



**UNIVERSIDADE DE SÃO PAULO**  
**Instituto de Ciências Matemáticas e de Computação**

**Departamento de Sistemas de Computação**

---

Ferramenta de Apoio à Análise e  
Mineração de Dados para Agricultura

*Ruan Felipe de Oliveira Neves*

---

São Carlos - SP

# Ferramenta de Apoio à Análise e Mineração de Dados para Agricultura

*Ruan Felipe de Oliveira Neves*

*Orientador: Elaine Parros Machado de Sousa*

Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina SSC0593 Projeto de Graduação I do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP para obtenção do título de Bacharel em Sistemas de Informação.

Área de Concentração: Mineração de dados

**USP – São Carlos**  
**31/05/2016**

*“Do or do not... There is no try.” – Master Yoda.*

# Dedicatória

Dedicaria este trabalho ao meu afilhado Noah, que nos seus três anos de vida, não tem ideia do quanto me ajudou e ajuda em tudo, mas na devida idade ele entenderá o porquê não o fiz.

Dedico então este trabalho ao grande mestre, e amigo que foi o Professor Milton Sonoda, tirado de nós tão inesperadamente e de maneira tão cruel. Fique em paz amigo, e pode deixar que aqui a gente “tá trabalhando”.

# Agradecimentos

Agradeço, primeiramente, a minha orientadora Elaine, que aceitou essa terrível missão de me orientar neste trabalho (ps.: ainda tenho medo do gato morto); ao professor Marcassa, que dispensa quaisquer comentários, grande pesquisador, exemplo, orientador e guia, sem dúvida, este trabalho jamais existiria sem a participação dele em todo o processo que o antecedeu; aos professores Tereza Mendes e Attilio Cucchieri, pessoas as quais nunca saberei como demonstrar a gratidão que tenho por eles; ao meu grande amigo de pesquisa e exemplo Professor Caio, que me ensinou muito e me fez crescer muito também sob sua orientação; a minha família, a todos da minha família, mas principalmente a minha Mãe Monalisa, por tudo, simplesmente tudo que me ofereceu e ainda oferece, em todos os quesitos, áreas, realmente, obrigado Mãe; A minha namorada, Tami, que passou apuros comigo em dias sem dormir, onde meu humor era insuportável até pra mim “valeu gorda”; Aos meus amigos, “ahh meus amigos”, Jéssica, Hipo, Mari, Marcela, Thiago, Rosa, Sam, Rafa, May, Mariane, Bodelha, Brasil, (desculpe se algum nome ficou fora) que me aguentaram tantos dias, tardes, noites, de mal humor, insuportável, valeu galera! o/

# Resumo

O Brasil é o maior produtor de citros da atualidade, seguido pelos Estados Unidos. Pesquisas na área de agricultura de precisão vêm sendo desenvolvidas desde o início dos anos 90, no intuito de mitigar perdas e aumentar lucros na lavoura. Doenças como *greening* e cancro cítrico, por serem altamente destrutivas, e que podem ser facilmente confundidas com deficiência de zinco e verrugose, respectivamente, vêm devastando cada vez mais plantações, acarretando a agricultores grandes ônus em seus pomares. Existem várias pesquisas ao redor do mundo, principalmente no Brasil e Estados Unidos, com o intuito de melhorar a forma de diagnóstico destas doenças. Uma parceria, então formada por pesquisadores do Instituto de Física de São Carlos, do Instituto de Ciências Matemáticas e Computação e da Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo, e por colaboradores da *University of Florida* vem trabalhando no desenvolvimento de uma técnica baseada em espectroscopia de imagem de fluorescência, combinada a métodos de mineração de dados para otimizar o processo de realização do diagnóstico de doenças que afetam os citros. Tomando como base o trabalho destes pesquisadores, no qual o método proposto e publicado obteve grande eficácia, e no intuito de auxiliar pesquisadores da área, este trabalho de conclusão de curso propõe o desenvolvimento de uma ferramenta que seja capaz de integrar todo o processo computacional do experimento (extração de características das imagens, testes de algoritmos de classificação e análise dos resultados) numa única ferramenta com interface gráfica, de fácil uso, para que usuários sem muito conhecimento na área de computação tenham maior facilidade de realizar seus experimentos. Esta ferramenta será feita de forma completamente modular, orientada a objetos, *Open Source* e utilizando a linguagem Python, possibilitando sua expansão de acordo com as necessidades do pesquisador. A ferramenta traz em seu primeiro protótipo, algoritmos para extração da homogeneidade, energia e contraste, e os algoritmos de classificação SVM e KNN, dando possibilidade do usuário alterar os parâmetros de cada um dos métodos de classificação. Traz também parâmetros padrões definidos empiricamente.

# Sumário

<b>CAPÍTULO 1: INTRODUÇÃO .....</b>	<b>1</b>
1.1. CONTEXTUALIZAÇÃO E MOTIVAÇÃO .....	1
1.2. OBJETIVOS .....	2
1.3. ORGANIZAÇÃO DO TRABALHO .....	3
<b>CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA .....</b>	<b>4</b>
2.1. CONSIDERAÇÕES INICIAIS .....	4
2.2. DIAGNÓSTICO DE DOENÇAS EM CITROS .....	4
2.3. EXTRATORES DE CARACTERÍSTICAS .....	5
2.4. ALGORITMOS DE CLASSIFICAÇÃO .....	6
2.4.1. SVM .....	6
2.4.2. KNN .....	7
<b>CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO .....</b>	<b>8</b>
3.1. CONSIDERAÇÕES INICIAIS .....	8
3.2. PROJETO .....	8
3.3. DESCRIÇÃO DAS ATIVIDADES REALIZADAS .....	9
3.3.1. <i>Processo de engenharia de software</i> .....	9
3.3.2. <i>Escolha da linguagem</i> .....	10
3.3.3. <i>Desenvolvimento dos algoritmos de extração de características</i> .....	10
3.3.4. <i>Desenvolvimento e testes dos algoritmos classificadores</i> .....	11

3.3.5. Desenvolvimento da interface gráfica .....	11
3.3.6. Integração da interface gráfica com os algoritmos.....	11
3.4. DIFICULDADES E LIMITAÇÕES .....	12
3.5. CONSIDERAÇÕES FINAIS .....	12
<b>CAPÍTULO 4: RESULTADOS.....</b>	<b>13</b>
4.1. EXTRAÇÃO DE CARACTERÍSTICAS .....	13
4.2. CLASSIFICADORES .....	14
4.3. RESULTADOS .....	15
4.4. EXEMPLO DE EXECUÇÃO.....	16
<b>CAPÍTULO 5: CONCLUSÃO .....</b>	<b>18</b>
5.1. CONTRIBUIÇÕES .....	18
5.3. RELACIONAMENTO ENTRE O CURSO E O PROJETO .....	18
5.4. CONSIDERAÇÕES SOBRE O CURSO DE GRADUAÇÃO.....	18
5.5. TRABALHOS FUTUROS .....	19
<b>REFERÊNCIAS.....</b>	<b>20</b>



# CAPÍTULO 1: INTRODUÇÃO

## 1.1. Contextualização e Motivação

No início dos anos 90, houve o surgimento do conceito de agricultura de precisão, [10] com o intuito de aumentar a produtividade e o lucro na lavoura para reduzir ao máximo as perdas causadas por pragas. No Brasil, o maior produtor de citros [10] no mundo, algumas doenças como o *greening citrus* (*Huanglongbing*, ou HLB) [2] e o cancro cítrico [3] têm se mostrado muito agressivas, o que torna cada vez mais interessante investimento e pesquisa na área.

O HLB, considerado uma das doenças mais letais e destrutivas, é causado pela bactéria “*Candidatus Liberibacter spp.*” [2]. Seus sintomas são folhas pequenas, com manchas amareladas e frutos pequenos, deformados e amargos [3]. O cancro, também altamente destrutivo, é causado pela bactéria “*Xanthomonas citri subsp. citri*” e afeta todas as espécies de citros de importância comercial, como espécies de laranja e limão. Seus sintomas são um pouco diferentes do HLB, apresentando pequenos pontos salientes de coloração marrom na parte inferior da folha, os quais vão aumentando e ganhando coloração mais escura com a evolução da doença [3]. Estas lesões podem atingir os frutos e levar à necrose do tecido afetado, conseqüentemente, desfolhamento das plantas, depreciação dos frutos por causa das lesões e redução na produção, por queda prematura.

Atualmente, o diagnóstico dessas doenças é realizado principalmente a partir de ensaios bioquímicos [1], que possuem alto custo e demandam muito tempo para obtenção dos resultados, em relação à velocidade de propagação da doença na lavoura. O HLB e o cancro cítrico também podem ser facilmente confundidos com as doenças deficiência de zinco e verrugose [11, 12], por possuírem sintomas visuais parecidos. Outros métodos de diagnósticos [10, 11, 12] estão sendo estudados no intuito de cortar custos e acelerar o processo de identificação e erradicação das doenças da lavoura. Alguns desses métodos vêm se mostrando muito eficazes e com custo bastante acessível.

A técnica de interesse neste trabalho, baseada no uso da espectroscopia de imagens de fluorescência (FIS, do inglês *Fluorescence Image Spectroscopy*) [10, 11, 12] para aquisição de imagens das folhas de plantas contaminadas foi desenvolvida no Laboratório de Interações Atômicas (LIA) do Instituto de Física de São Carlos (IFSC), em parceria com pesquisadores de outras universidades e tem se mostrado muito eficiente. Estas imagens são submetidas a algoritmos de classificação [5, 7, 8, 9], que as diferenciam com base em suas características de textura [4]. Em geral, a taxa de acerto na diferenciação de doenças passa dos 95%, considerada excelente nesta área.

Atualmente, pesquisadores do IFSC e colaboradores que utilizam FIS para detecção de doenças realizam todo o processo de extração de características e aplicação de algoritmos de classificação por meio de código desenvolvido, por exemplo, em MATLAB<sup>®</sup>, e são executados por linhas de comando, o que torna seu uso complicado para o agricultor ou pesquisador que não possua algum conhecimento na área de computação. Para tentar reduzir esta limitação, neste trabalho é proposta uma ferramenta que dispõe de uma interface gráfica organizada e de fácil utilização. Esta permite que o usuário, independentemente de conhecimento específico, consiga realizar todo o processo de análise das imagens para, por exemplo, dar apoio a sua pesquisa.

## **1.2. Objetivos**

Este trabalho tem como objetivo principal desenvolver uma ferramenta que torne fácil e simples a tarefa de reconhecer doenças em plantas, por meio de um sistema prático, barato, rápido e eficiente, em comparação às técnicas atualmente utilizadas. O objetivo é que o software para diagnóstico de doenças possa ser usado facilmente por pesquisadores que desejarem realizar estudos relacionados a imagens de plantas. O aplicativo deverá disponibilizar recursos, tais como: extratores de características, ferramentas de visualização de imagens e de resultados, algoritmos de mineração de dados (principalmente algoritmos de classificação), dentre outros que possam ser úteis para a área de aplicação.

Para implementação do primeiro protótipo da ferramenta, que será capaz apenas de lidar com conjuntos já rotulados, e utilizado somente como ferramenta de apoio à pesquisa, foram selecionados: extratores de características de textura, em particular contraste, energia e homogeneidade. Os algoritmos de classificação *Support Vector Machine* (SVM) e *k-Nearest Neighbors* [5, 7, 8] e a medida de avaliação de qualidade *k-fold cross validation* foram adotados. Tais escolhas foram baseadas em trabalhos anteriores que utilizaram essas técnicas para a classificação de doenças, como cancro cítrico e verrugose [10, 11, 12]. A ferramenta também traz como uma de suas principais características a alta modularidade, ou seja, ela pode ser expandida facilmente para incorporar novas técnicas e suporte à identificação de outras doenças em culturas distintas.

No que diz respeito aos resultados, a ferramenta traz um primeiro resumo sobre a classificação, onde são apresentados a matriz de confusão [5, 7, 8] e a precisão do teste. Entretanto, deixa à disposição uma área na qual podem ser vistos resultados mais detalhados. Dentre os resultados mais detalhados estão novamente a precisão, matriz de confusão, seguidos da especificidade, sensibilidade e uma tabela mostrando a quantidade de falsos positivos e negativos e verdadeiros negativos e positivos.

### **1.3. Organização do Trabalho**

No Capítulo 2 é apresentado o conteúdo teórico utilizado no projeto, bem como trabalhos de literatura relacionados ao mesmo. A seguir, no Capítulo 3, são descritas as atividades realizadas ao longo do desenvolvimento do projeto. No Capítulo 4, são apresentados os resultados obtidos, dentre eles, algumas telas da interface da ferramenta, gráficos gerados e outros dados relacionados especificamente à aplicação. Por último, no Capítulo 5, encontram-se as conclusões após o desenvolvimento do projeto, discussões dos resultados obtidos e dificuldades encontradas durante o desenvolvimento do trabalho.

# CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

## 2.1. Considerações Iniciais

A busca por métodos mais precisos e com o menor número de falsos negativos é cada vez mais um desafio, dado o grande prejuízo que a detecção tardia das pragas podem acarretar. Neste capítulo, serão abordados conceitos envolvidos no método desenvolvido por pesquisadores do IFSC, em conjunto com pesquisadores do Instituto de Ciências Matemáticas e Computação da Universidade de São Paulo (ICMC), da Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo (ESALQ) e *University of Florida*, no qual este trabalho se baseia, e também conceitos que foram explorados no decorrer do projeto.

O Capítulo está organizado da seguinte forma: as técnicas mais usadas de diagnósticos na atualidade; os métodos de extração de características; os algoritmos de classificação.

## 2.2. Diagnóstico de doenças em citros

Atualmente, os métodos mais comuns para diagnóstico e confirmação de pragas no cultivo de cítricos, tais como do Cancro Cítrico, Verrugose, *Greening* e Deficiência de Zinco, baseiam-se em duas técnicas bioquímicas: reação em cadeia da polimerase (PCR, do inglês *Polymerase Chain Reaction*) [1] e ELISA (do inglês, Enzyme-Linked Immunosorbent Assay) [1], consiste em realização do isolamento e identificação do patógeno em meio de cultura adequado.

A PCR é um método de biologia molecular para detecção indireta de genes. Consiste em replicar várias vezes a mesma parte de interesse do DNA, por repetição de ciclos de replicação de um gene determinado, com *primers* iniciadores da replicação (desenhados especificamente para este gene), em uma reação realizada *in vitro*. Estes ciclos consistem em uma etapa de desnaturação do DNA a alta temperatura, para que ocorra separação da dupla fita, anelamento dos *primers* com DNA, que servem de molde

para início da replicação direta e reversa pela DNA polimerase, e uma etapa de extensão, em que ocorre de fato a replicação. Posteriormente, o produto de PCR pode ser analisado por eletroforese em gel de agarose, no qual o tamanho do gene é confirmado [1].

ELISA é um teste que permite detecção indireta de anticorpos específicos, portanto, de várias doenças. É baseado na especificidade da ligação antígeno-anticorpo. Para tanto, os antígenos de interesse são imobilizados em uma placa ou plataforma e neles são aderidos anticorpos primários, que se ligam com a proteína de interesse. Posteriormente, são tratados com anticorpos secundários acoplados a uma enzima, específica para ligar aos anticorpos primários. Esta enzima produz uma substância corada ao se ligar ao substrato. Portanto, quando adicionado substrato à plataforma, é possível medir a intensidade da coloração do meio, para determinar de forma indireta, se há ou não presença do antígeno da doença procurada [1].

Outro método que vem sendo muito estudado é o uso de algoritmos de mineração de dados combinado a técnica de espectroscopia de imagem de fluorescência. A folha da planta infectada é submetida a uma excitação luminosa por meio de um diodo emissor de luz, e então são geradas imagens a partir da fluorescência dela, em comprimentos de ondas específicos. Essas imagens são submetidas a filtros computacionais para correção de ruídos experimentais e submetidas a classificadores que são capazes de distinguir, com base em modelos previamente criados de amostras já rotuladas, qual a doença daquela amostra. [11, 12]

### **2.3. Extratores de características**

Extratores de características são algoritmos baseados em operações matemáticas matriciais, capazes de extrair métricas de imagens, denominadas características. Existem diversos tipos de características [4, 11, 12] e cada tipo traz consigo grandes grupos de características, cada qual com suas mais diversas aplicações. Dentre as possíveis aplicações, estão processamento gráfico [4], reconhecimento de padrões [4, 5, 7, 8, 9], detecção de falhas e exames médicos [3].

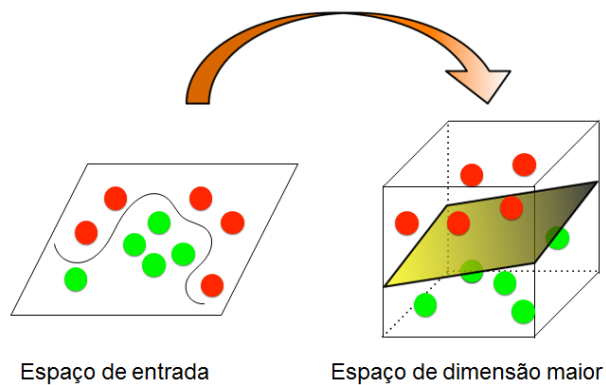
## 2.4. Algoritmos de classificação

Algoritmos de classificação definem uma classe de algoritmos pertencentes à mineração de dados [5, 7], capazes de, a partir de um conjunto de amostras rotuladas, rotular outras amostras, tomando como base suas características. O processo é realizado por meio da combinação de operações matemáticas e heurísticas muito bem definidas e têm sua qualidade medida a partir de métricas como precisão, taxa de acerto, taxa de erro, número de falsos negativos ou positivos, dentre outros. Possuem aplicações nas mais diversas áreas do conhecimento, dentro e fora do mundo da computação, como por exemplo, na biologia, medicina e física.

### 2.4.1.SVM

SVM são algoritmos de classificação que dividem os dados em diferentes classes [5, 7, 8], baseados no conceito de planos de decisão, os quais são construídos a partir de uma função chamada “função de *kernel*”, que possui alguns parâmetros de entradas que devem ser dados. Estes planos por sua vez são planos n-dimensionais, chamados de hiperplanos, onde n é o número de dimensões dos dados.

Neste trabalho, foi utilizado o modelo clássico do algoritmo SVM, que se trata de um classificador binário que diferencia os dados em duas classes, de acordo com o lado que o dado se encontra em relação ao hiperplano. A classificação funciona da seguinte forma: uma vez que o modelo é criado, quando um novo ponto é adicionado, é atribuído a ele uma classe, de acordo com o lado que ele ocupar referente ao hiperplano (Figura 1).



**Figura 1 - Separação do SVM**

## 2.4.2. KNN

O KNN [5, 7, 8] se trata de um algoritmo de classificação baseado em funções de distâncias. É, sem dúvida, um dos mais simples, mais citados e usados na literatura, com um método de aprendizado chamado de “preguiçoso”. O mesmo aprende e cresce com cada ponto que é adicionado a ele, não gerando assim um modelo matemático fixo baseado num conjunto de treinamento. Ele classifica novos pontos a partir dos  $k$  pontos mais próximos do mesmo, no qual o próximo refere-se ao que apresentar menor distância, de acordo com a função de distância adotada. Existem uma infinidade de funções de distância, as mais comuns são euclidiana [5, 7] e manhattan [5, 7]. Em alguns casos, são criadas novas funções, que se adequam de forma ideal ao domínio de aplicação.

# CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

## 3.1. Considerações Iniciais

Neste capítulo será apresentado o projeto como um todo, incluindo técnicas e ferramentas utilizadas, a forma que foram usadas e aplicadas, os objetivos esperados e também os objetivos que foram atingidos com cada parte do processo. Serão também apresentados, em detalhes, os passos do desenvolvimento dos códigos responsáveis pela extração das características, pela mineração dos dados e os métodos utilizados para se realizar a análise dos resultados. E, por último, expor a maneira pela qual a interface foi construída, qual a ideia por trás da mesma e os resultados alcançados.

## 3.2. Projeto

Com o intuito de desenvolver uma ferramenta eficiente e de fácil utilização, foi necessário fazer uso de recursos de programação que possibilitassem a elaboração de toda a parte “científica” do código, somada a uma interface agradável. Para tanto, foi escolhida Python como linguagem a ser usada para o desenvolvimento do trabalho, pois possibilitava uma implementação clara e limpa dos métodos, tanto para a extração de características das imagens, quanto para o processo de mineração de dados. É também capaz de prover um ambiente agradável para o desenvolvimento de uma interface gráfica de boa usabilidade.

Os métodos de extração de características das imagens foram primeiramente implementados e testados. Estes são responsáveis por extrair informações das imagens, por meios de métodos matemáticos, para que essas possam ser classificadas por um algoritmo classificador.

Uma vez que tínhamos os métodos de extração implementados e testados, passamos para a implementação dos algoritmos de classificação, começando pelo KNN. Foram realizados testes iniciais em que verificamos a eficiência do método, tanto no quesito de tempo, quanto de precisão. Na sequência, foi implementado o SVM e o mesmo foi



submetido a vários testes, levando em conta a variação tanto do *Kernel*, como dos parâmetros C e Gamma.

Por último, foi construída e implementada a interface gráfica da ferramenta, na qual foram organizados todos os códigos já existentes, de maneira que os mesmos ficassem fácil de usar, acessar e exibissem resultados, de maneira que qualquer usuário com pouco conhecimento na área fosse capaz de entendê-los e utilizá-los.

### **3.3. Descrição das Atividades Realizadas**

Nesta seção, serão descritos, passo a passo, todos os procedimentos realizados durante o desenvolvimento do projeto. Será abordada cada etapa, no intuito de explicitar métodos, ferramentas e recursos utilizados, referentes a conhecimento, linguagem e técnicas de programação.

#### **3.3.1. Processo de engenharia de software**

A partir da consulta de especialistas da área, como por exemplo, o Professor Caio Wetterich, criamos a seguinte lista de requisitos: 1) Seleção da pasta contendo as imagens (dados); 2) Visualização de como os dados se distribuem no espaço de parâmetros; 3) Seleção de quais características devem ser extraídas, por meio de uma lista de múltipla seleção, para a mineração; 4) Opção de salvar as características atribuídas, atribuindo-lhes rótulos (*labels*); 5) Fácil seleção de diversos algoritmos de mineração; 6) Fácil seleção de diversos algoritmos de verificação de qualidade; 7) Ajuste de parâmetros dos algoritmos de mineração e de verificação de qualidade; 8) Visualização dos resultados em resumo e completo para impressão; 9) Fácil expansão para novos algoritmos e funcionalidades; 10) *Graphical User Interface* (GUI) amigável; 11) Planejamento da *User Experience* (UE) baseado na facilidade.

Com isso modelamos a ferramenta (Figura 2)

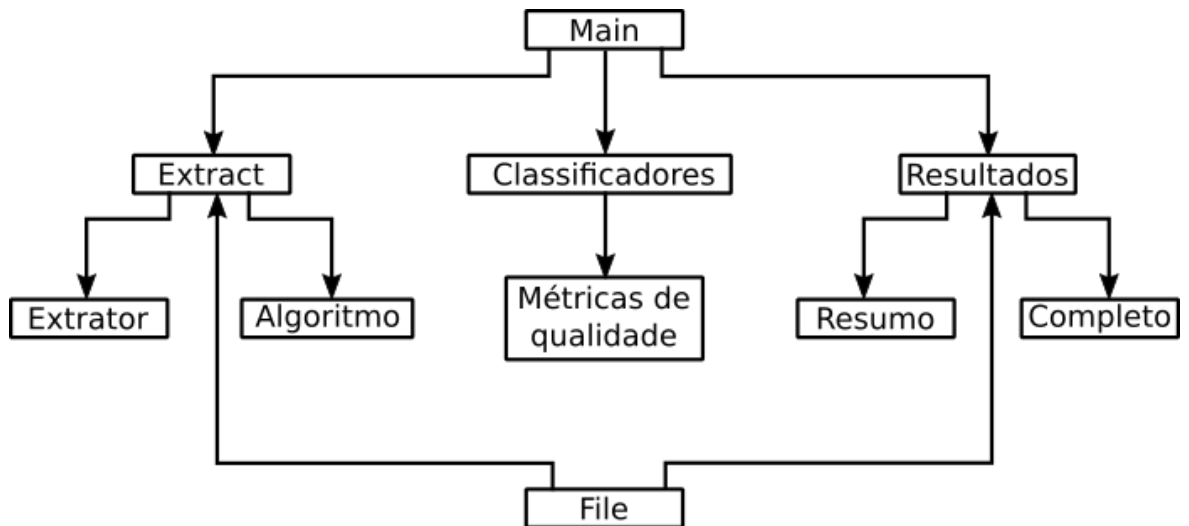


Figura 2: Diagrama simplificado de classes do projeto

### 3.3.2. Escolha da linguagem

Após a realização de testes com outras linguagens, como Java e C++, a linguagem Python foi escolhida para o desenvolvimento do trabalho, por se mostrar muito mais conveniente, no quesito de unir todos os requisitos necessários para a ferramenta desejada. Esta dispõe de conjunto de bibliotecas muito bem documentadas e trata-se de uma linguagem que segue a filosofia da legibilidade, fácil implementação e de alto nível (códigos complicados, densos e de difícil interpretação são dispensados, reduzindo o tempo de implementação) [13].

### 3.3.3. Desenvolvimento dos algoritmos de extração de características

Um dos requisitos para o trabalho era ter um código capaz de extrair características, tanto de textura, como estatísticas das imagens, de maneira rápida e eficiente. Para tanto, foram utilizados recursos da biblioteca *skimage* [8], para se extrair, num primeiro protótipo a homogeneidade, a energia e o contraste. Esta biblioteca fornece recursos para extração de ambos os tipos de características e uma documentação clara e bem consolidada, que possibilita tratar os retornos das funções e utilizá-los da forma desejada.

### **3.3.4. Desenvolvimento e testes dos algoritmos classificadores**

Utilizamos os métodos KNN e SVM, por se tratarem de métodos de fácil interpretação, além de possuírem eficácia já comprovada, para conjuntos de dados do mesmo domínio de aplicação[6, 7]. Estes métodos apresentaram uma acurácia muito boa. Ambos foram implementados usando a biblioteca *sklearn* [8], que permite total controle de todos os parâmetros que serão utilizados e também função de paralelismo, quando se trata de um conjunto altamente massivo de dados. O usuário pode alterar os parâmetros principais dos algoritmos ou pode optar por utilizar os valores que a ferramenta já traz por padrão. Esses valores foram obtidos a partir de testes realizados durante o desenvolvimento do projeto. No caso do KNN, o algoritmo foi testado variando-se a função de distância entre a euclidiana e a Manhattan e também o K. No SVM, foi testado o *kernel* “rbf” com diversas combinações de C e Gamma.

### **3.3.5. Desenvolvimento da interface gráfica**

A interface gráfica foi desenhada e implementada, independente do que se havia feito até agora. Foi projetada buscando facilitar a usabilidade de um usuário com pouco ou quase nenhum conhecimento na área de computação, com intuito de sempre organizar as janelas e botões de maneira simples e intuitivas. Buscamos ajuda com especialistas da área, pesquisadores do IFSC e da ESALQ, para fazer o levantamento de requisitos e desenhar uma interface que, de fato, fosse agradável, prática e realmente facilitasse o trabalho.

A interface foi escrita em *Tkinter*, uma biblioteca do Python feita exclusivamente para criação de interfaces gráficas, com diversos recursos que possibilitam a criação de telas simples e com todos os recursos necessários para que a ferramenta cumpra com seus objetivos.

### **3.3.6. Integração da interface gráfica com os algoritmos**

Por último, foi necessário realizar a integração entre todos os algoritmos (de extração, classificação e geração de gráficos) com a interface gráfica, já que foram feitas de forma independente. A integração demandou certo esforço e houve a necessidade de

adequar algumas implementações às funcionalidades da interface, além da exibição de resultados, em seus devidos locais.

### **3.4. Dificuldades e Limitações**

A maior dificuldade encontrada foi o desenvolvimento da interface gráfica, seguido da integração entre a interface e a implementação. Tal dificuldade se deu por conta do tempo utilizado para aprender a utilizar e dominar todas as funcionalidades da linguagem.

### **3.5. Considerações Finais**

O trabalho exigiu organização e sistematização rigorosa, uma vez que toda fase posterior era altamente dependente da anterior. Mesmo durante o desenvolvimento da interface gráfica, completamente independente do restante do trabalho, houve necessidade das outras etapas para planejar o conteúdo de cada janela e sua organização, e também, pensar na melhor maneira de se fazer as integrações.

No capítulo resultados, serão exibidas algumas telas da ferramenta, ilustrando o uso da mesma por especialistas e algumas das funcionalidades que a esta dispõe, junto com alguns resultados obtidos.

# CAPÍTULO 4: RESULTADOS

Para simplificar o uso da ferramenta, ela foi dividida em três grandes etapas: extração das características; geração e aplicação dos classificadores e visualização dos resultados. A seguir, descrevemos o uso de cada uma dessas etapas, ilustradas por imagens das respectivas abas.

## 4.1. Extração de características

A extração de características (Figura 3) consiste em 2 etapas: 1) seleção do diretório contendo as imagens, que devem ser todas da mesma doença, e 2) seleção de quais características serão utilizadas pelo algoritmo de classificação.

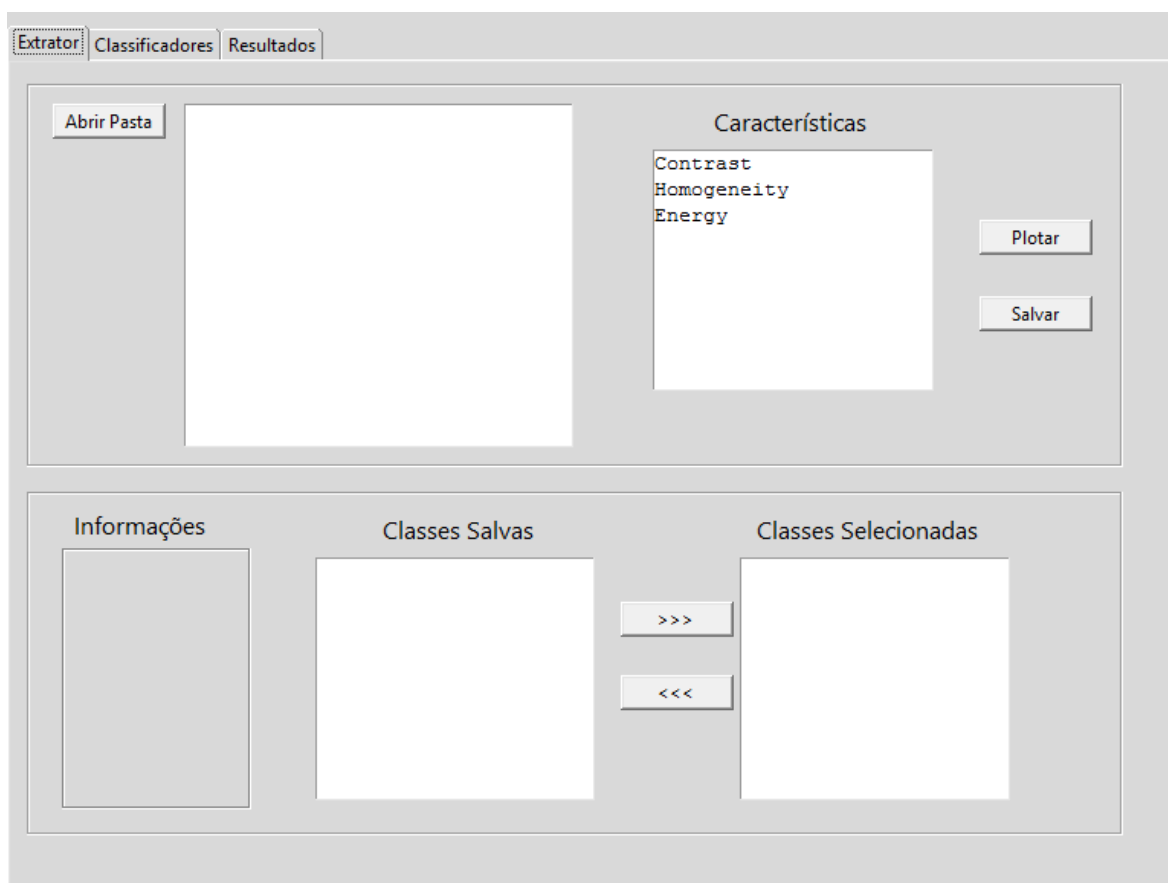
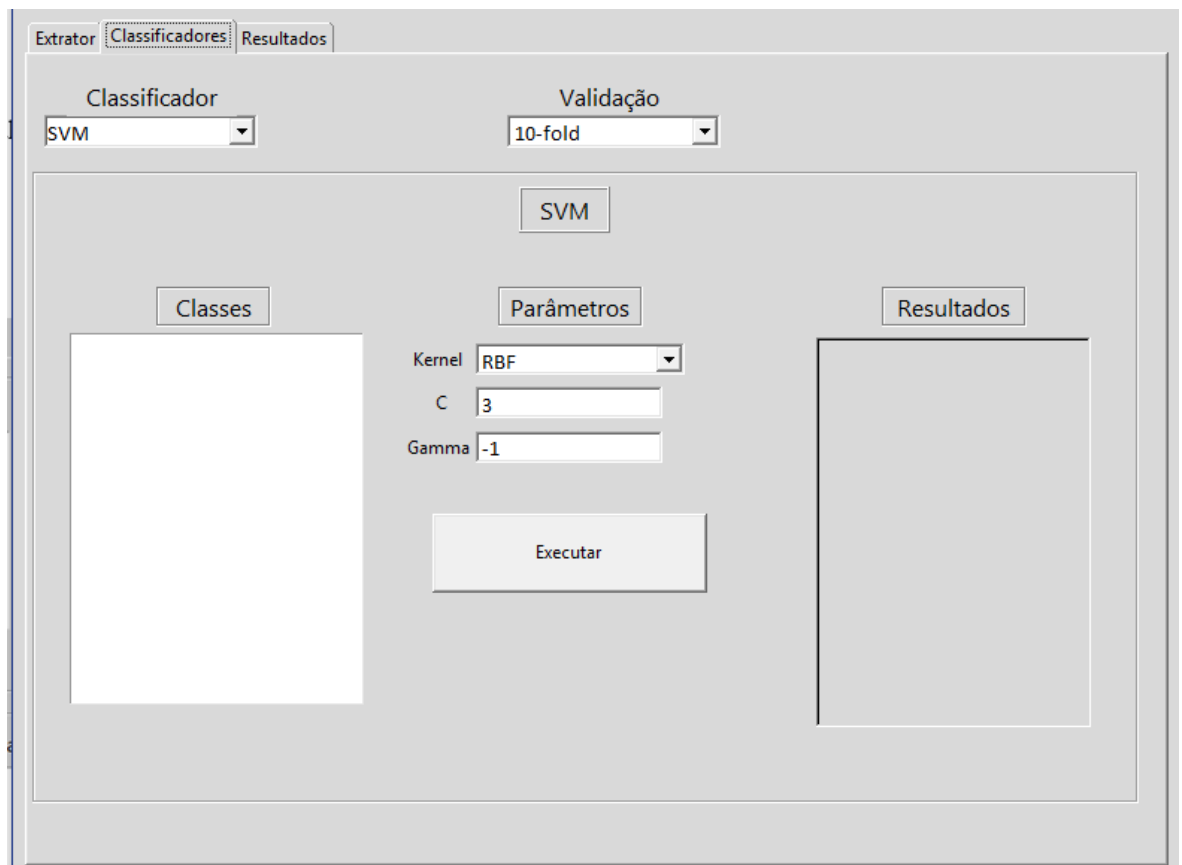


Figura 3: Interface para a extração de características das imagens

Para facilitar a seleção das características, é possível gerar um gráfico de até 3 dimensões, para visualizar como estas componentes estão distribuídas pelo espaço. A interface foi implementada de forma a facilitar a inclusão de novas características a serem extraídas. Após a escolha de quais características serão utilizadas, os dados extraídos são salvos em uma classe (*label*), cujo nome é dado pelo usuário. À direita, está um local onde são exibidas informações, como características extraídas e número de amostras, quando uma classe é selecionada com clique do mouse.

## 4.2. Classificadores

O uso dos classificadores (Figura 4) deve ser feito em 3 etapas: seleção do classificador desejado, método de validação e parâmetros desejados.

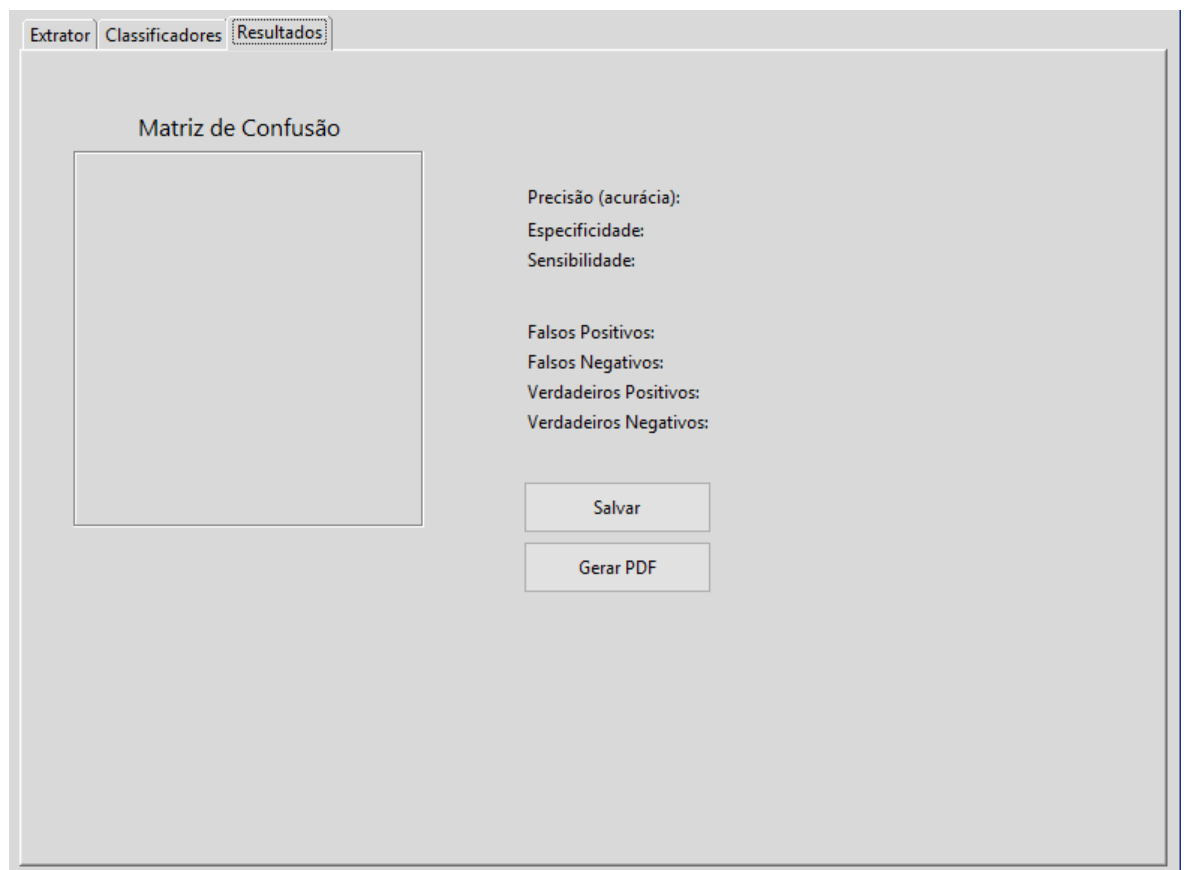


**Figura 4: Interface para uso dos classificadores**

Para o uso dos classificadores, o usuário deve selecionar o classificador desejado. Escolhido o classificador, deve-se então definir um método de validação (por padrão traz o *10-fold cross validation* [7, 8]) e, então definir os parâmetros do classificador ou usar os parâmetros iniciais, definidos empiricamente. Após, deve selecionar quais classes (geradas na etapa de extração de características) submeter ao classificador, e então, executar. O resumo dos resultados, contendo a precisão (ou acurácia) e a matriz de confusão gerada, será exibido à direita da tela. O relatório completo pode ser visualizado na aba “Resultados”.

### 4.3. Resultados

Na aba “Resultados” são exibidas mais informações sobre o processo de classificação, como pode ser visto na Figura 5.

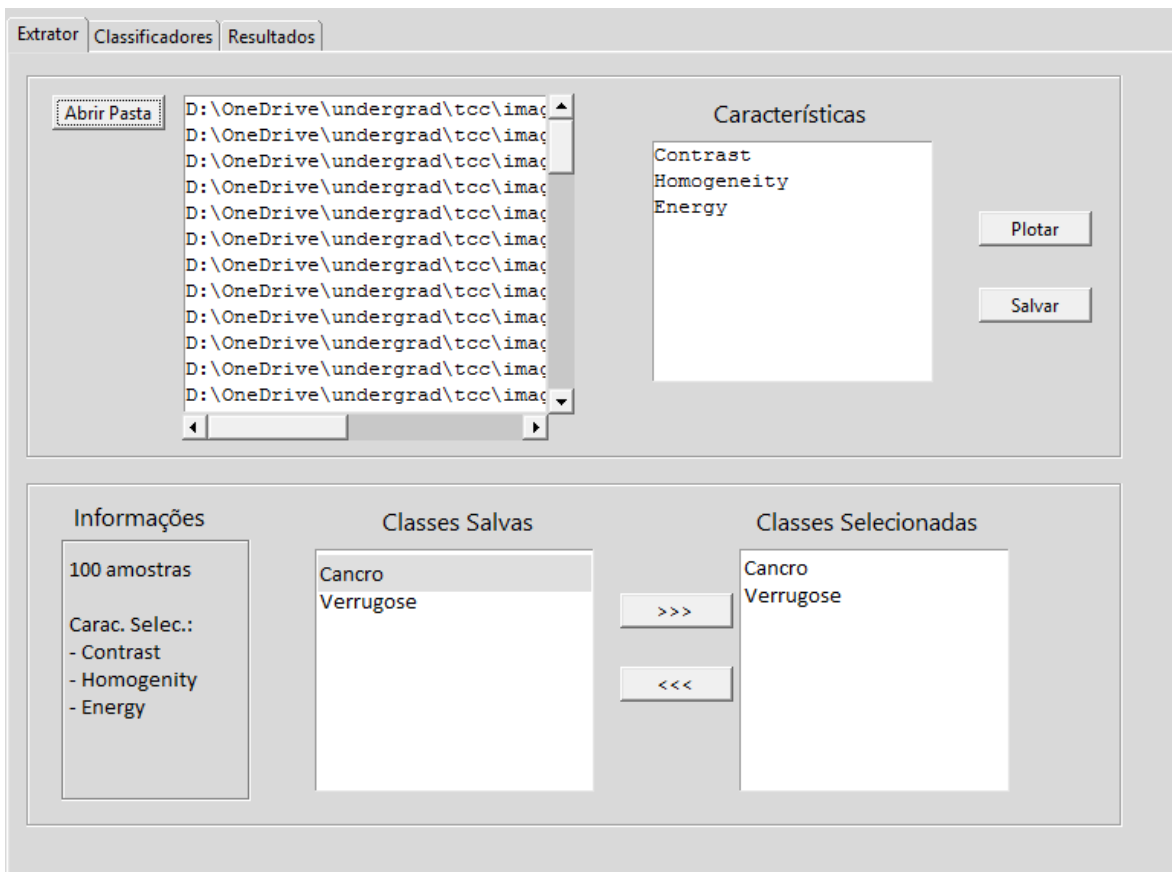


**Figura 5: Interface para visualização dos resultados**

Na tela, são exibidos a matriz de confusão novamente, além da acurácia. Ela também traz a sensibilidade e a especificidade [5, 7, 8] do teste, logo abaixo os dados usados para gerar a matriz de confusão, e dois botões, um que permite o usuário salvar o classificador gerado, e outro, que permite ao usuário gerar um relatório em PDF contendo aquelas informações.

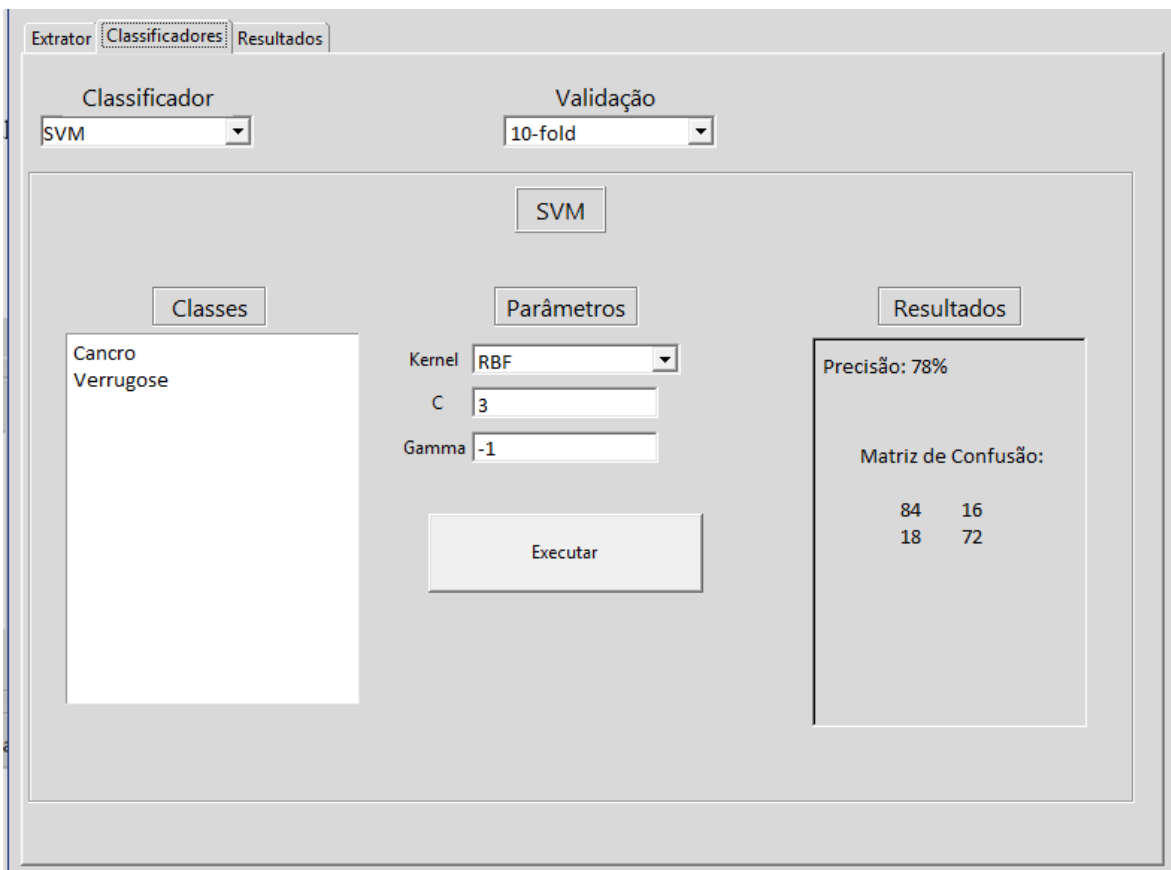
#### 4.4. Exemplo de execução

A seguir descrevemos um exemplo da execução do programa com extração das características (Figura 6), classificadores (Figura 7) e resultados (Figura 8).

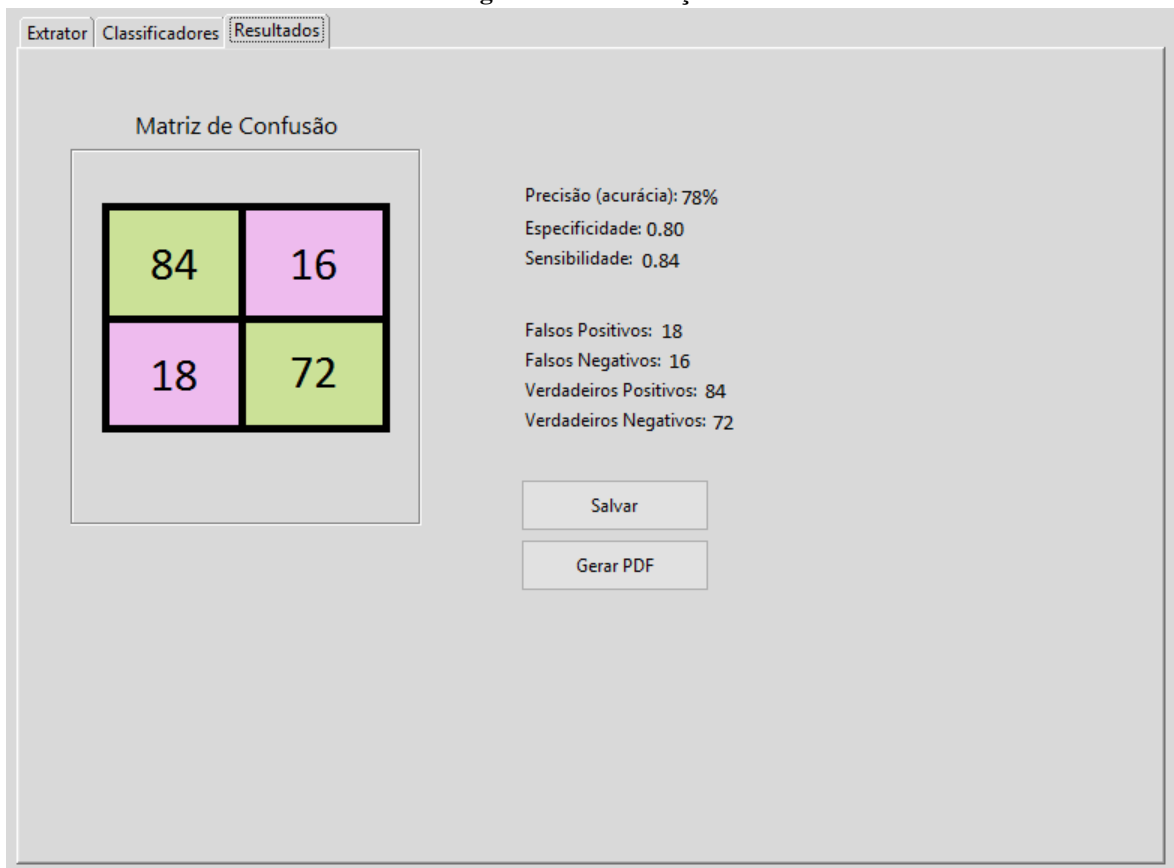


**Figura 6:** Tela de extração de características com pasta carregadas, características extraídas e salvas com respectiva classe.





**Figura 7: Classes submetidas ao classificador SVM, onde podem ser vistos os parâmetros e o algoritmo de validação.**



**Figura 8: Aba com resultados mais explícitos e detalhados.**

# CAPÍTULO 5: CONCLUSÃO

## 5.1. Contribuições

A interface desenvolvida, além de facilitar a utilização de algoritmos de mineração de dados pelos pesquisadores sem conhecimento profundo em computação, integra nela todo o processo, desde a extração de características até a visualização dos resultados. Por ter sido implementada de forma completamente modular, permite fácil expansão para novas metodologias e características. Tais fatores tornam possível utilizar esta ferramenta para aumentar a eficiência de investigações futuras em mineração de dados de imagens. Todo *software* desenvolvido é *Open Source*, distribuído gratuitamente e em linguagem livre.

## 5.3. Relacionamento entre o Curso e o Projeto

O curso forneceu grande base de conhecimento para realização do projeto. Disciplinas como “Programação Orientada a Objeto”, “Algoritmos e Estruturas de Dados II”, “Mineração de Grandes Bases de Dados”, “Inteligência Artificial”, “Engenharia de Software”, “Interação Usuário-computador”, “Programação para Alto Desempenho”, dentre outras, foram cruciais para elaboração deste projeto. Por meio delas pude adquirir o conhecimento e maturidade necessários para o desenvolvimento do mesmo, no que se diz respeito a conhecimento e técnicas sobre planejamento do desenvolvimento, técnicas de programação, algoritmos e eficiência de algoritmos, técnicas e métodos de pré-processamento e mineração de dados (base e motivação de todo esse trabalho).

## 5.4. Considerações sobre o Curso de Graduação

Acredito que o curso tem tudo para ser um curso de excelência. No geral, dispõe de excelentes professores, preparando o aluno com uma formação sólida, tanto para o mercado, quanto para a academia. Os pontos fortes são muitos, e os fracos bem poucos, mas dentre eles, o mais gritante, que acredito que vale ser ressaltado, é a discrepância entre mesmas disciplinas dadas por professores diferentes e a organização de alguns professores

com respeito a nota e assiduidade. Acredito que esses dois pontos deveriam receber mais atenção de imediato. Do mais, acredito que o curso tem tudo para despontar e se firmar com a grade atual.

## **5.5. Trabalhos Futuros**

Como trabalho futuro, a ser realizado na disciplina “Trabalho de conclusão de curso II”, a ferramenta será expandida, tentando torná-la uma ferramenta que, além de auxiliar o pesquisador da área, possa também colaborar com o agricultor, apresentando algumas ferramentas como um classificador previamente treinado, para imagens não classificadas. Também está no planejamento aumentar o número algoritmos extratores e classificadores, melhorias na interface gráfica, na parte de resultados e adicionar a realização de análises comparativas sobre os dados e classificadores.

# REFERÊNCIAS

[1] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21054/>

[2] BATOOL, A. et al. Citrus greening disease—a major cause of citrus decline in the world—a review. Hort. Sci.(Prague), v. 34, n. 4, p. 159-166, 2007.

[3] DAS, A. K. Citrus canker-A review. Journal of Applied Horticulture, v. 5, n. 1, p. 52-60, 2003.

[4] HARALICK, Robert M.; SHANMUGAM, Karthikeyan; DINSTEN, Its' Hak. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on, n. 6, p. 610-621, 1973.

[5] JIAWEI HAN, M. K.; PEI, Jian. Data Mining: Concepts and Techniques: Concepts and Techniques. 2011.

[6] LINS, Emery C.; BELASQUE JR, José; MARCASSA, Luis G. Detection of citrus canker in citrus plants using laser induced fluorescence spectroscopy. Precision agriculture, v. 10, n. 4, p. 319-330, 2009.

[7] MINER, Gary; NISBET, Robert; ELDER IV, John. Handbook of statistical analysis and data mining applications. Academic Press, 2009.

[8] MUCHERINO, Antonio; PAPAJORGJI, Petraq J.; PARDALOS, Panos M. Data mining in agriculture. Springer Science & Business Media, 2009.

[9] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

[10] WETTERICH, Caio Bruno. Uso de imagens de fluorescência para monitoramento da evolução do cancro cítrico. 2012. Dissertação (Mestrado em Física Básica) - Instituto de Física de São Carlos, University of São Paulo, São

Carlos, 2012. Disponível em:  
<<http://www.teses.usp.br/teses/disponiveis/76/76131/tde-23032012-151751/>>.  
Acesso em: 2016-05-31.

[11] WETTERICH, Caio Bruno et al. Detection of citrus canker and Huanglongbing using fluorescence imaging spectroscopy and support vector machine technique. *Applied Optics*, v. 55, n. 2, p. 400-407, 2016.

[12] WETTERICH, Caio Bruno et al. Detection of Citrus Canker and Huanglongbing Using Fluorescence Imaging Spectroscopy (FIS) Technique and Two Different Computational Methods: Support Vector Machine and Neural Network. In: *Frontiers in Optics*. Optical Society of America, 2015. p. FW5E. 4.

[13] Zen of Python. 2004. Disponível em:  
<<https://www.python.org/dev/peps/pep-0020/>>. Acesso em 31 mai. 2016.

