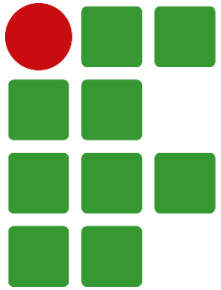


INSTITUTO FEDERAL
Paraíba
Campus Campina Grande

COMPUTANDO *ENSEMBLE METHODS* PARA PREDIZER EVASÕES ESTUDANTIS

RODOLFO BOLCONTE DONATO

Orientadora: Samara Martins Nascimento

Coorientador: Gustavo Wagner Diniz Mendes

SUMÁRIO

- Introdução:
  - Evasão Estudantil;
  - Justificativa;
  - Objetivos;
  - Classificação Automática de Dados;
  - Aprendizado de Máquina;
- Planejamento:
  - Preparação dos Dados;
  - Algoritmos Utilizados;
  - Métricas Estatísticas;
  - Metodologia Experimental.
- Resultados Obtidos:
  - Comparação dos Testes 1 e 2;
  - Comparação dos Testes 3 e 4;
  - Comparação dos Testes 5 e 6.
- Considerações Finais e Pesquisas Futuras

EVASÃO ESTUDANTIL

Fonte: https://sambatech.com/

EVASÃO ESTUDANTIL NO IFPB *CAMPUS* CAMPINA GRANDE

- Custo de uma evasão:
  - 1 aluno custa R$ 3,7 mil por mês;
  - 10 alunos custam R$ 444 mil por ano.
- Alunos do Curso Superior de Tecnologia (CST) em Telemática de 2007 a 2016:
  - 839 matrículas realizadas;
  - 439 matrículas evadidas.

JUSTIFICATIVA

- Trabalho de Mestrado da Universidade Federal de Campina Grande (UFCG) para a previsão de evasões estudantis, testando duas estratégias computacionais [Melo 2016];
- Projeto de pesquisa para previsão de evasões estudantis nos cursos do Instituto Federal da Paraíba (IFPB) *campus* Campina Grande, com dois grupos de testes [Bolconte e Mendes 2017].

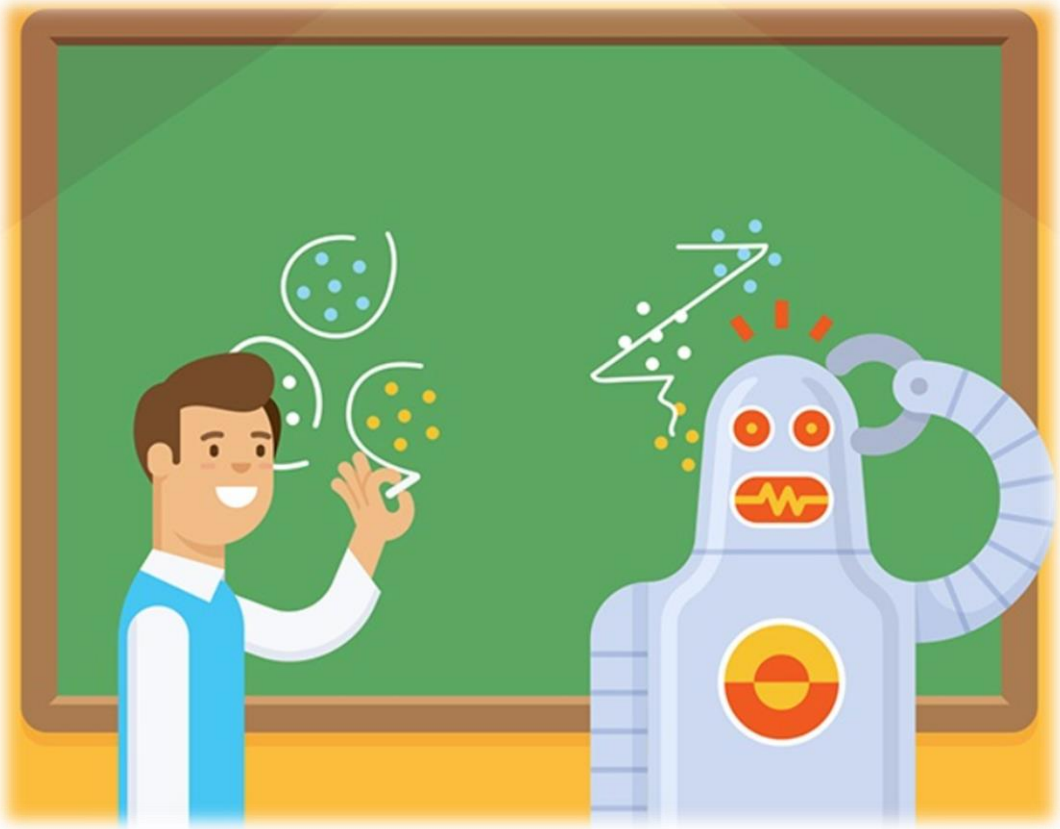
OBJETIVOS

- Geral:
  - Comparar e definir o *Ensemble Method* mais adequado para a previsão automática de evasões estudantis do CST em Telemática do IFPB *campus* Campina Grande.
- Específicos:
  - Definir atributos descritivos de evasões;
  - Caracterizar o funcionamento dos algoritmos utilizados;
  - Testar e comparar o desempenho dos *Ensemble Methods* utilizando o mesmo Algoritmo de Classificação;
  - Definir o *Ensemble Method* mais adequado para a previsão de evasões num conjunto específico.

CLASSIFICAÇÃO AUTOMÁTICA DE DADOS

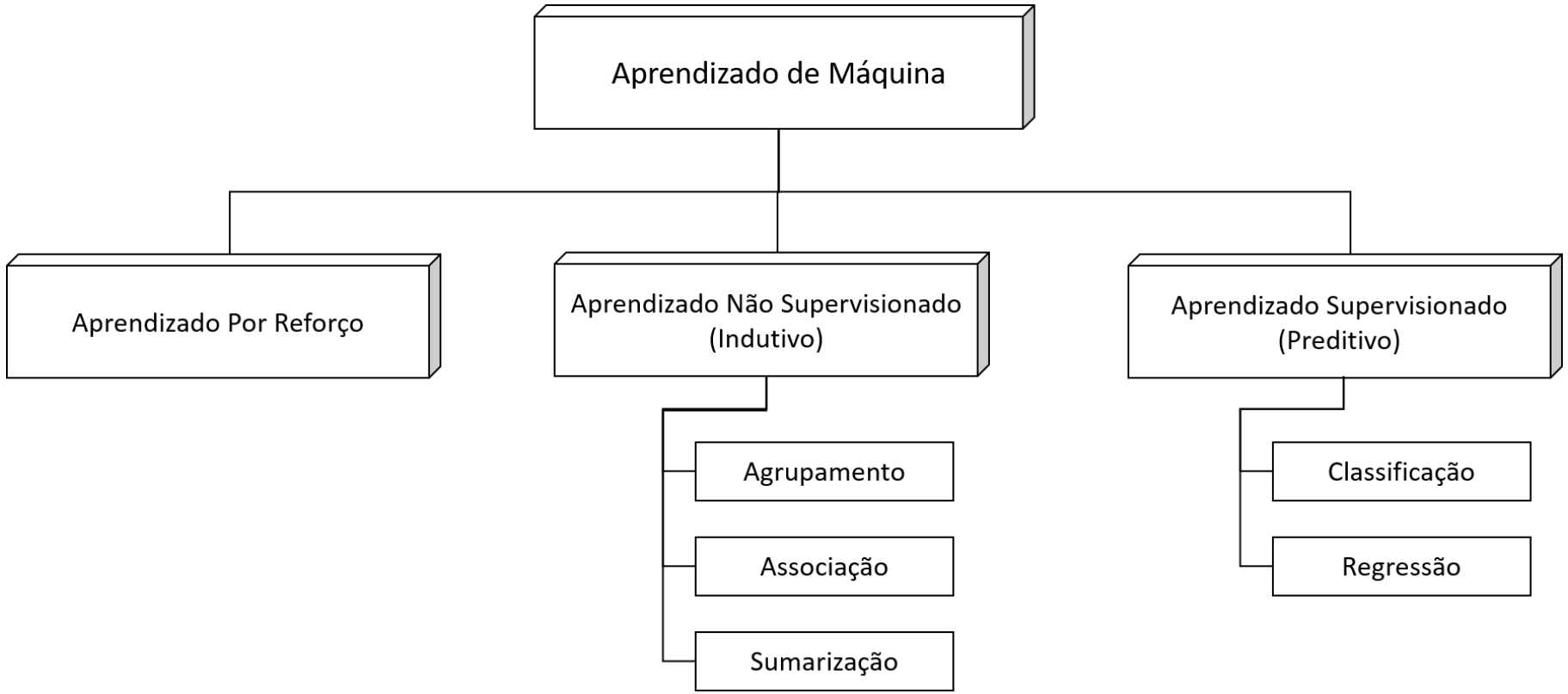
Fonte: https://resultato.com.br/

APRENDIZADO DE MÁQUINA



Fonte: https://lamfo-unb.github.io/

APRENDIZADO DE MÁQUINA



Fonte: Elaboração Própria.

PREPARAÇÃO DOS DADOS

- Dados da plataforma *QAcadêmico* disponibilizados como *backup* pelo Instituto Federal da Paraíba:
  - Dados de todos os *campi* e seus cursos até o ano de 2016;
  - Garantia de dados não redundantes através de correção de valores pelo sistema.
- Realização de filtragem dos dados de alunos referentes ao CST em Telemática (de 2007.1 a 2015.1):
  - Organização de atributos descritivos de evasão estudantil ou não;
  - Uma tupla de informações para uma matrícula;
  - Total de 720 matrículas, 429 evadidas e 291 não evadidas.

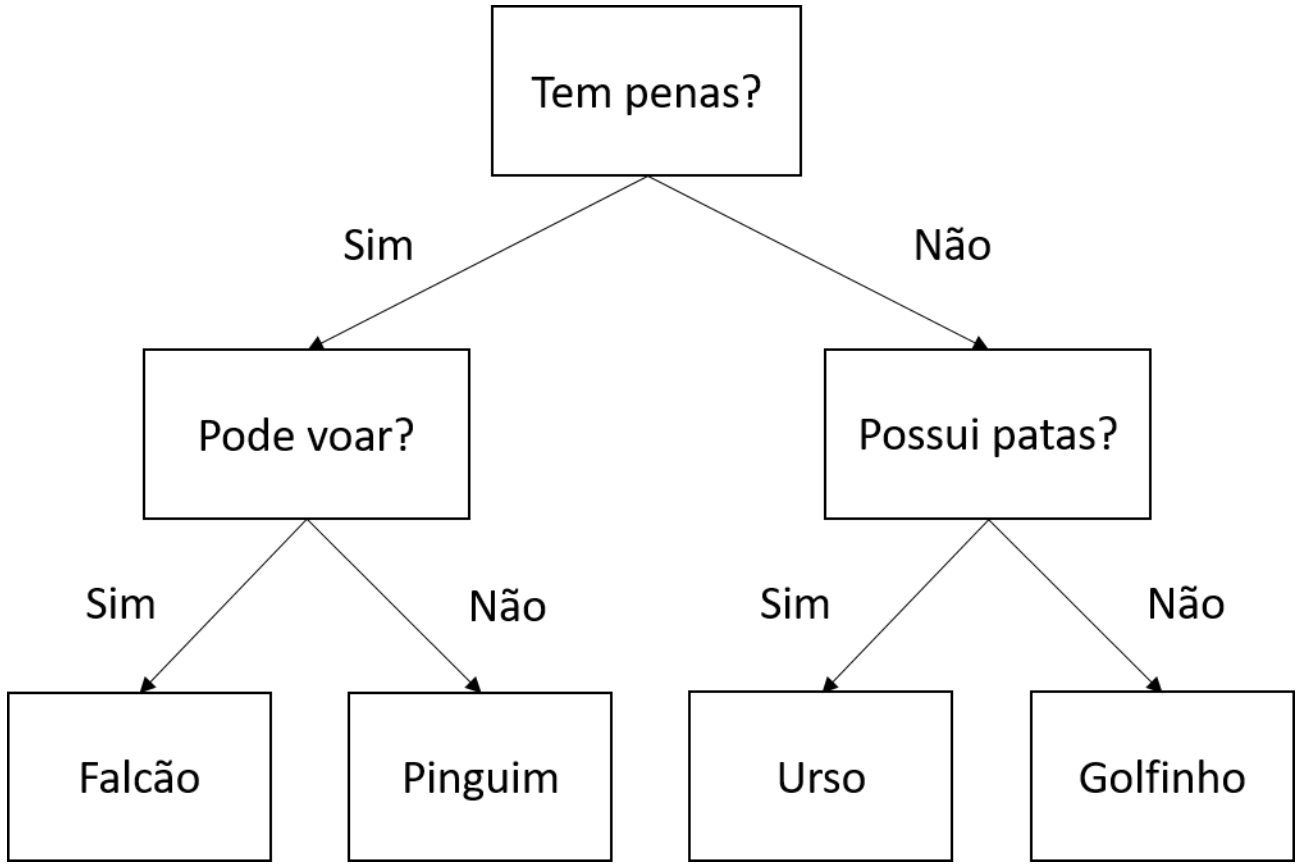
ATRIBUTOS DESCRITIVOS

- Atributos Quantitativos e Qualitativos;
- Atributos utilizados:
  - Porcentagem do Curso;
  - Coeficiente de Rendimento do Aluno;
  - Quantidade de Períodos Letivos;
  - Quantidade de Disciplinas do Curso;
  - Quantidade de Disciplinas (Aprovadas, Reprovadas por Nota, Reprovadas por Falta, Canceladas e Trancadas);
  - Evasão.

ALGORITMOS DE CLASSIFICAÇÃO

- Possibilidade de identificar a qual categoria já assimilada uma amostra pode pertencer;
- Algoritmos conhecidos:
  - *K-Nearest Neighbor;*
  - *Naive Bayes;*
  - *Support Vector Machine;*
  - Árvore de Decisão.

ALGORITMO ÁRVORE DE DECISÃO

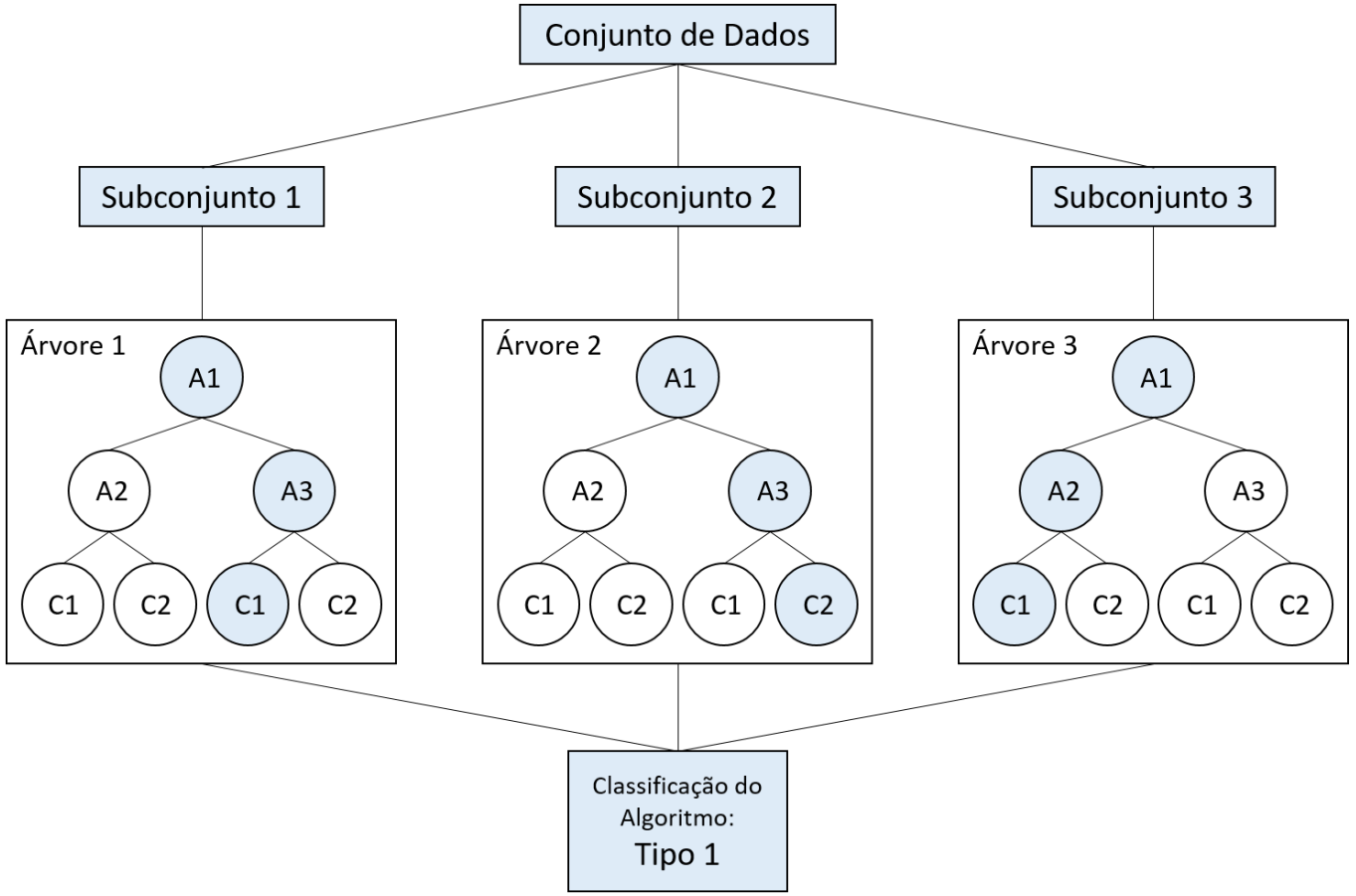


Fonte: Elaboração Própria.

ENSEMBLE METHODS

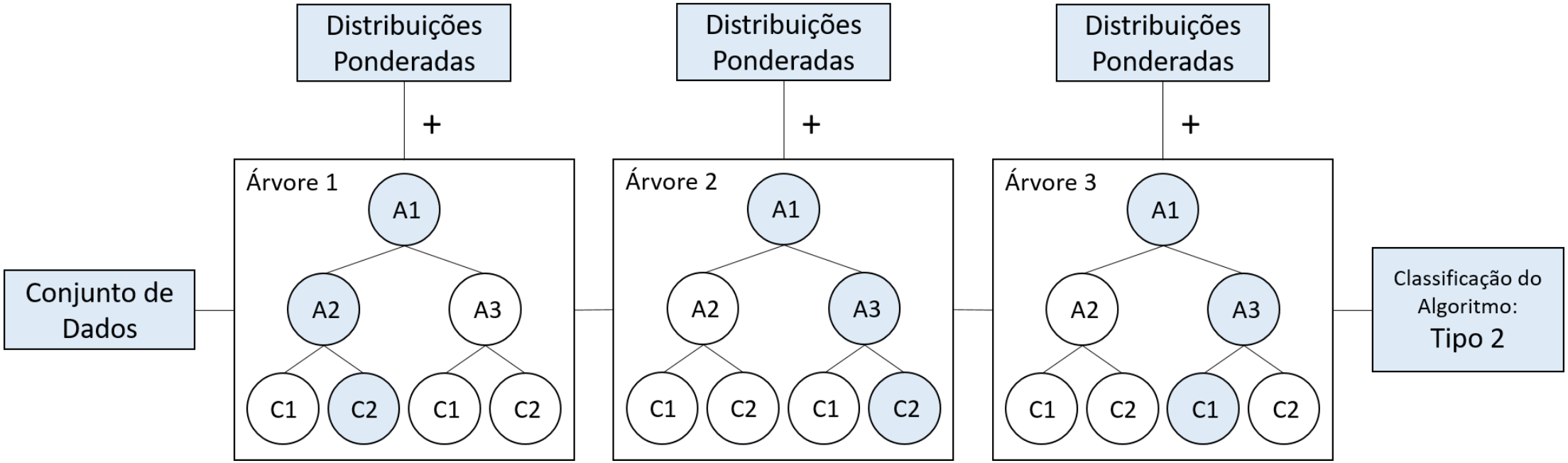
- Utilização de vários modelos para obter desempenho maior que apenas um, trabalhando dentro de limitações, como velocidade de processamento e tempos de retorno [Julian 2016];
- Se dividem em:
  - Algoritmos de Agregação;
  - Algoritmos de Impulso.

ALGORITMO FLORESTA ALEATÓRIA



Fonte: Elaboração Própria.

ALGORITMO AUMENTO DE GRADIENTE



Fonte: Elaboração Própria.

MATRIZ DE CONFUSÃO

		VALOR PREVISTO	
		POSITIVO	NEGATIVO
VALOR REAL	POSITIVO	(VP) VERDADEIRO POSITIVO	(FP) FALSO POSITIVO
	NEGATIVO	(FN) FALSO NEGATIVO	(VN) VERDADEIRO NEGATIVO

Fonte: Elaboração Própria.

- Para previsão de evasões:
  - VP: Evasões previstas corretamente;
  - FP: Não evasões previstas como evasões;
  - FN: Evasões previstas como não evasões;
  - VN: Não evasões previstas corretamente.

MÉTRICAS ESTATÍSTICAS

$$Acurácia = \frac{(VN + VP)}{(VN + FP + FN + VP)}$$

$$Precisão = \frac{VP}{(VP + FP)}$$

$$Sensibilidade = \frac{VP}{(VP + FN)}$$

$$Taxa\ de\ Falsa\ Previsão\ Positiva = \frac{FP}{(FP + FN)}$$

$$Tempo\ de\ Processamento = T_f - T_i$$

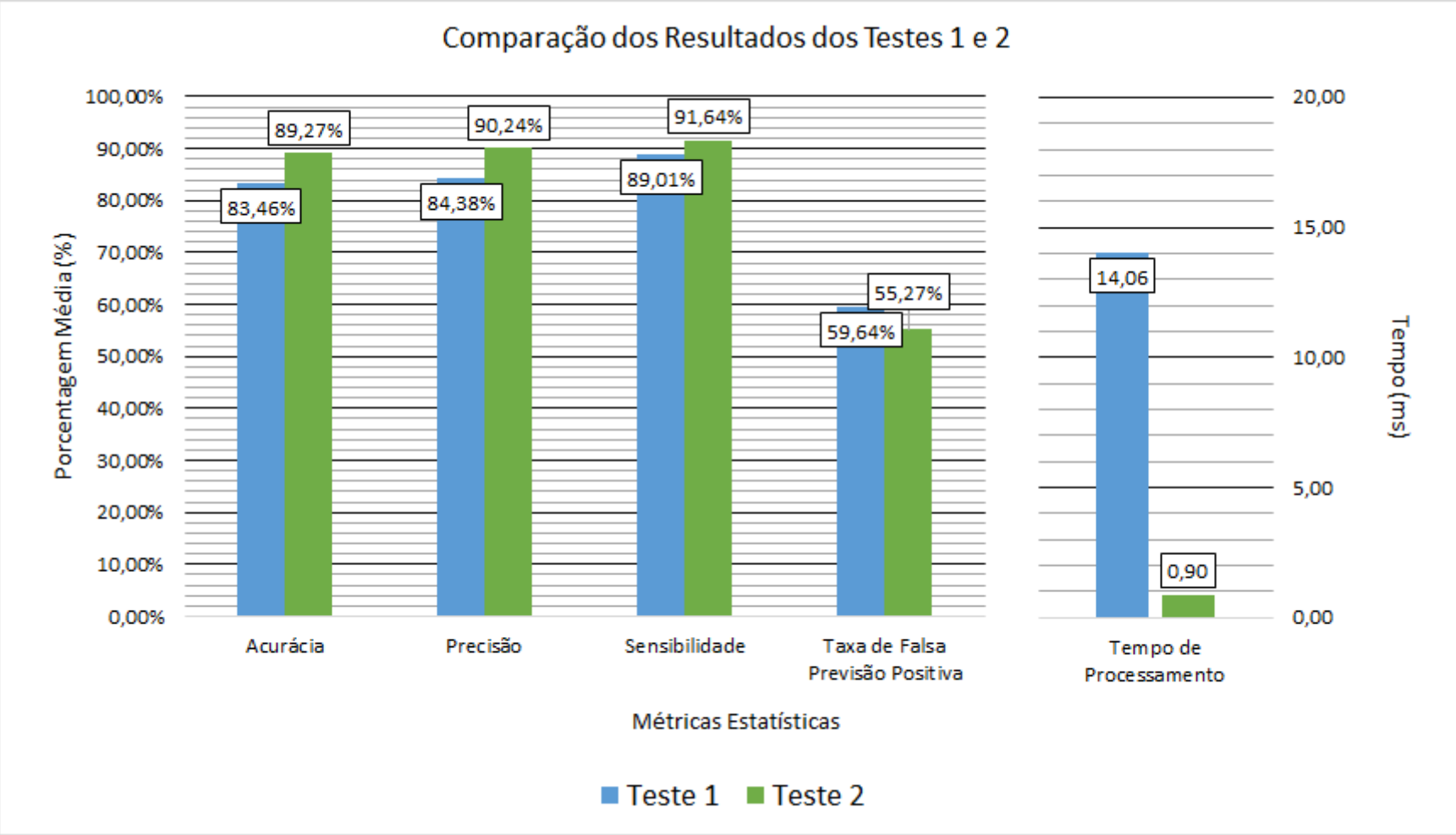
METODOLOGIA EXPERIMENTAL

- População do *QAcadêmico* com amostras referente ao CST em Telemática de 2007.1 a 2015.1;
- Conjunto de Dados divido com o método *Bootstrap* a cada Teste:
  - Dois Subconjuntos: um com 63,2% dos dados para Treino (455 amostras) e 36,8% para Teste (265 amostras).
- Cada Teste é executado 10 vezes, estendendo o método *k-fold*.

TESTES DESBALANCEADOS

- Testes com o Conjunto de Dados sem utilização de técnicas de Balanceamento de amostras:
  - Teste 1: Floresta Aleatória sem balanceamento de dados;
  - Teste 2: Aumento de Gradiente sem balanceamento de dados.

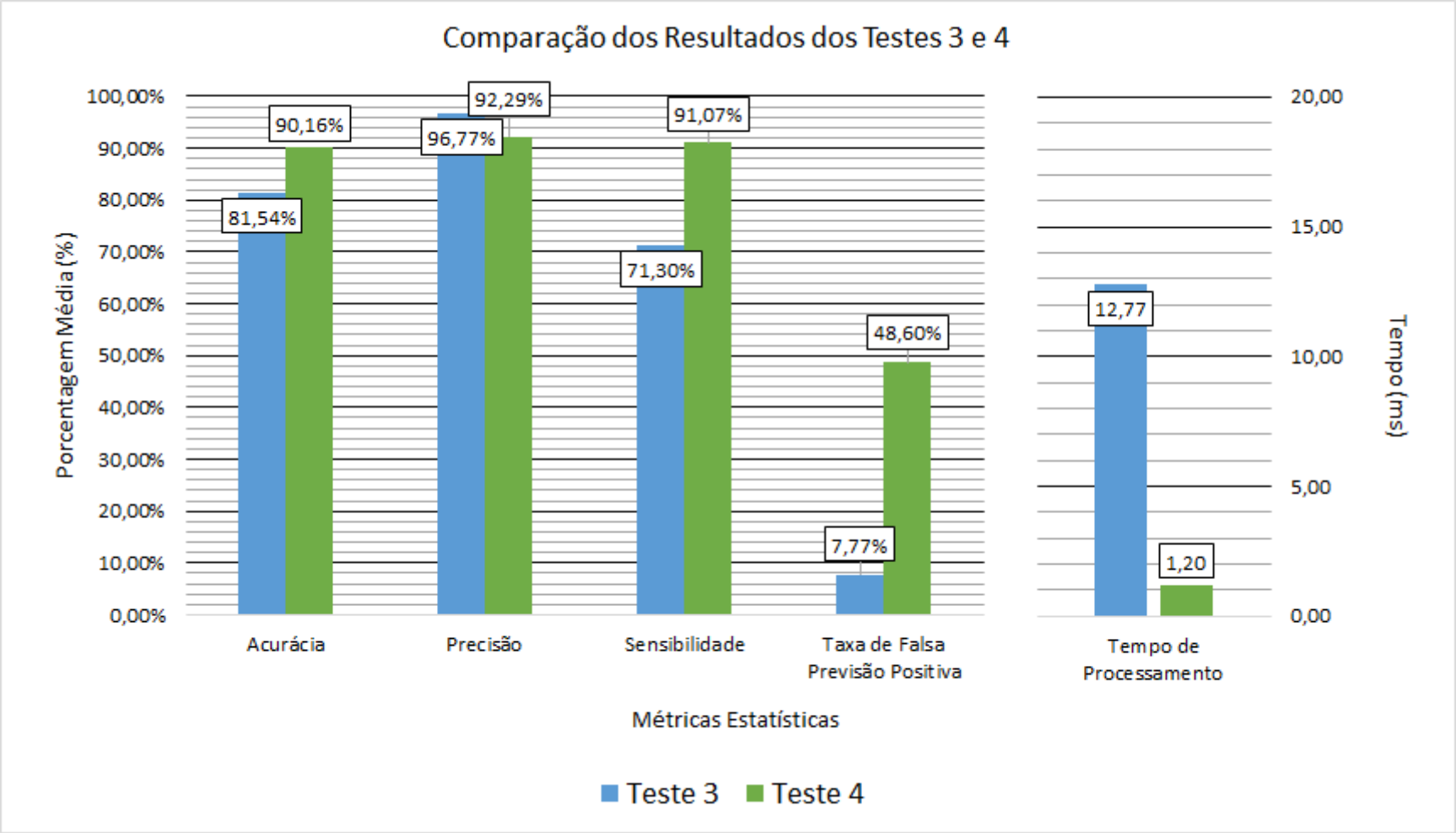
COMPARAÇÃO DOS TESTES 1 E 2



TESTES COM *OVERSAMPLING*

- Testes com o Conjunto de Dados balanceado com o método *Oversampling*:
  - Teste 3: Floresta Aleatória com balanceamento do tipo *Oversampling*;
  - Teste 4: Aumento de Gradiente com balanceamento do tipo *Oversampling*.

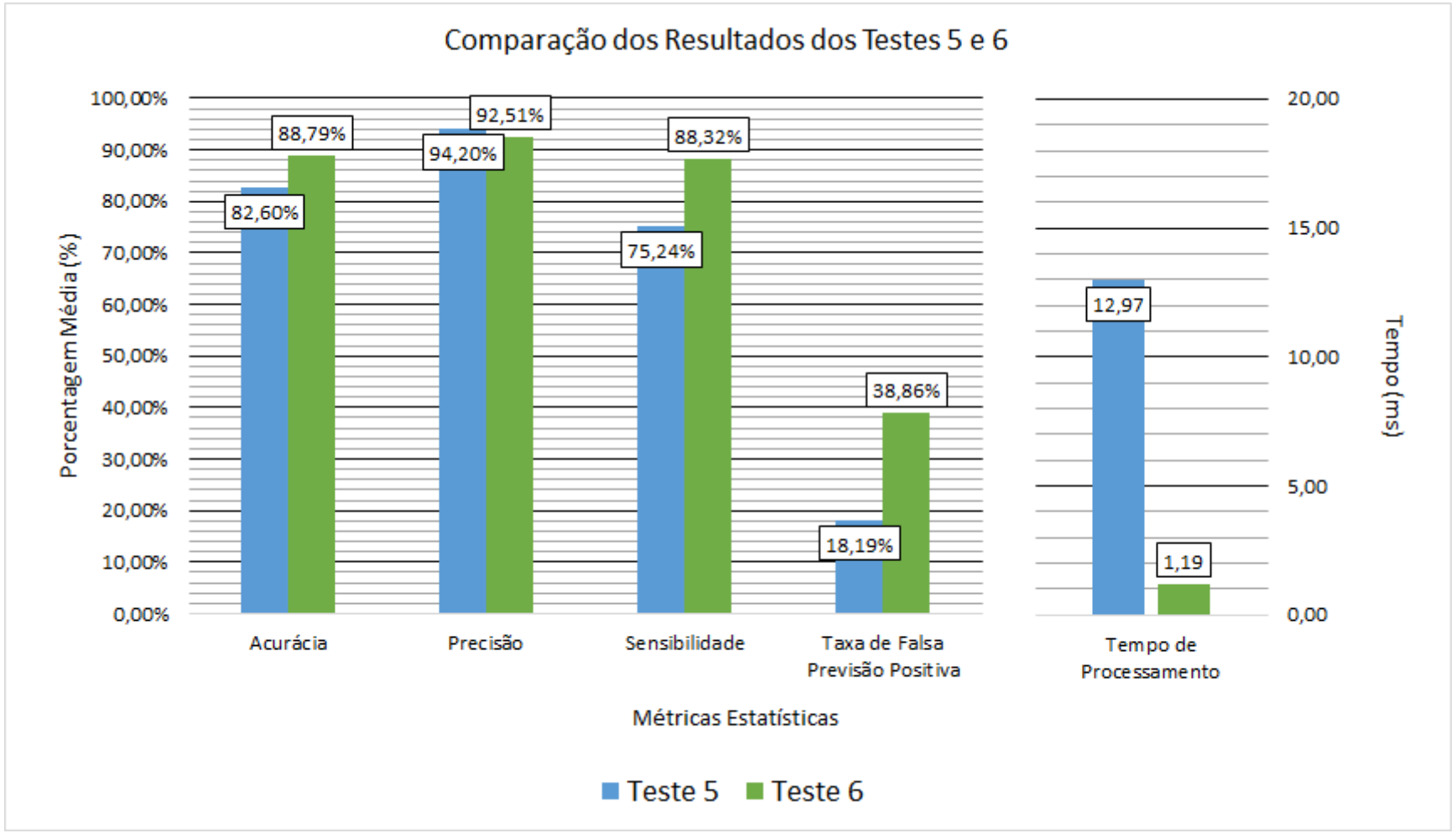
COMPARAÇÃO DOS TESTES 3 E 4



TESTES COM *UNDERSAMPLING*

- Testes com o Conjunto de Dados balanceado com o método *Undersampling*:
  - Teste 5: Floresta Aleatória com balanceamento do tipo *Undersampling*;
  - Teste 6: Aumento de Gradiente com balanceamento do tipo *Undersampling*.

COMPARAÇÃO DOS TESTES 5 E 6

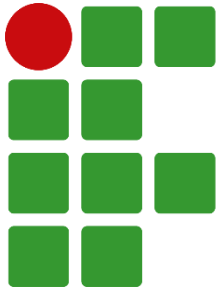


CONSIDERAÇÕES FINAIS

- Aumento de Gradiente ligeiramente melhor que o Floresta Aleatória:
  - Tempo de Processamento determinante;
  - Menor custo computacional;
  - Menor inclinação na previsão de dados.
- Algoritmos trabalham melhor com Conjunto de Dados sem balanceamento, se utilizar *Bootstrap.*

PESQUISAS FUTURAS

- Utilização do Aumento de Gradiente em pesquisas similares;
- Definir os atributos que mais descrevem uma evasão ou não:
  - *Least Absolute Shrinkage and Slection Operator (LASSO);*
  - *Recursive Feature Elimination (RFE).*
- Utilização de atributos socioeconômicos para a classificação;
- Amostras referentes a um período de matrícula;
- Utilizar dados mais atuais;
- Informar os possíveis evasores aos setores de apoio ao estudante para a realização de medidas preventivas.



INSTITUTO FEDERAL
Paraíba
Campus Campina Grande

COMPUTANDO *ENSEMBLE METHODS* PARA PREDIZER EVASÕES ESTUDANTIS

RODOLFO BOLCONTE DONATO

Orientadora: Samara Martins Nascimento

Coorientador: Gustavo Wagner Diniz Mendes