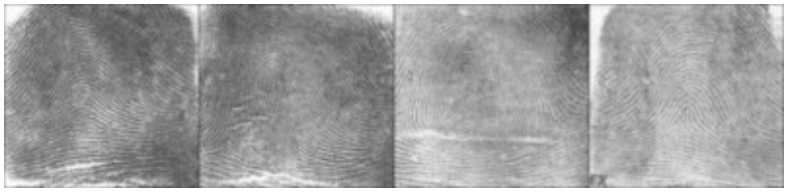


Figura 18 - Exemplo de diferentes impressões digitais

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.

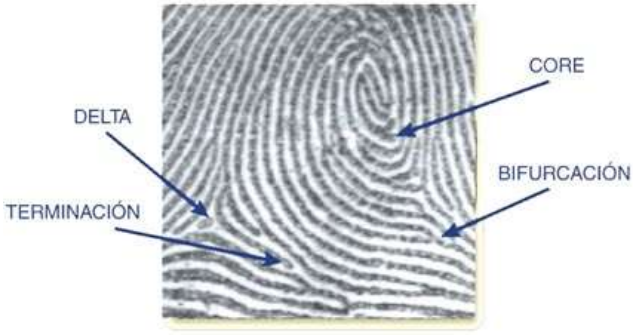


Figura 19 - Pontos que identificam uma digital

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.

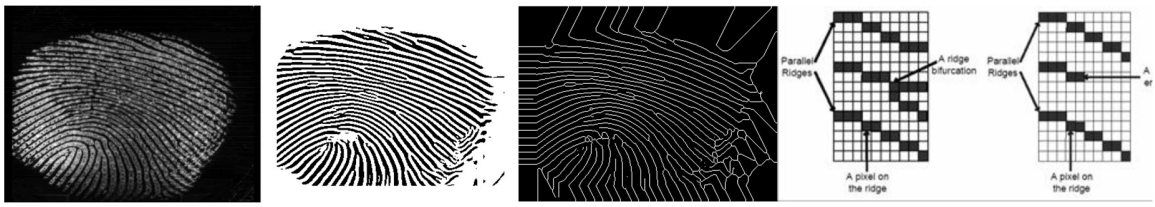


Figura 20 - Processamento de uma digital:

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;

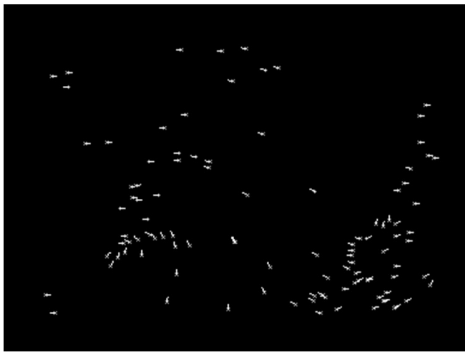


Figura 21 - Marcação dos pontos que serão comparados com o banco de dados

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

Nesta seção estão dois exemplos de gráficos de dispersão dos dados desse projeto, como podemos ver em alguns casos é relativamente simples separar os dados enquanto em outros pontos é uma tarefa mais complexa para os algoritmos, o que pode justificar alguns dos erros cometidos.

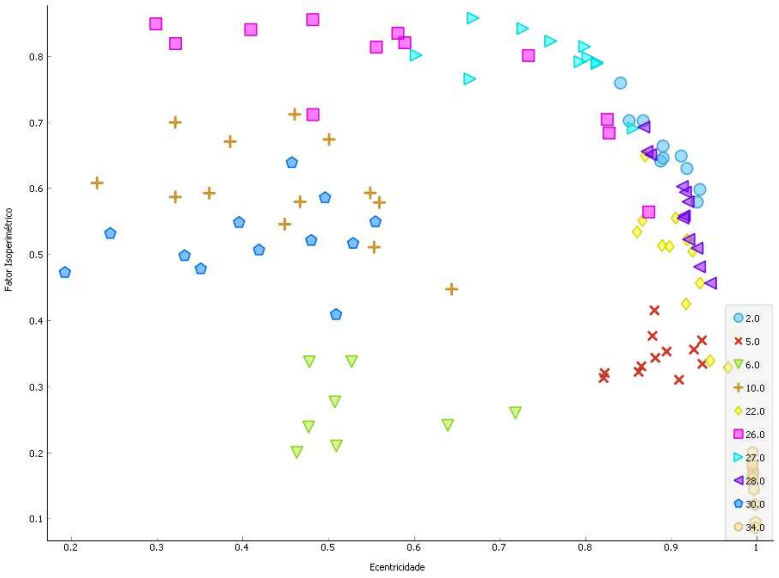


Figura 22 - Gráfico de dispersão Excentricidade x Fator Isoperimétrico

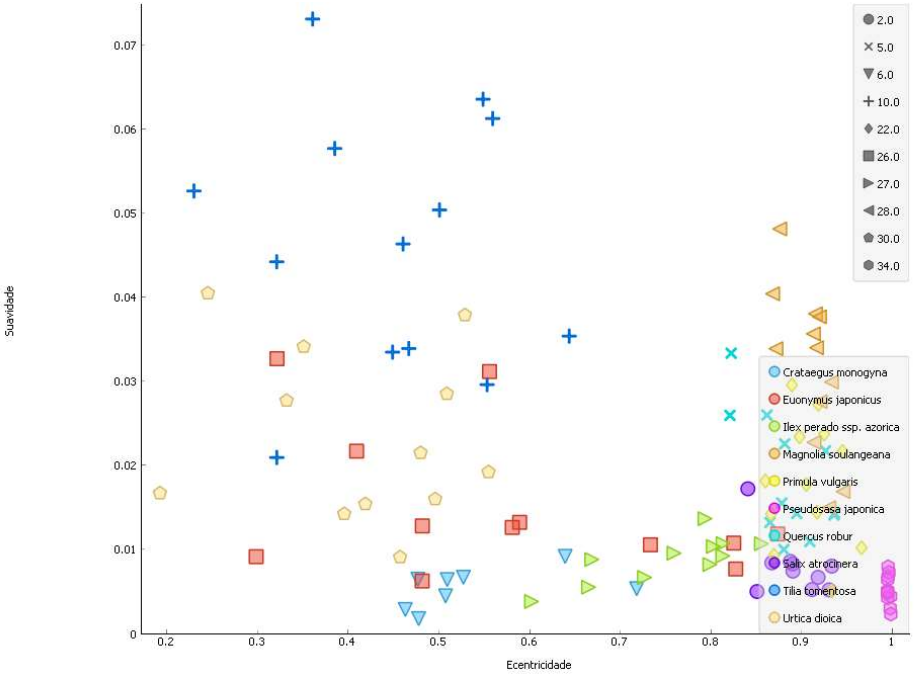


Figura 23 - Gráfico de dispersão Excentricidade x Suavidade