Projeto Condutor

Gabriel T. C. Hrysay[1], Raissa A. N. Higa[1],

[1]Departamento de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)

gabrielhrysay@alunos.utfpr.edu.br, rhiga.2018@alunos.utfpr.edu.br

Resumo. *Este artigo pretende propor soluções computacionais viáveis para as questões políticas, econômicas e sociais da cidade de Curitiba, Paraná, decorrente dos problemas gerados pela pandemia, fazendo uso das vantajosas técnicas de aprendizagem de máquina. Tendo em vista que nem sempre a melhor solução pode ser encontrada, com as limitações dos recursos computacionais, em tempo favorável, mesmo com uma possível evolução significativa nas velocidades de execução dos computadores, serão abordados problemas cujas soluções podem ser exploradas através de árvores de decisão e redes neurais.*

1. Introdução

1.1. Cenário da determinação de bandeiras epidemiológicas

O sistema de bandeiras é uma ferramenta importantíssima para a contenção do crescimento do número de casos que tem sido adotada por diversos municípios. As unidades administrativas selecionam atributos representativos das situações dos sistemas de saúde, das taxas de transmissão e gravidade dos casos registrados em dado período de tempo para determinar as restrições civis que serão legalmente impostas aos habitantes da região no próximo período. Deste modo, os órgãos governamentais buscam equilibrar as atividades socioeconômicas essenciais e não essenciais, de modo a não impedir o funcionamento urbano, ao mesmo tempo que tenta conter explosões de casos superiores à capacidade do sistema de saúde disponível.

A determinação da bandeira ocorre periodicamente e afeta atividades diversas. Desse modo, é de interesse dos cidadãos conhecer os motivos de uma determinada bandeira ser selecionada. Este problema pode ser abordado através de uma árvore de decisão. Um algoritmo como o ID3 permite elaborar uma árvore com base nos parâmetros relevantes aos municípios sem abrir mão da transparência no processo de decisão.

1.2. Cenário 2

Com o avanço mundial da Covid-19 e a circulação prolongada do vírus, novas variantes da doença vão surgindo através de mutações.A variante mais mencionada atualmente é a Delta, porém entre outras principais mais encontradas no Brasil estão a Alpha, a Beta e a Gamma. Cada variante possui características próprias que podem as deixar, por exemplo, mais letais ou mais transmissíveis.[Pinheiro 2021] Fazer o mapeamento e classificar essas variantes por área e período de tempo é útil para entender melhor o espalhamento delas, sendo possível assim estipular medidas para tentar conter seu avanço
.

2. Metodologia

2.1. Cenário da determinação de bandeiras epidemiológicas

A escolha por árvores de decisão garante a transparência aos cidadãos sobre as decisões tomadas pelo poder público, além de refletir fatores com alta probabilidade de impacto no agravamento da situação da crise de saúde causada pela doença

O sistema de bandeiras empregado na cidade de Curitiba, por exemplo, leva em conta nove atributos. De acordo com a prefeitura do município, eles são divididos em dois grupos, com 50% de peso para cada um deles.

O primeiro grupo tenta representar os níveis de propagação da doença a partir dos indicadores abaixo, retirados do site da prefeitura municipal de Curitiba. O segundo grupo, retirado da mesma fonte, tenta representar a capacidade de atendimento do município:

1. Indicadores de propagação, com peso total 5:
   (a) Indicadores com peso total 1,5:
      i. Número de casos novos confirmados nos últimos sete dias em relação ao número de casos novos confirmados nos sete dias anteriores.
      ii. Número de internados por SRAG (Síndrome Respiratória Aguda Grave) em UTIs no dia em relação ao mesmo número de sete dias atrás.
      iii. Número de pacientes de covid-19 confirmados em leitos de UTI no dia em relação ao mesmo número de sete dias atrás.
      iv. Número de pacientes de covid-19 confirmados em leitos clínicos no dia em relação ao mesmo número de sete dias atrás.
   (b) Indicadores com peso total 3,5:
      i. Número de casos confirmados nos últimos sete dias para cada 100.000 habitantes.
      ii. Número de óbitos nos últimos sete dias para cada 100.000 habitantes.
2. Indicadores de capacidade de atendimento, com peso total 5:
   (a) Número de leitos de UTI disponíveis para atender covid-19 no dia.
   (b) Número de leitos de UTI disponíveis para atender covid-19 no dia em relação ao mesmo número de sete dias atrás.
   (c) Número de leitos de enfermaria disponíveis para atender covid-19 no dia em relação ao mesmo número de sete dias atrás.

O cálculo final da bandeira é antecipado pelo cálculo de bandeiras parciais, determinada por intervalos específicos para cada atributo, como descrito na seção de apêndices.

O treinamento das árvores envolveu a obtenção de dados que garantissem o equilíbrio do conjunto de testes. Para tal, partiu-se da fórmula empregada pela prefeitura:

$$((A_{1ai} + A_{1aii} + A_{1aiii} + A_{1aiv}) * 0.375 + (A_{1bi} + \\ + A_{1bii}) * 1.75 + A_{2a} + 2 * A_{2b} + 2 * A_{2c})/10$$

Em que A_x representa o atributo de "índices"x. Intervalos específicos determinam a bandeira final da cidade a partir do resultado da equação acima:

Referência para a bandeira final	
Amarela	[0, 2]
Laranja	(2, 2.7]
Vermelha	(2.7, ∞)

2.1.1. Tratamento dos dados

Para o modelo de árvore de decisão, propõe-se utilizar os mesmos atributos que Curitiba como entrada para o algoritmo. A árvore realizará a classificação dos casos de entrada em três categorias de bandeira: amarela, laranja e vermelha. Alguns dados foram convertidos para inteiros 100 vezes maiores a fim de simplificar a produção dos dados.

Foi escrito um código que gerou todas as 3^9 combinações possíveis de "sub-bandeiras". A partir dos dados gerados, seleciona-se aleatoriamente a quantidade desejada de amostras de cada bandeira. Com os dados de bandeiras parciais já adquiridos, são gerados, para cada atributo (coluna) valores aleatórios que garantirão a obtenção da bandeira parcial naquele atributo. Isso é feito observando os intervalos que geram aquela bandeira naquele atributo, conforme tabela no apêndice.

Todos os atributos de entrada são do tipo numérico e contínuo. Deste modo, a árvore buscará o ponto de divisão dos valores tentando maximizar o ganho de informação. O algoritmo implementado pode obter o ganho de informação a partir da entropia ou da impureza de Gini. Para este artigo, o algoritmo foi executado com a última opção. O ponto negativo desta busca é que ela poderá encarecer o processo de treinamento da árvore[Russel 2013]. A maioria dos valores pode ser expressa de modo percentual. Aqueles que não puderam foram simulados dentro de um intervalo fictício, mas razoável.

Para a floresta aleatória, seleciona-se o percentual do conjunto de teste que será utilizado por cada uma das árvores. Também é possível definir a quantidade de árvores e a profundidade delas. Enfim, as árvores votam em uma bandeira, e a moda da votação é considerada a resposta final.

2.1.2. Requisitos

Requisito funcional:

1. O sistema deve devolver uma bandeira ao usuário a partir de um vetor numérico de entrada.

Requisito não funcional:

1. O sistema devera apresentar a árvore de decisão construída.
2. O sistema deverá ser capaz de construir uma floresta aleatória.

2.2. Cenário 2

Para este cenário de classificação das variantes do vírus pela área e época de contágio decidimos utilizar uma abordagem de aprendizado de máquina baseada em Redes Neurais Artificiais (RNA) em um modelo de aprendizado supervisionado utilizando backpropagation. O sistema recebe como entrada neurônios que correspondem à

localização x e y do local de contágio, e a semana em que ocorreu a infecção, e retorna na saída a probabilidade do caso infectado ser de cada variante. É utilizada um Multilayer Perceptron, onde cada camada intermediária de neurônios recebe o somatório das saídas dos neurônios das camadas anteriores multiplicadas por seus respectivos pesos seguida de uma função de ativação, (forward pass).

$$y = activationfunction(\Sigma xi * wi)$$

Em cada ciclo de aprendizagem, é calculado o erro de classificação , sendo assim, os pesos são ajustados retropropagando o erro através da rede neural, utilizando a regra da cadeia com o cálculo do gradiente, de acordo com a taxa de aprendizado especificada. Após isso, é aplicada a função softmax na camada de saída para retornar a probabilidade entre 0 e 1 do caso ser de cada variante.

2.2.1. Requisitos Funcionais

- RF1 - O sistema deve permitir cadastrar dados de ocorrências de variantes para o treinamento.
- RF2 - O sistema deve permitir retreinar a rede neural.
- RF3 - O sistema deve permitir ajustar o parâmetro da taxa de aprendizado.
- RF4 - O sistema deve permitir escolher as funções de ativação.
- RF5 - O sistema deve permitir ajustar por quantas épocas fazer o treinamento.

2.2.2. Requisitos Não-Funcionais

- RNF1 - O sistema deve utilizar os dados de treinamento para treinar a rede neural.
- RNF2 - O sistema deve prever de qual variante seria um novo caso baseado na localização e semana de contágio .

3. Resultados

3.1. Cenário 1

O algoritmo implementado é capaz de gerar dados para treinamento e para teste, de construir árvores de decisão, florestas de árvores de decisão e de testar as árvores e florestas geradas. Alguns resultados são apresentados em anexos, a saber: taxas de acerto por profundidade da árvore, para o caso de uma árvore única, e taxas de acerto por tamanho da floresta. Há diversos exemplos, considerando alterações em parâmetros. Em alguns casos, observa-se que com uma árvore é possível obter cerca de 80% de acerto enquanto sua profundidade é por volta de 6. Em outros casos, não é possível superar 75% de acerto. Já no caso de florestas, não foi difícil beirar 90% de acerto.

Espera-se obter na saída uma decisão da bandeira mais apropriada para a cidade em certo momento, a ser escolhida entre três categorias: amarela, laranja ou vermelha. Apesar da impossibilidade de se encontrar a menor árvore de decisão, a heurística do

algoritmo de aprendizado permite encontrar uma árvore pequena. Considerando que os dados de treinamento são baseados no algoritmo que elabora uma pontuação para cada avaliação já utilizado pelo município, é esperado que atributos com maior peso na nota no algoritmo original sejam apresentados na árvore como portadores do maior ganho de informação. Como trabalho futuro, é possível averiguar a funcionalidade do algoritmo a partir de dados reais.

3.2. Cenário 2

Para os testes foram gerados dados aleatórios artificiais com um fator bias, de forma que os casos de mesma variante fiquem mais agrupados. Para simplificar, os dados das localizações foram gerados no formato de um plano x * y, agrupados por semana em um intervalo de 1 a 8 semanas.

Realizando testes, a função de ativação do tipo ReLU, apresentou resultados melhores do que as do tipo Sigmoid e LeakyReLu. Com um learning rate de 0.02, observou-se a partir do valor do erro que um número adequado de iterações do treinamento seria por volta de 500 épocas, pois apresenta um módulo da derivada baixa, com o erro quase constante. Utilizar mais que isso não apresentaria classificações muito mais precisas e estaria sujeita à ocorrência de overfitting, quando o modelo não generaliza os dados, decorando cada característica usada para teste que o torna impreciso para nova entrada de dados reais.

4. Impactos sociais

4.1. Cenário 1

A representação visual da árvore de decisão permitiria identificar quais atributos costumam ser mais decisivos na tomada de decisão, uma vez que o algoritmo obtém essa informação a partir do cálculo de entropia. Essa visualização seria útil para cidadãos sem conhecimento matemático suficiente compreenderem como se dá o processo de escolha da bandeira. De modo similar, municípios pequenos que não contam com profissionais especializados na análise de dados ou na contenção de doenças infectocontagiosas podem ser favorecidos um método relativamente simples, mas funcional, para entender os riscos para suas populações.

Deste modo, a árvore poderia ter um fim didático, ajudando a população a compreender como as atitudes coletivas de distanciamento social podem influenciar na situação da saúde coletiva e das atividades econômicas em um futuro próximo.

4.2. Cenário 2

Cada nova variante do vírus da Covid-19 traz mutações cada vez mais perigosas e transmissíveis, além de poderem se tornar resistentes às vacinas criadas e possibilitarem a reinfecção de pessoas que já possuem anticorpos para outras variações. Sendo assim, mapear áreas de comum incidências das variantes permitem aos governos locais focar seus esforços e adotar medidas específicas para uma região.

Referências

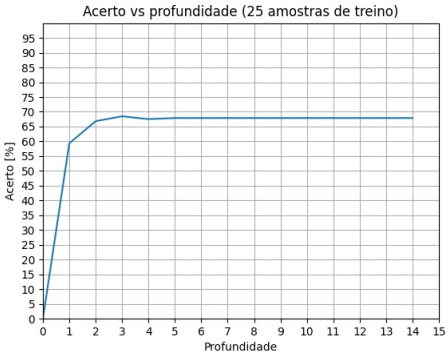
Pinheiro, C. (2021). Variantes do coronavírus quem são e como se comportam. https://saude.abril.com.br/medicina/variantes-do-coronavirus-quem-sao-e-como-se-comportam/.

Russel, S. Norvig, P. (2013). Inteligência artificial. 3 ed.

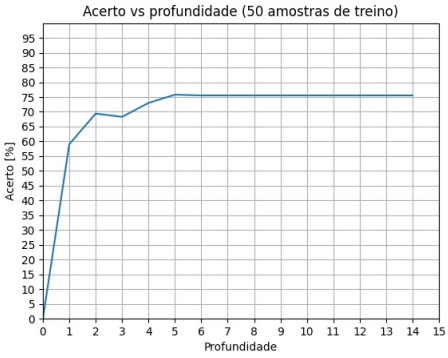
A. Valores de referência por atributo para determinação de bandeiras parciais

Referência para parâmetros do tipo 1(a)	
Amarela	[0, 1]
Laranja	(1, 2]
Vermelha	(2, ∞)
Referência para parâmetros do tipo 1(b)i	
Amarela	[0, 5]
Laranja	(5, 15]
Vermelha	(15, ∞)
Referência para parâmetros do tipo 1(b)ii	
Amarela	[0, 1]
Laranja	(1, 2.5]
Vermelha	(2.5, ∞)
Referência para parâmetros do tipo 2(a)	
Amarela	(55, ∞)
Laranja	(28, 55]
Vermelha	[0, 28)
Referência para parâmetros do tipo 2(b) e 2(c)	
Amarela	[0, 91]
Laranja	(91, 95]
Vermelha	[95, ∞)

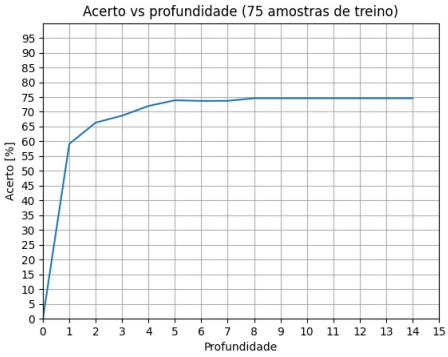
B. Plotagens de taxa de acerto por profundidade da árvore para o caso de árvore única



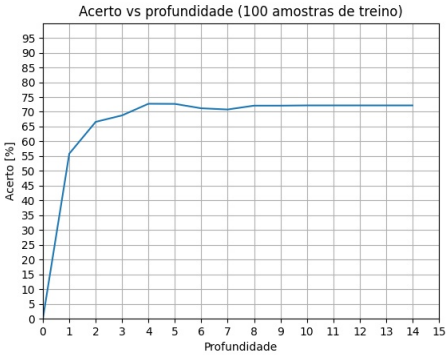
(a) Figura 1: treino com 25 amostras por bandeira



(b) Figura 2: treino com 50 amostras por bandeira

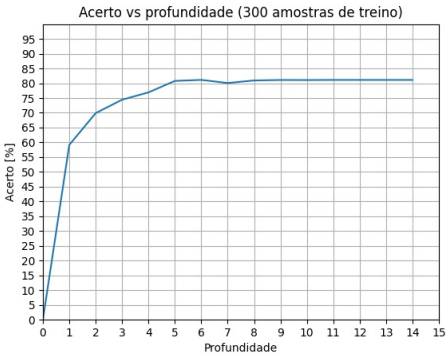


(c) Figura 3: treino com 75 amostras por bandeira

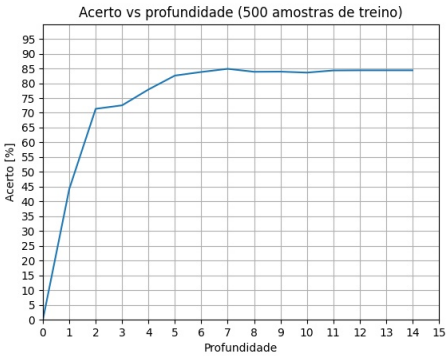


(d) Figura 4: treino com 100 amostras por bandeira

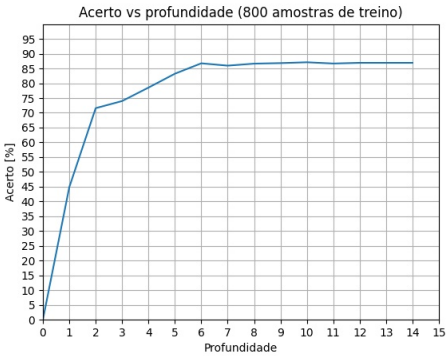
Figura 1. Plotagens de taxa de acerto por profundidade da árvore para o caso de árvore única, considerando o tamanho do conjunto de treinamento da árvore (parte 1)



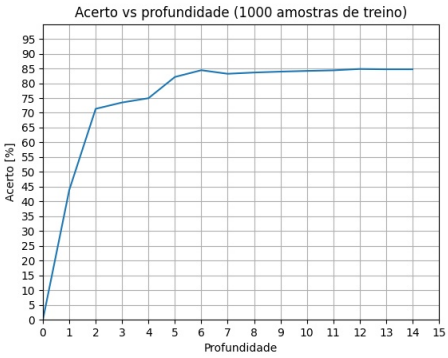
(a) Figura 1: treino com 300 amostras por bandeira



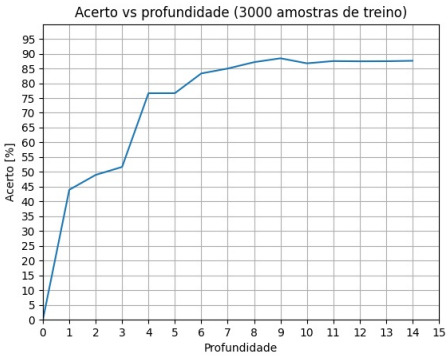
(b) Figura 2: treino com 500 amostras por bandeira



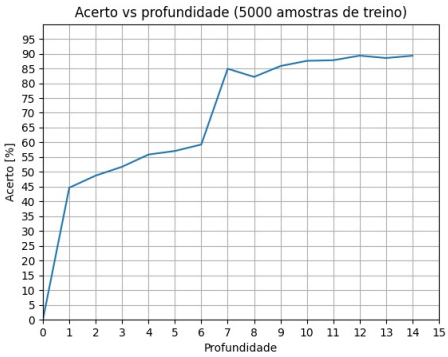
(c) Figura 3: treino com 800 amostras por bandeira



(d) Figura 4: treino com 1000 amostras por bandeira



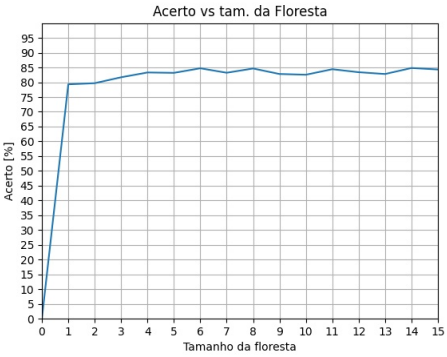
(e) Figura 5: treino com 3000 amostras por bandeira



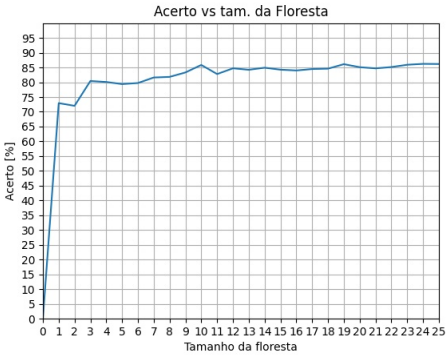
(f) Figura 6: treino com 5000 amostras por bandeira

Figura 2. Plotagens de taxa de acerto por profundidade da árvore para o caso de árvore única, considerando o tamanho do conjunto de treinamento da árvore (parte 2)

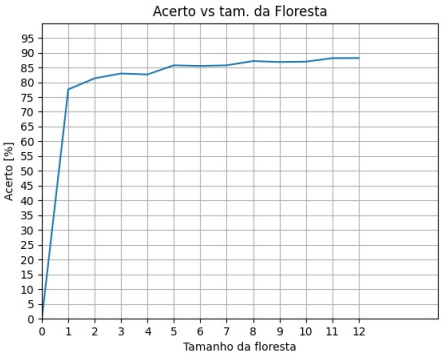
C. Taxas de acerto por tamanho da floresta



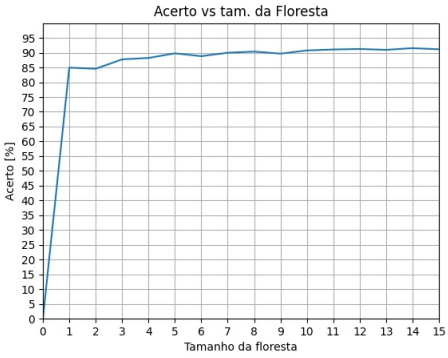
(a) Árvores com profundidade D = 4; subconjunto de treinamento de cada árvore é 75% do conjunto de treinamento completo



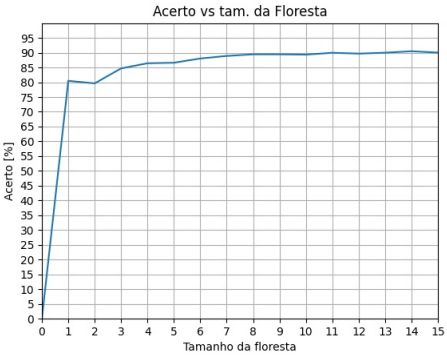
(b) Árvores com profundidade D = 4; subconjunto de treinamento de cada árvore é 20% do conjunto de treinamento completo



(c) Árvores com profundidade D = 6; subconjunto de treinamento de cada árvore é 40% do conjunto de treinamento completo



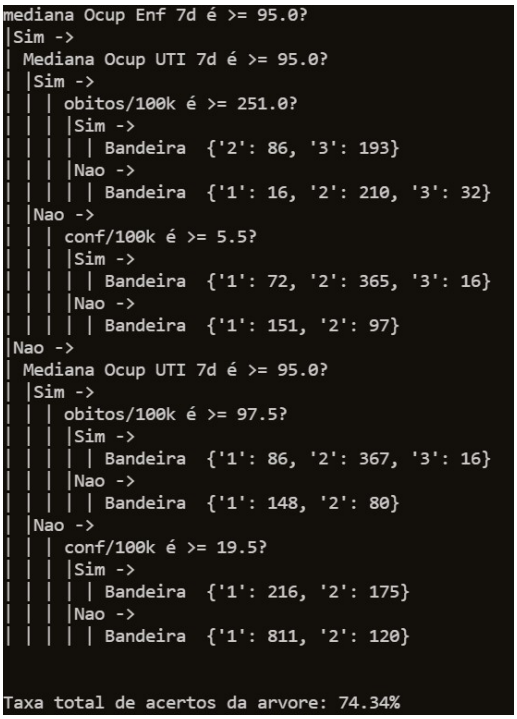
(d) Árvores com profundidade D = 6; subconjunto de treinamento de cada árvore é 75% do conjunto de treinamento completo



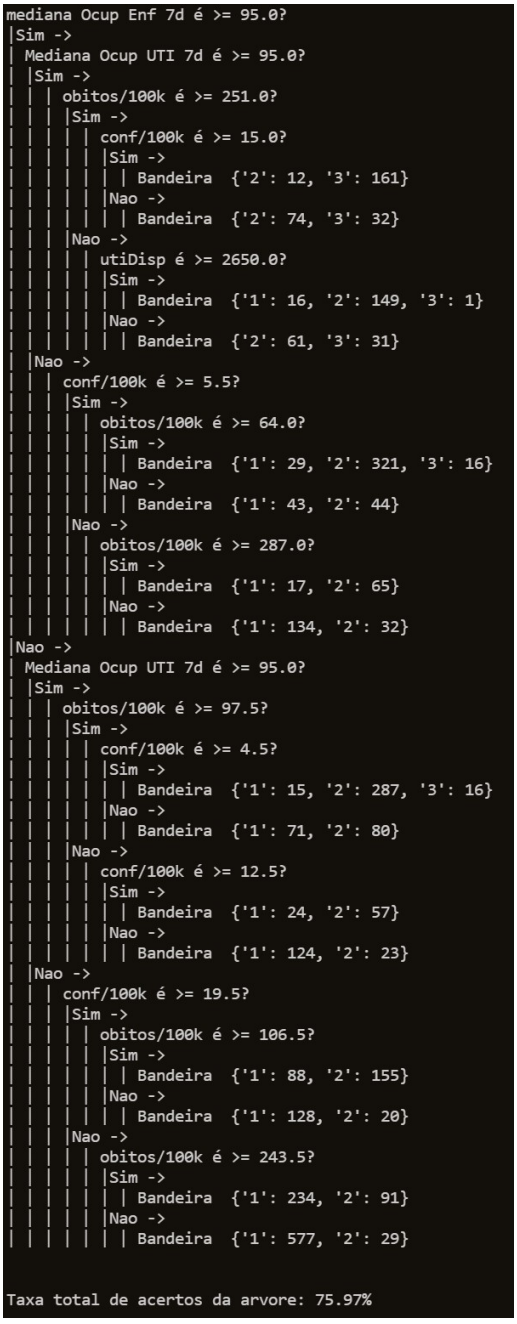
(e) Árvores com profundidade D = 10; subconjunto de treinamento de cada árvore é 60% do conjunto de treinamento completo

Figura 3. Plotagens de taxas de acerto por tamanho da floresta, considerando o tamanho do conjunto de treinamento das árvores da floresta e a profundidade das árvore

D. Exemplos de visualização das árvores geradas

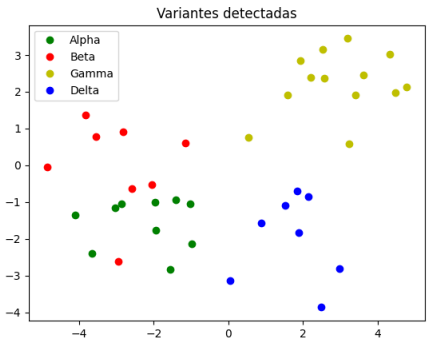


(a) Árvore com profundidade D = 4

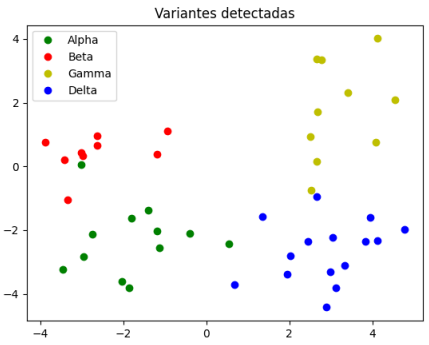


(b) Árvore com profundidade D = 5

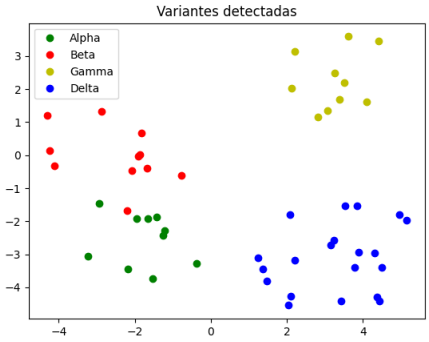
E. Dados Gerados das Variantes por Semana



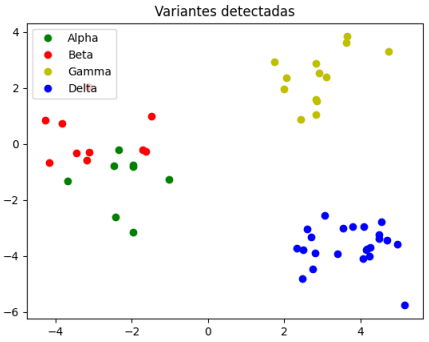
(a) Semana 1



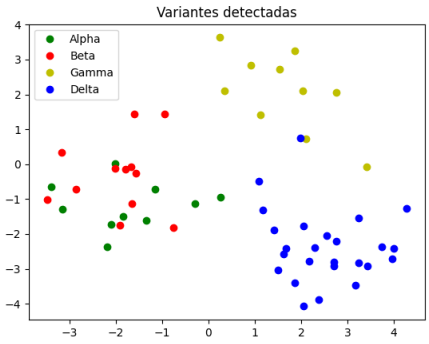
(b) Semana 2



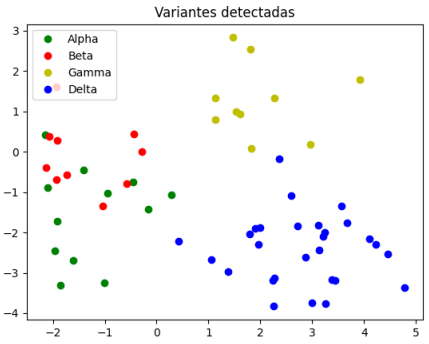
(c) Semana 3



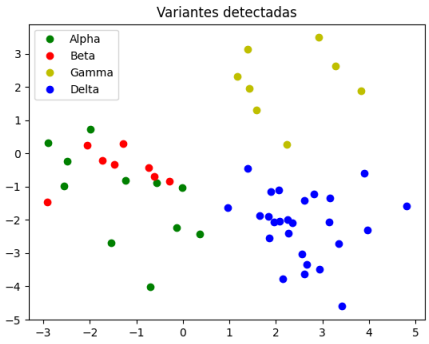
(d) Semana 4



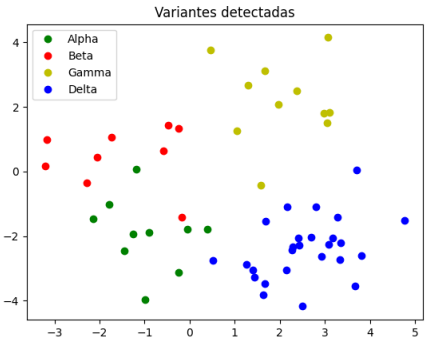
(e) Semana 5



(f) Semana 6



(g) Semana 7



(h) Semana 8

Figura 5. Avanço das variantes ao decorrer das semanas

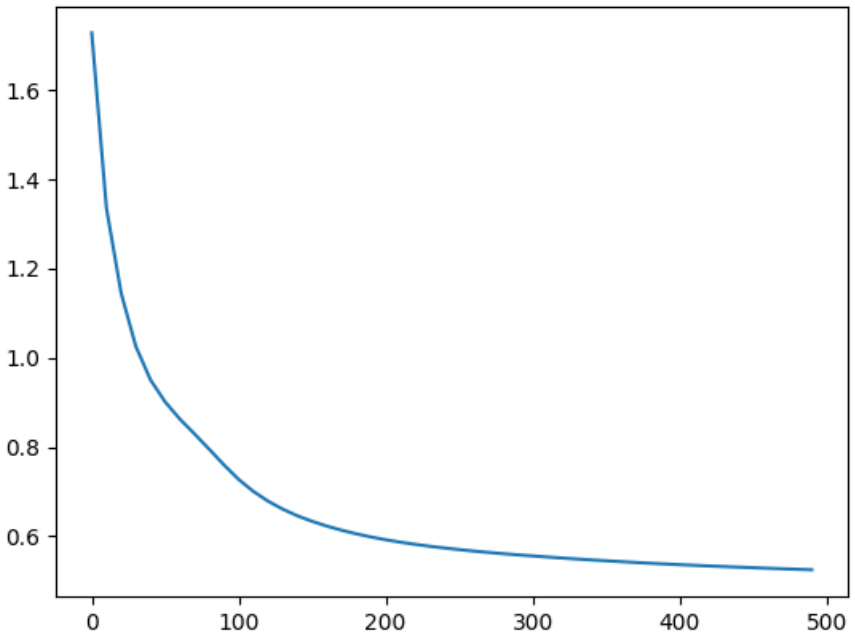


Figura 6. Erro x Número de Épocas

F. Erro por época