UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
APRENDIZADO DE MÁQUINA - INF01017

Trabalho 1 - Florestas Aleatórias

Juei Hao Weng - 218768
Leonardo Barlette de Moraes - 219826
Leonardo Heitich Brendler - 218766

Sumário

Sumário

Objetivo

- Implementação do algoritmo de **Florestas Aleatórias (*Random Forests*)** para tarefas de classificação:
  - Seguindo o paradigma de aprendizado *ensemble* (múltiplos modelos);

- Utilização da metodologia de **validação cruzada estratificada (*cross-validation*)**:
  - O objetivo é avaliar o desempenho do modelo e o efeito de diferentes valores de parâmetros no aprendizado do algoritmo.
    - Parâmetro otimizado neste trabalho é o **número de árvores no *ensemble* (*ntree*)**;

Sumário

Implementação - Descrição geral

2.1.1 Estruturas de dados utilizadas para armazenar as árvores

Para a construção das árvores de decisão foi utilizada a biblioteca *data.tree.* O bloco básico das árvores são os **objetos do tipo *Node*** que possuem:

atributo: Pode ser um *ativo*, um campo ou método;
ativo: Um campo no nodo que pode ser chamado como um atributo, mas se comporta como um método sem argumentos (e.g. node$position);
campo: Um valor no nodo (e.g. node$cost ← 2500)
método: Um método agindo em um objeto, nesse caso um nodo (e.g. node$revert()).
herança: Quando o nodo herda um atributo de um de seus ancestrais.

Implementação - Descrição geral

2.1.2 Classificação de novas instâncias

- A instância de entrada a ser avaliada é utilizada para percorrer cada árvore do ensemble:
  - Para percorrer a árvore, os nomes dos nodos (bifurcações) são lidos.

- Se o atributo da instância de entrada com nome igual ao nodo for **numérico**:
  - O nome de um de seus filhos é lido e o valor do atributo é testado (maior ou igual ou menor);
    - Dependendo do resultado da comparação, um dos nodos filhos é escolhido e o método é chamado novamente, de maneira recursiva.

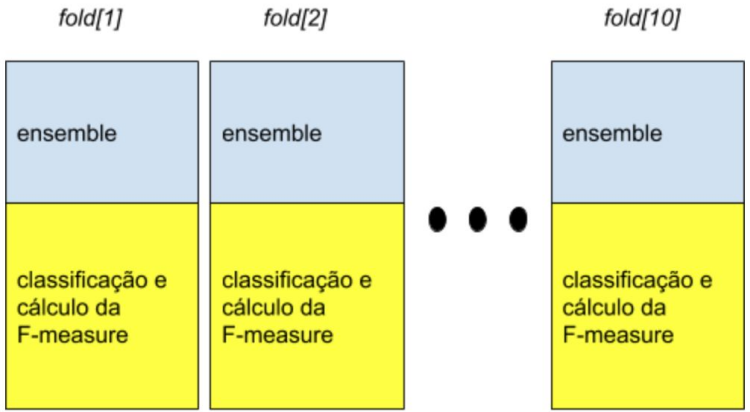
- Se o atributo da instância de entrada com nome igual ao nodo é **categórico**:
  - O nodo filho que possuir o mesmo valor que o atributo da instância de teste é escolhido, e o método é chamado novamente de forma recursiva;
    - Consequentemente, caso o nodo seja folha, o seu valor é retornado (no caso, a classe a qual a instância foi classificada).

- Por fim, as classificações de cada árvore do *ensemble* são reunidas e é realizado uma **votação majoritária** decidindo a classificação final do algoritmo.

Implementação - Descrição geral

2.1.3 Otimizações no algoritmo

Não foram realizadas otimizações de baixo nível ao algoritmo. Contudo, as classificações para cada *fold* de teste diferentes (k=10) foram realizadas em paralelo utilizando as bibliotecas *parallel, iterators, foreach e doParallel*. O cálculo da F-measure também é realizado em paralelo, depois de cada geração de *ensemble*.



Cada geração de ensemble é realizada em paralelo para diferentes folds de treinamento, com a classificação das entradas de treinamento e cálculo do desempenho do ensemble sendo realizado em seguida.

Sumário

Resultados - Estrutura final da árvore induzida

Árvore resultante do *Benchmark*:

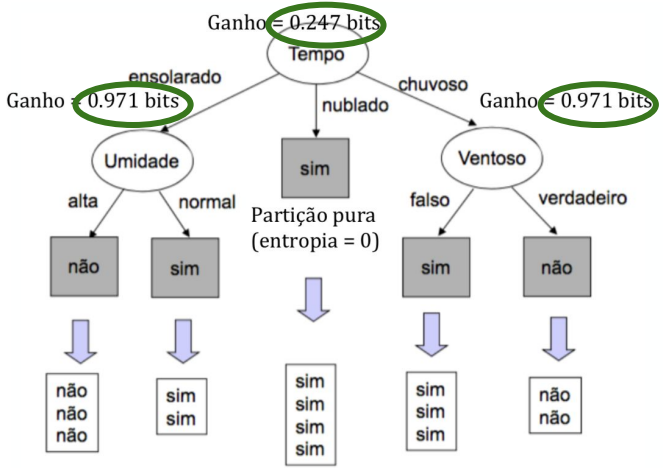
```
Tempo
 ¦--Ensolarado
 ¦   °--Umidade
 ¦       ¦--Alta
 ¦       ¦   °--Nao
 ¦       °--Normal
 ¦           °--Sim
 ¦--Nublado
 ¦   °--Sim
 °--Chuvoso
     °--Ventoso
         ¦--Falso
         ¦   °--Sim
         °--Verdadeiro
             °--Nao
```

Atributo Escolhido	Ganho de Informação
Tempo	0.246749819774439
Umidade	0.970950594454669
Ventoso	0.970950594454669

Resultados - Estrutura final da árvore induzida

Árvore resultante do *Benchmark*:

Tempo



°--Nao

Atributo Escolhido	Ganho de Informação
Tempo	0.246749819774439
Umidade	0.970950594454669
Ventoso	0.970950594454669

ESTRUTURA CORRETA

Sumário

Resultados - Análise de desempenho

Valores do **parâmetro *ntree***: 10, 25, 50, 100, 125, 150;

Uso da F1-measure;

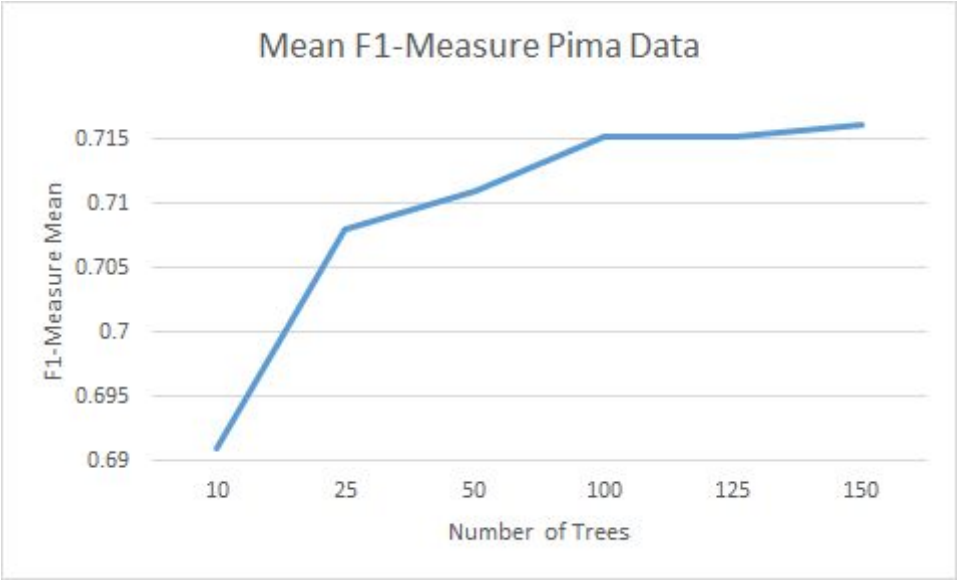
Duas análises:

Desempenho do algoritmo através das **médias da F1-measure** para diferentes valores do *ntree*;

Distribuição da F1- measure em relação ao *ntree*;

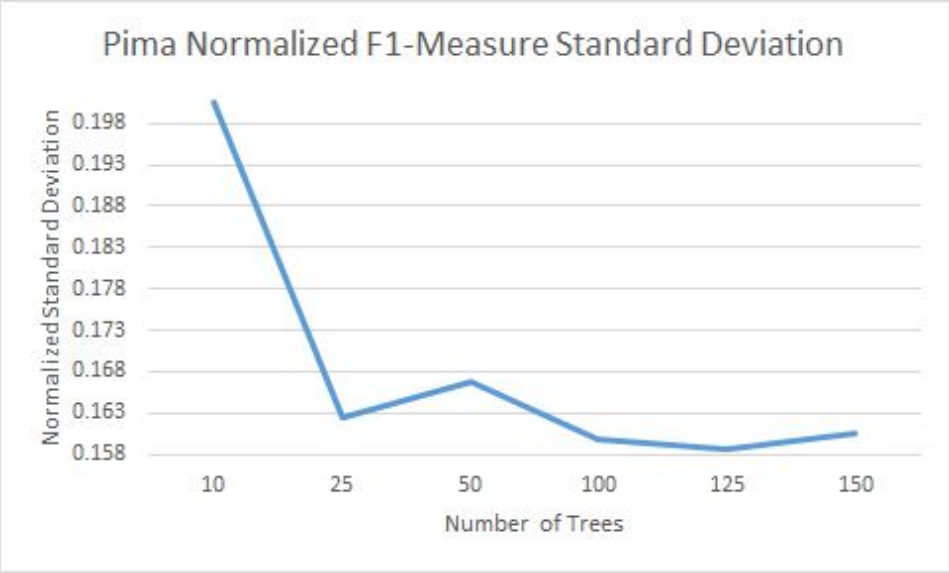
Resultados - Análise de desempenho

3.2.1 Pima Indian Diabetes Data Set (8 atributos, 768 exemplos, 2 classes)



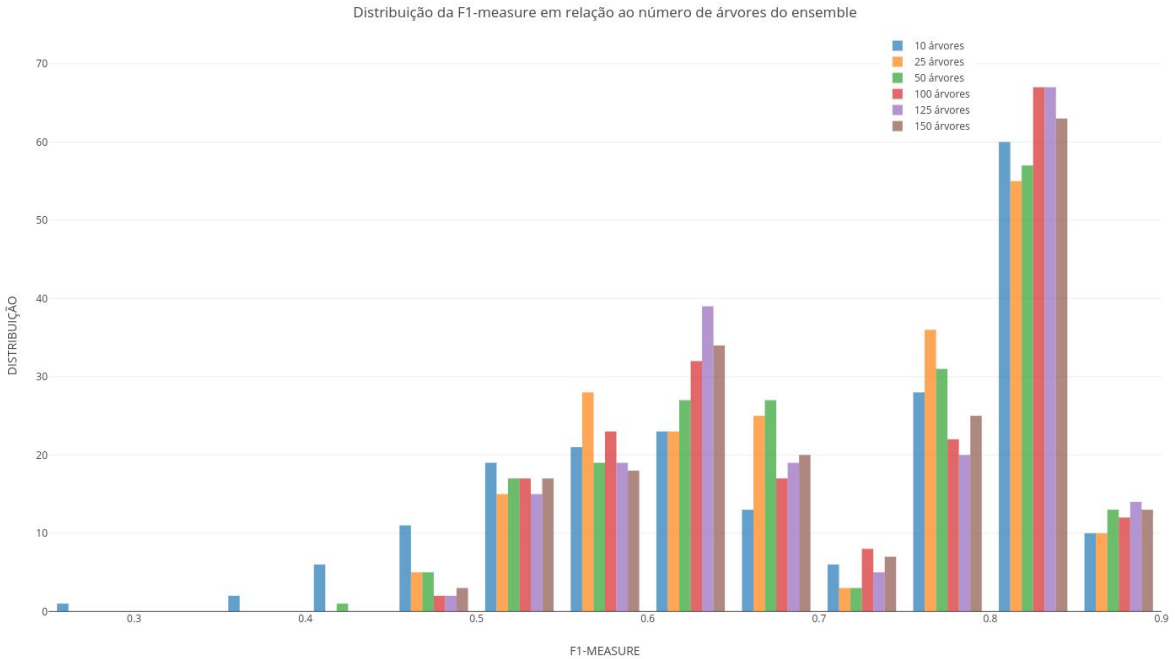
Resultados - Análise de desempenho

3.2.1 Pima Indian Diabetes Data Set (8 atributos, 768 exemplos, 2 classes)



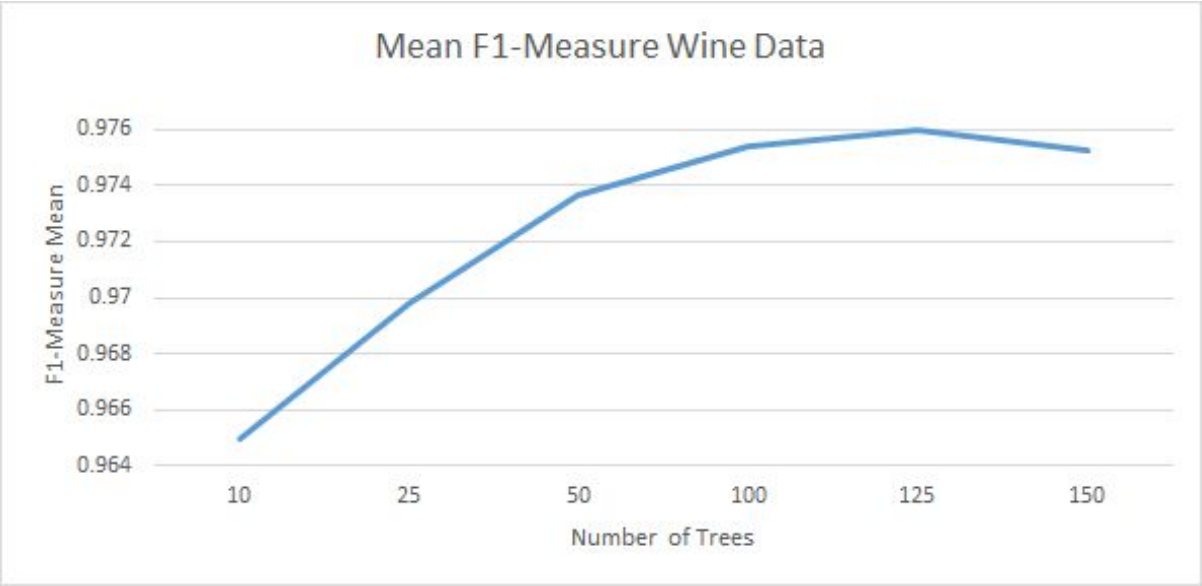
Resultados - Análise de desempenho

3.2.1 Pima Indian Diabetes Data Set (8 atributos, 768 exemplos, 2 classes)



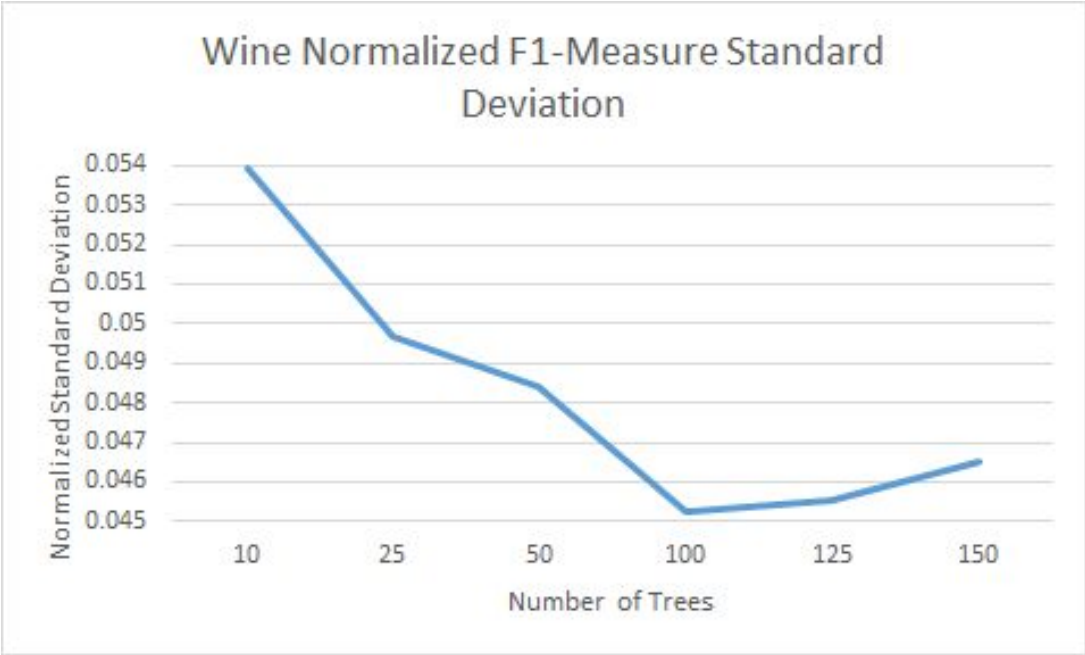
Resultados - Análise de desempenho

3.2.2 Wine Data Set (13 atributos, 178 exemplos, 3 classes)



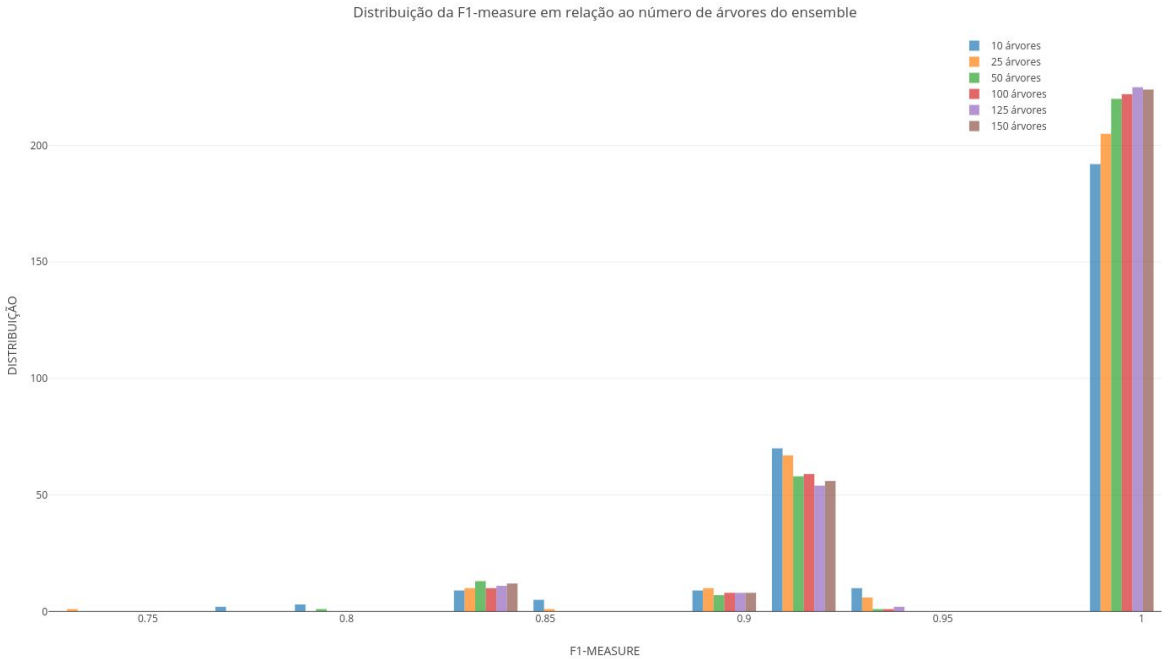
Resultados - Análise de desempenho

3.2.2 Wine Data Set (13 atributos, 178 exemplos, 3 classes)



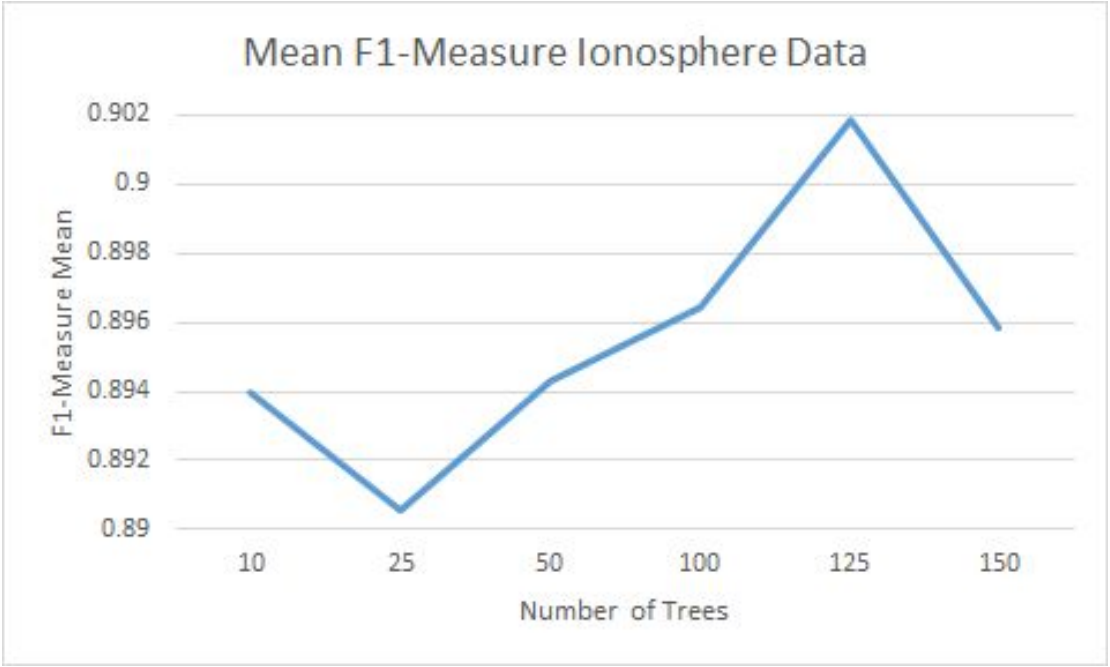
Resultados - Análise de desempenho

3.2.2 Wine Data Set (13 atributos, 178 exemplos, 3 classes)



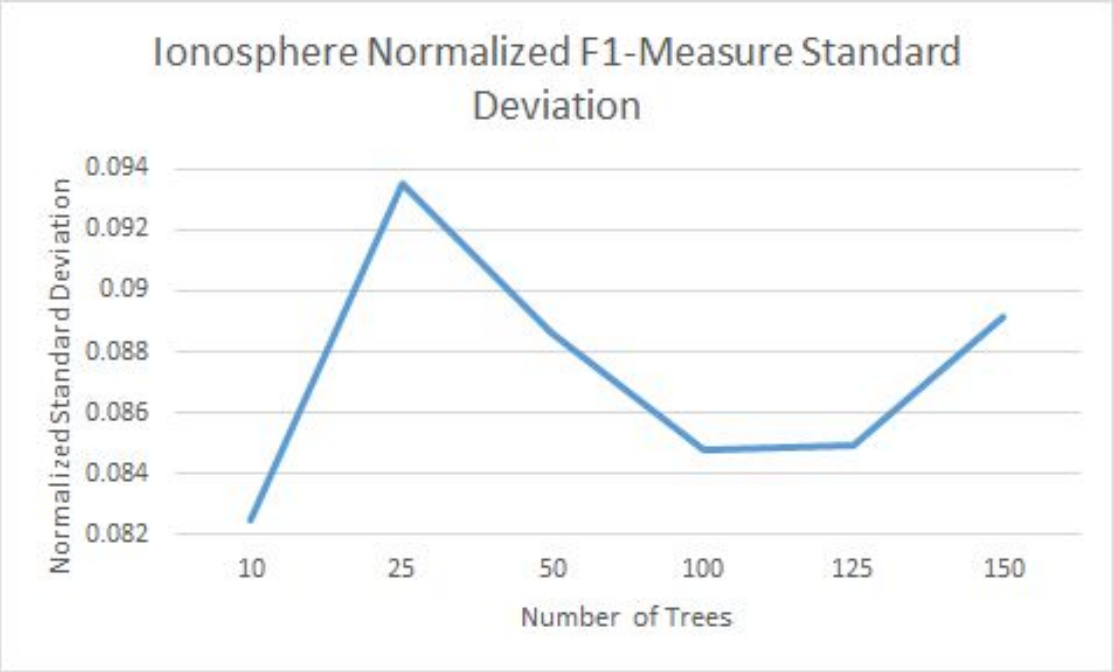
Resultados - Análise de desempenho

3.2.3 Ionosphere Data Set (34 atributos, 351 exemplos, 2 classes)



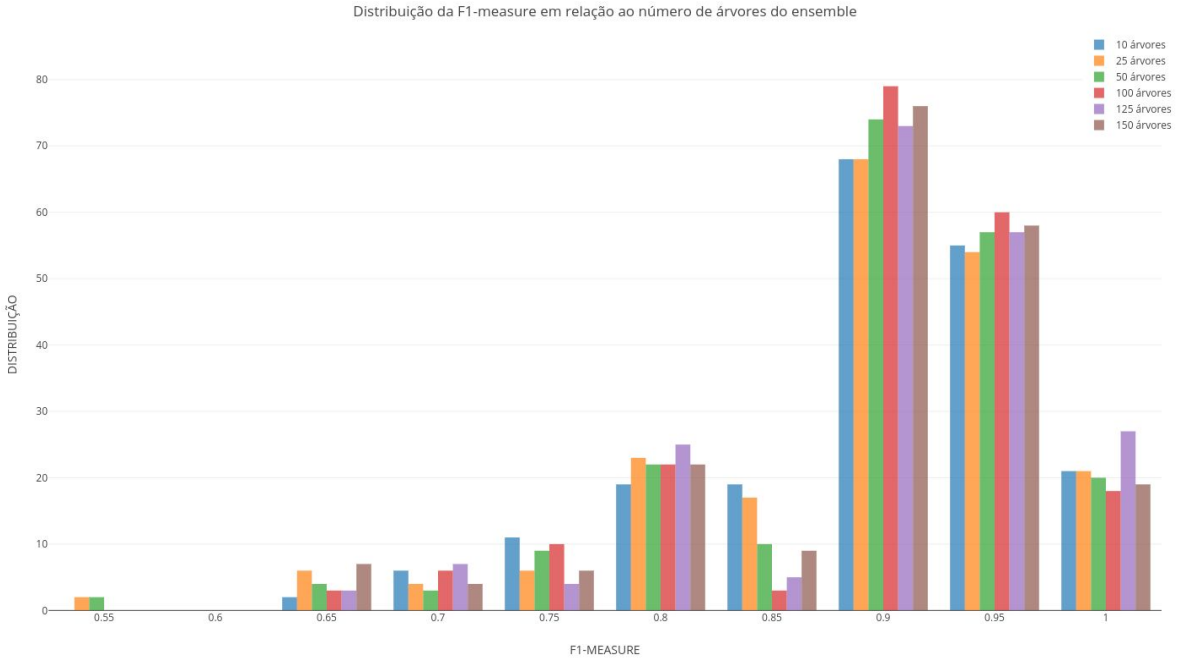
Resultados - Análise de desempenho

3.2.3 Ionosphere Data Set (34 atributos, 351 exemplos, 2 classes)



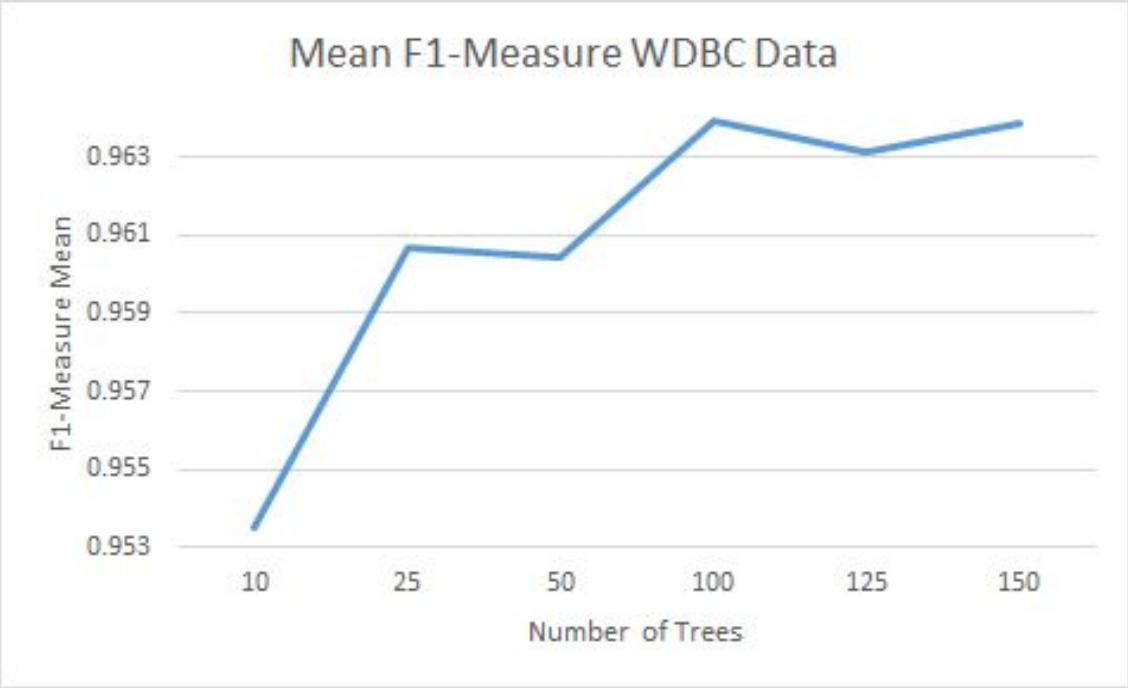
Resultados - Análise de desempenho

3.2.3 Ionosphere Data Set (34 atributos, 351 exemplos, 2 classes)



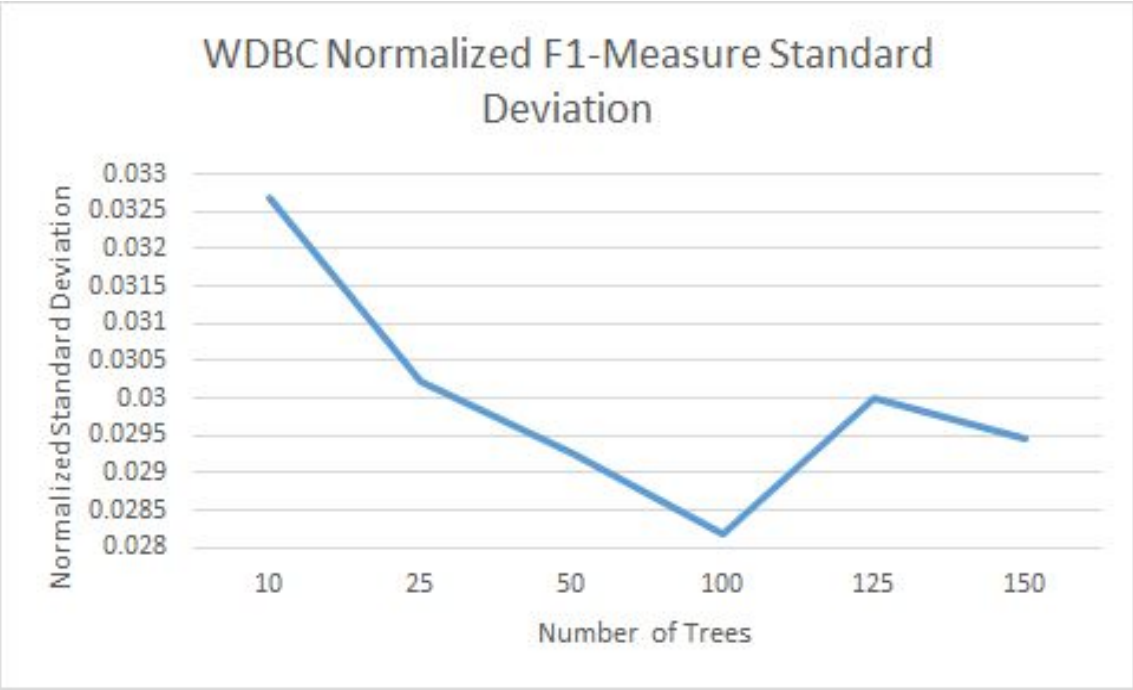
Resultados - Análise de desempenho

3.2.4 Breast Cancer Wisconsin (32 atributos, 569 exemplos, 2 classes)



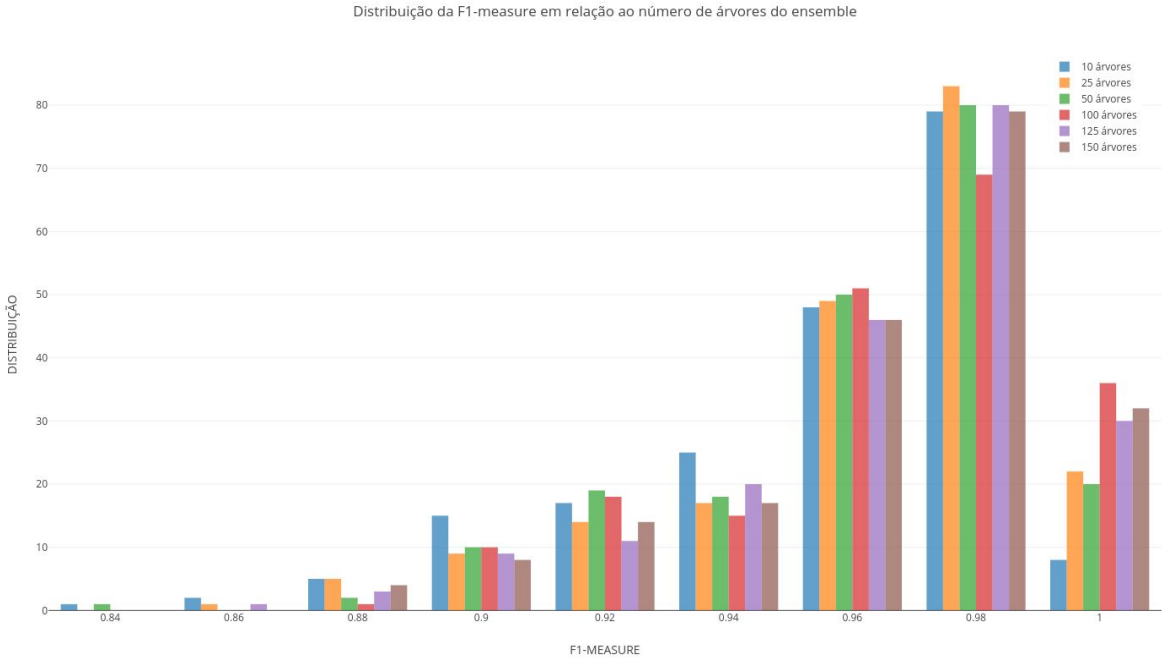
Resultados - Análise de desempenho

3.2.4 Breast Cancer Wisconsin (32 atributos, 569 exemplos, 2 classes)



Resultados - Análise de desempenho

3.2.4 Breast Cancer Wisconsin (32 atributos, 569 exemplos, 2 classes)



Sumário

Conclusões

- Desempenho do algoritmo através das **médias da F1-measure** para diferentes valores do *ntree*:
    - Diminuição no erro de classificação de acordo com o aumento no número de árvores no modelo de Florestas Aleatórias;
    - Em alguns casos houve diminuição da F-Measure devido às escolhas aleatórias de atributos a cada bifurcação da árvore onde atributos com baixo ganhos de informação foram escolhidos;
    - Desvio padrão diminui de acordo com o aumento do número de árvores:
        - E acompanha as más escolhas de atributos aumentando seu valor
        - Em alguns casos há o aumento da F-measure e o aumento do desvio, é possível concluir que quanto maior o tamanho do *ensemble,* menor o impacto que a má escolha de atributos tem sobre o desempenho.

Conclusões

- **Distribuição da F1- measure** em relação ao *ntree*:
  - Quanto menor o número de árvores, maior a distribuição de resultados mais baixos da F1-measure;
  - Identificação da presença das maiores distribuições de resultados mais altos da F1-measure:
    - Nos modelos observados os resultados estão presentes para valores mais altos de *ntree*, o que era esperado.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
APRENDIZADO DE MÁQUINA - INF01017

Trabalho 1 - Florestas Aleatórias

Juei Hao Weng - 218768
Leonardo Barlette de Moraes - 219826
Leonardo Heitich Brendler - 218766