Projeto 2 - Fundamentos em Sistemas Inteligentes

Renan Godoi de Medeiros
Matricula: 15/0146612

***Abstract*—O projeto consiste na aplicação do algoritmo de Florestas Randômicas para classificação supervisionada em espécies de plantas com base em 14 parâmetros de suas folhas.**

I. Introdução

Este relatório tem por objetivo descrever o projeto da disciplina de Fundamentos em Sistemas Inteligentes e demonstrar as capacidades e limitações do algoritmo de classificação de Florestas Randômicas no âmbito de Sistemas Inteligentes.

Os tópicos foram divididos em 6 partes, o tópico II explica de forma clara os problemas que acarretaram no desenvolvimento do projeto assim como as dificuldades encontradas no mesmo.O tópico IV explica a razão pela qual tentei resolver os problemas citados em II.

Ainda no tópico III eu descrevo os materiais utilizados para a realização do projeto, assim como os métodos que eu utilizei no desenvolvimento do mesmo, em seguida, no tópico V é apresentada de forma clara como o projeto foi desenvolvido e também como ele foi dividido dentro do código do software. No tópico VI descrevo os resultados encontrados após uma série de testes e pesquisas a respeito.E por fim em VII apresento as minhas conclusões a respeito do trabalho, assim como as experiências que este trabalho me proporcionou.

II. Problema

Os problemas que acarretaram o trabalho foi a forma com que seriam analisados os fatores de classificação para as árvores, no geral para a construção das árvores é necesario ter uma plena idéia da quantidade de árvores a serem construidas assim como ás métricas das quais as àrvores vão ser analisadas pelo algoritmo, e dependendo do caso, perde-se a eficácia do processamento do algoritmo assim como a precisão dos dados aos quais estão sendo analisados.

No caso da linguagem que escolhi, me permitiu utilizar diversas formas para analisar os resultados do algoritmo, oque me deixou um pouco confuso em relação a escolha da metodologia,entretanto escolhi a que mais me convinha e que achava que tinha mais relação com oque estava sendo pedido no trabalho.

A documentação foi algo que não me ajudou muito neste trabalho, tive que fazer uma pesquisa acirrada em fórums e sites para conseguir entender de forma clara as funcionalidades e métodos que a linguagem fornecia para uso.

O tempo não foi um grande problema, já que o prazo foi razoável, mas as outras tarefas de outras disciplinas além desta me prejudicaram bastante, fora os problemas que tive com minha máquina que estava trabalhando, apesar disso consegui realizar o trabalho me dedicando bastante em algumas partes do dia.

III. Materiais e Métodos

Para este trabalho, utilizei uma máquina windows para desenvolver o código, escolhi a linguagem R para implementação, pois ela é a mais adequada para o uso estatístico em geral, assim como a ferramenta RStudio, para desenvolver o software. Utilizei também o software de modelagem Asta para modelar a sua arquitetura.

Também fiz o uso do github para garantir caso algum problema acontecesse na minha máquina, utilizei a biblioteca *randomForest* assim como o metódo randomForest para a classificação dos dados.

Para um bom treinamento e validação dos dados, fiz a divisão de 70:30, aos quais 70% foram destindados ao treinamento do algoritmo e 30% destinados à validação.

Utilizei também a classe das plantas como fator de classificação, pois na maioria das vezes quando tentava utilizar os outros atributos normalmente ocorriam erros incomuns quando os resultados eram processados.

IV. Porque

As razões pelas quais quis resolver este problema foi meu interesse na implementação do projeto, pois de certa forma inteligência artificial implementada na prática é algo que não é visto muito no curso, e se formos analisar, praticamente inteligência artificial é utilizada em muitas áreas da tecnologia da informação, desde pequenos sites, até coisas mais complexas, como as novas arquiteturas de placas gráficas fabricadas pela empresa Nvidia a qual possui uma inteligência artificial responsável por ajudar a renderizar os gráficos de jogos e melhorar a imersão dos jogadores.

A existência de várias soluções para o problema deste projeto, foi um dos fatores que me incentivou em minhas escolhas pois para o problema de florestas randômicas, a aplicabilidade deste algoritmo é imensa, logo é possível reunir muitas soluções para diferentes linguagens, inclusive linguagens que não são necessáriamente voltadas a estatística.

Apesar das florestas randômicas não ser a melhor escolha para a solução de problemas desse tipo já que existem muitos algoritmos capazes de resolver este problema de forma talvez até mais eficiente, a proposta deste trabalho é mostrar que as florestas randômicas podem ser úteis talvez não necessariamente na eficácia mas sim na precisão, trazendo resultados coerentes com a menor possibilidade de erro possível.

V. Implementação

Para implementação, primeiro pensei em deixar uma arquitetura equivalente com que havia feito no projeto anterior,com o processamento herdando da leitura de arquivos,e

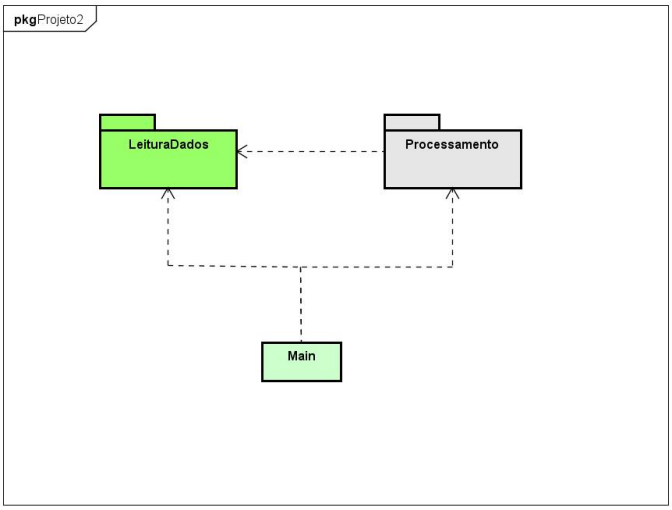

Fig. 1. Diagrama geral

a main herdando tanto o processamento quanto a leitura, com algumas diferenças na parte tanto da leitura dos dados, como também no processamento, já que os algoritmos e dados processados seriam relativamente diferentes do projeto anterior.

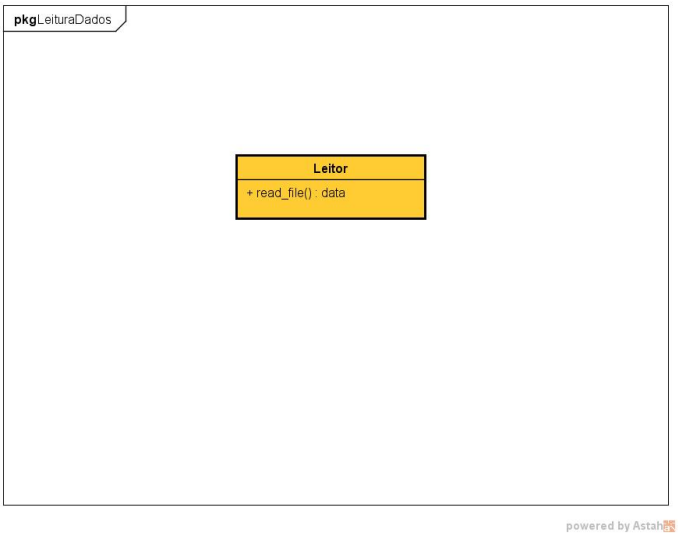

Fig. 2. Esquematização da leitura dos dados

Começando pela leitura representado na figura 2, o arquivo em formato .CSV contendo a tabela juntamente com os dados que serão lidos, basicamente o leitor fará a leitura de todos os dados contidos na tabela através da função *read_file()*, e retorna uma matriz correspondente com os dados lidos, contudo esta função retornará a matriz com as colunas nomeadas correspondentes aos atributos dos dados das folhas que serão analisados pelo módulo de processamento.

Para o processamento dos dados na figura 3, havia descrito apenas uma função inicialmente, entretanto como foi requerida

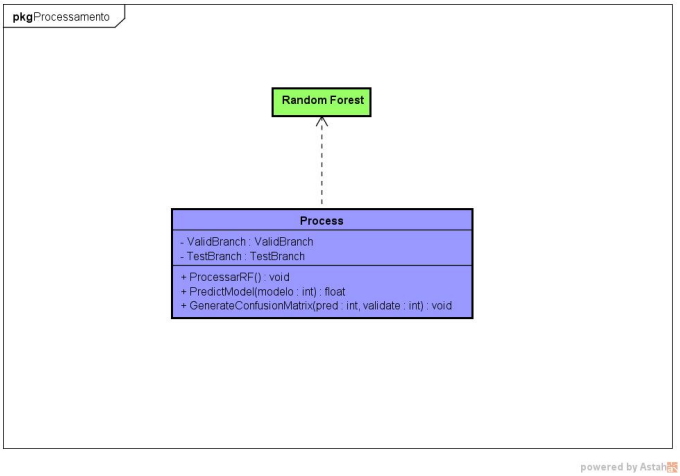

Fig. 3. Esquematização da processamento dos dados

uma validação cruzada de vários modelos, resolvi criar mais funções para facilitar a implementação do programa.Para o funcionamento correto é necessário que o módulo de processamento herde os métodos e atributos contidos na biblioteca *randomForest* disponível na linguagem R e descrito no modelo.

Eu criei 3 métodos para este módulo, que são *ProcessarRF()* que será responsável pelo processamento geral das florestas randômicas assim como a divisão de dados lidos no módulo de leitura, aos quais serão divididos entre teste e validação, e armazenados nas variáveis ValidBranch e TestBranch contidas dentro do próprio módulo de processamento. O método *PredictModel(modelo)*, ao qual é responsável por fazer a predição dos modelos processados pela função de processamento das florestas, esta função também retorna um valor correspondente a precisão da análise naquele modelo ao qual foi recebido de argumento.

O último método *GenerateConfusionMatrix(pred,valid)* como a sua própria nomenclatura sugere, ela é responsável por gerar uma matriz de confusão referente ao modelo processado pelo método de processamento da floresta, recebendo como parâmetro a predição do algoritmo e os valores correspondentes a validação.

VI. Resultados

Os resultados saíram como o esperado com quase todos os modelos tendo uma taxa de acerto acima dos 60%, chegando a casa dos 70%, para a obtenção de diferentes resultados na análise das florestas, utilizei diferentes parâmetros na função randomForest do R, alterando dois parâmetros que corresponderiam ao número de árvores a serem criadas, e o número correspondente às váriaveis randômicas que são selecionadas pelo algoritmo a medida que a árvore é processada.

Foram criados e testados no total 10 modelos diferentes para a classificação, é perceptível que na figura 4 ao qual apresenta um grafico relacionando a taxa de acerto com o índice dos modelos, para cada modelo era modificada o parametro de

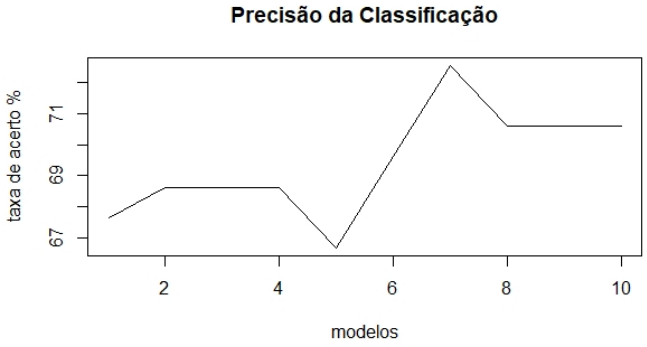

Fig. 4. Precisão da Classificação de florestas randômicas

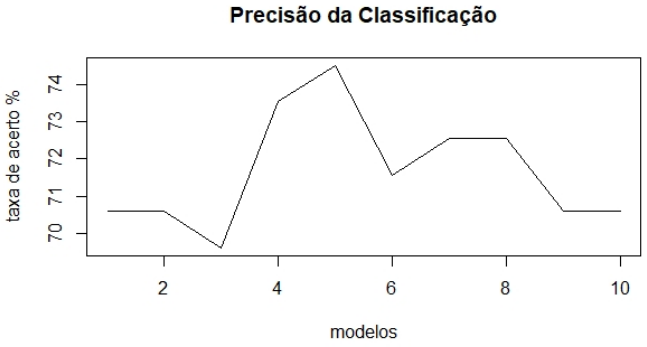

Fig. 6. Precisão da Classificação de florestas randômicas para uma semente diferente

atributos aleatórios e o numero de arvores processados, é perceptível que o algoritmo ficou mais preciso não por causa do número de árvores analisadas, mas sim possivelmente por conta dos atributos aleatórios analisados.No caso para os modelos próximos a 7 foram analisados cerca de 6 a 12 atributos, oque sugere que esse seria o equilíbrio mais balanceado para se analisar os resultados do algoritmo, entretanto dependendo do caso essas taxas podem variar a medida de vezes que os dados são processados pois as florestas são analisadas de forma aleatória então pode ser que o algoritmo obtenha mais precisão por exemplo no décimo ou no nono.

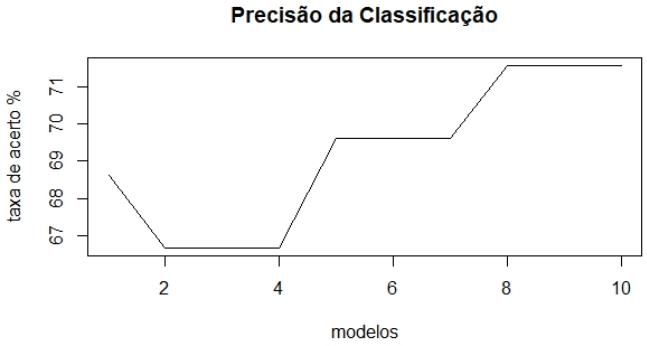

Fig. 5. Precisão da Classificação de florestas randômicas

Como é o caso da figura 5, onde ouve uma variação significativa das taxas de acerto, é perceptível que as mudanças entre os modelos 2 a 4 foram relativamente baixas em comparação com o teste anterior, entretanto ouve um pico alto entre os modelos 4 a 6, oque sugere uma melhora significativa, e é notório que agora o modelo mais preciso anteriormente não é o mais preciso neste gráfico, e também podemos perceber que existem mais de um modelo que seria considerado o mais preciso do gráfico.

Neste caso da figura 6, com uma simples alteração no valor da semente processada, foi possível obter resultados complemente diferentes, uma variação muito grande em relação aos dois gráficos anteriores, como pode-se observar a taxa de acerto dos últimos modelos onde tecnicamente teriam a maior quantidade de variáveis aleatórias obteve os piores resultados em relação aos outros, o melhor observado no gráfico seria o modelo 5 que teria a maior taxa de acerto, no entanto, ele não possui a maior quantidade de variáveis aleatórias calculadas pelo algoritmo, oque sugeriria que a semente pode influênciar nos resultados.

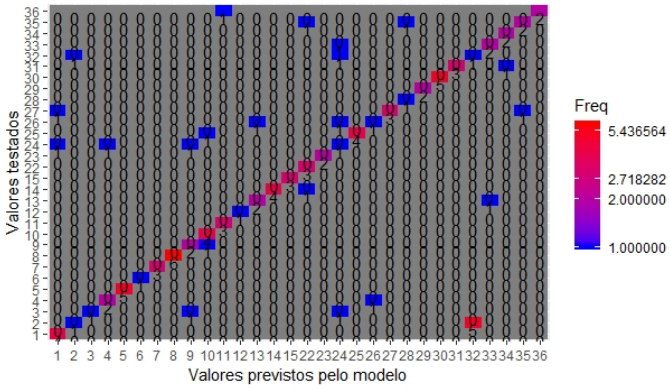

Fig. 7. Matriz de Confusão do melhor modelo

Agora observando-se a matriz de confusão do melhor modelo na figura 7, apesar da ilegibilidade devido ao seu tamanho, podemos analisar a frequência de acerto de cada uma das classes de folhas testadas, é notório que o maior erro se deu na classe 2, onde aparentemente o algoritmo confundiu com a classe 32,presumindo que a classe 2 e 32 pertenceriam a mesma classe de folha,entretanto para os outros valores testados, não ouve problemas na classificação dos resultados.

VII. Conclusão

Neste trabalho abordei o assunto de sistemas inteligentes, com o objetivo de mostrar o funcionamento do algoritmo de classificação de Florestas Randômicas, analisando os resultados obtidos, analisando a performance e se eram satisfatoriamente bons ou não.Também foi feita uma visão geral de como o projeto foi realizado,além dos softwares e IDE's que foram utilizados na sua construção.

Felizmente consegui cumprir todos os requisitos do trabalho, apesar das dificuldades encontradas em relação ao algoritmo, foi possível a sua realização graças não somente ao material disponibilizado na plataforma moodle da disciplina, mas também aos fórums e blogs de pesquisa e ajuda em relação ao assunto.

Este trabalho foi muito importante para o meu conhecimento, aprendi várias coisas em relação a árvores randômicas .O projeto levou bastante tempo para ser construido, mas apesar das dificuldades foi interessante, a àrea de inteligência artificial é muito vasta principalmente quando pensamos em sistemas.

REFERENCES

[1] Gareth James, Daniela Witten, Trevor Hastle, Robert Tibshirani An Introduction to Statistical Learning with applications in R, 2014

[2] How To Implement Random Forests in R https://www.r-bloggers.com/how-to-implement-random-forests-in-r/