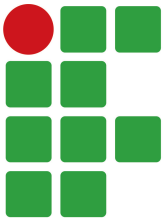

INSTITUTO FEDERAL
Goiás
Câmpus Anápolis

Pedro Henrique Silva Rodrigues

SELEÇÃO DE ÁRVORES EM ALGORITMOS DE FLORESTAS ALEATÓRIAS

Anápolis-GO

Dezembro - 2018

Pedro Henrique Silva Rodrigues

SELEÇÃO DE ÁRVORES EM ALGORITMOS DE FLORESTAS ALEATÓRIAS

Relatório Científico apresentado ao curso de Bacharelado em Ciências da Computação, como exigência da Disciplina de Inteligência Artificial ao prof^{o} D^{r} $Daniel Xavier Sousa$

Instituto Federal de Goiás - Campus Anápolis

Anápolis-GO

Dezembro - 2018

Lista de ilustrações

Sumário

Resumo

Florestas Aleatórias é um método de Aprendizado de Máquina para predição com base no aprendizado de instâncias de treino. A partir desse treino é gerado um conjunto de Árvores de Decisão, o voto majoritário define a classe resultante de uma predição, no caso de classificação, ou a média dos valores resultantes define o valor final, para o casa de regressão. Entretanto o custo computacional para se construir uma árvore e treiná-la é elevado assim como o custo para se fazer uma predição com base no modelo gerado por esse método. Um outro problema relacionado ao tamanho do conjunto de Árvores é a acurácia da predição, um aumento excessivo da quantidade de árvores pode tornar o método enviesado, ou seja cair em overfitting, dessa forma é importante estudar e desenvolver heurísticas/métodos para escolha das árvores que serão utilizadas na predição. Entre os conceitos que podem ser levados em consideração para a criação dessas heurísticas estão a correlação entre as features dos dados, profundidade da árvore, altura de uma feature na árvore, entre outros.

Introdução

Modelos preditivos são um dos principais ramos de pesquisa modernos, visto que essas técnicas de Análise de Dados revolucionaram a indústria com o desenvolvimento de produtos de alta tecnologia e devido ao seu alto potencial de negócio. Atualmente muitas dessas têm sido utilizados por grandes empresas como Google, Amazon e Netflix, de forma a otimizar os gastos, encontrar padrões em compras, fazer recomendações aos seus usuários/clientes com o intuito de se ter maior conhecimento e dessa forma manter-se forte no mercado (DEB; JAIN; DEB, 2018).

Embora os métodos tradicionais de Aprendizado de Máquina obtenham ótimos valores de acurácia durante a predição, ainda há erros inerentes ao método, e sempre é possível cair no overfitting (SHALEV-SHWARTZ; BEN-DAVID, 2014). Dessa forma devem ser adotados estratégias para otimizar a taxa de acerto baseado nas características intrínsecas desses métodos.

As Florestas Aleatórias, que é um método de Aprendizado de Máquina que cria como modelo um conjunto de Árvores de decisão (RUSSELL; NORVIG, 2016), possuem muitos parâmetros que podem ser ajustados de acordo com heurísticas de forma a aumentar a acurácia da predição. Entre esses parâmetros estão profundidade da árvore, quantidade de features por árvore, quantidade de instâncias de treino por árvore, entre outros. Entretanto há pouca informação acerca da seleção de Árvores após o modelo ter sido gerado baseando-se em suas características.

Esse estudo tem por objetivo realizar uma avaliação dos Algoritmo de Florestas Aleatórias considerando Heurísticas na seleção das árvores de decisão criadas pelo modelo. Espera-se com essas heurísticas obter redução da quantidade de Árvores para realizar uma predição, entretanto sem reduzir a taxa de acerto do método, ou seja pretende-se reduzir o custo computacional na predição mantendo a acurácia.

1 Referencial Teórico

1.1 Inteligência Artificial

Inteligência Artificial é a área de estudo que tem por objetivo representar entidades do mundo real através de modelos computacionais. Sua aplicabilidade é ampla nos mais diversos ramos com Engenharia, Biologia, Estudos Sociais, e principalmente Ciência da Computação, abrangendo subcampos de aprendizagem, simulação, percepção entre outros, até problemas específicos, como a solução de um jogo, desenvolvimento de equações matemáticas e diagnósticos médicos (RUSSELL; NORVIG, 2016).

O progresso recente da Inteligência Artificial se tornou viável a medida que sistemas reais estão implantando-a, como é o caso da Google, Amazon, Netflix entre outras grandes empresas. Os subcampos de Inteligência Artificial se tornaram mais integrados, e a Inteligência Artificial tem encontrado uma área de concordância com as demais disciplinas visto sua aplicabilidade.

1.1.1 Machine Learning

Machine Learning é um subcampo de Inteligência Artificial que dá a um computador a habilidade de aprender sem ser explicitamente programado para isso (SAMUEL, 1959). Ou seja baseado um conjunto de dados, o computador detecta padrões, encontra tendências, faz associações e a partir disso gera um modelo. Baseado nesse modelo o método é capaz de realizar predições.

Diferente do conhecimento humano, esses métodos necessitam de grandes quantidades de exemplos para que possam ter taxas de acerto elevadas. A quantidade de dados é diretamente proporcional ao custo computacional para analisá-los, e consequentemente o modelo gerado se torna mais complexo (RUSSELL; NORVIG, 2016).

1.2 Random Forest

Nesse algoritmo são geradas um conjunto (ensemble) de árvores sejam elas de decisão ou regressão conforme necessidade do problema. A predição se baseia nas avaliações individuais de cada árvore, com base na média, maior ocorrência e pesos das árvores (Figura 1). Os hiper-parâmetros desse método são: profundidade das árvores, tamanho da floresta, quantidade de features por árvores (BREIMAN, 2001).

Os modelos de árvore de decisão ou regressão consistem em fazer subdivisões nos conjuntos de dados com base em características semelhantes. Os grupos gerados por

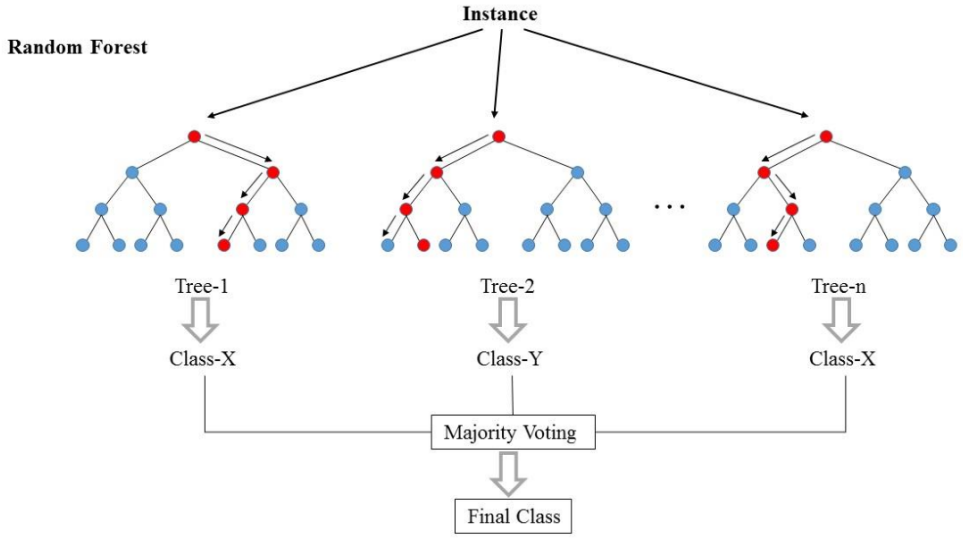

Figura 1 – Esquematização Random Forest (DIMITRIADIS et al., 2018)

essas subdivisões possuem informações em comum e possivelmente pertencem a uma mesma classe (classificação), ou possuem um variação linear semelhante (regressão). Essas subdivisões acontecem nos nós da árvores com base em alguma métrica que tenha por objetivo maximizar a divisão desses grupos da forma mais simples.

Quanto maior a quantidade de nós e profundidade da árvore mais o árvore gerada é específica para o conjunto de dados de treino, dessa forma não foi capaz de generalizar as classes e caiu em overfitting. Uma forma de evitar esse problema é limitar o tamanho e profundidades de uma árvore.

2 Heurísticas Propostas

Foram implementadas e testadas três hipóteses de heurísticas de seleção de Árvores para os algoritmos de Florestas Aleatórias. A primeira se baseia no quão alto uma feature é pra uma instância. A segunda e a terceira se baseiam na correlação entre as features de uma árvore, uma comparando-as entre si, e a outra comparando-as com o valor de predição.

2.1 1ª Hipótese de Heurística: Seleção durante o teste

A primeira Hipótese de Heurística consiste em realizar uma seleção das árvores que serão utilizadas durante a predição, com base nas features que possuem maiores valores para a instância em questão.

O procedimento para isso consiste em: Aplicar escala no conjunto de dados, para que as features pudessem ser analisadas de forma equivalente precisa-se adicionar uma escala. A Escala Padrão pela média consiste em redimensionar os dados de forma que a média seja 0 e os demais valores fiquem em torno desse conforme figura 2. Uma distribuição de valores que se encontram próximos de uma média (chamada de "valor esperado") é conhecida como uma distribuição "normal" (REIS et al., 1999).

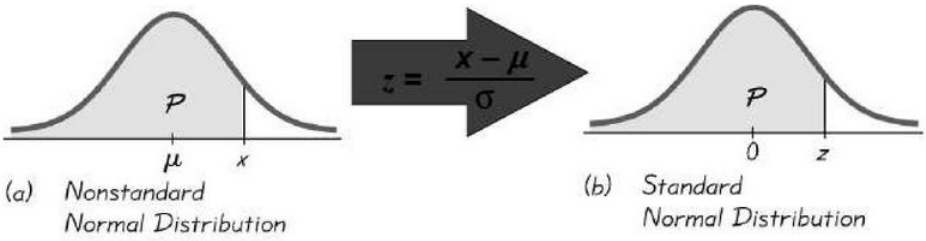

Figura 2 – Exemplo de Normalização pela Média (REIS et al., 1999).

Após aplicar a escala no conjunto de dados é feita a seleção de Árvores que serão utilizadas para a predição. As Árvores que possuem features cujos valores nas instâncias estão mais próximos de 0 e que estão próximas ao topo são pouco representativas dessa instância, portanto as features mais representativas (ou seja com valores mais distantes do 0, média nesse caso) estarão mais abaixo na árvore e serão por isso mais impactantes na classificação da instância em um conjunto.

Essa classificação pode ser ordenada conforme a profundidade desses nós na árvore, e após isso definindo um hiper-parâmetro como a porcentagem de árvores que serão

utilizadas é possível fazer uma predição com as árvores das primeiras posições.

2.2 2ª Hipótese de Heurística: Seleção com base na correlação das features da Árvore

A segunda Hipótese de Heurística consiste em realizar uma seleção das árvores que serão utilizadas durante a geração do modelo, com base na correlação entre cada par de features do conjunto de treino que será aplicado.

Correlação pode ser entendido como o grau de dependência entre duas variáveis (REIS et al., 1999). É possível ver uma interpretação gráfica da correlação entre duas variáveis na figura 3.

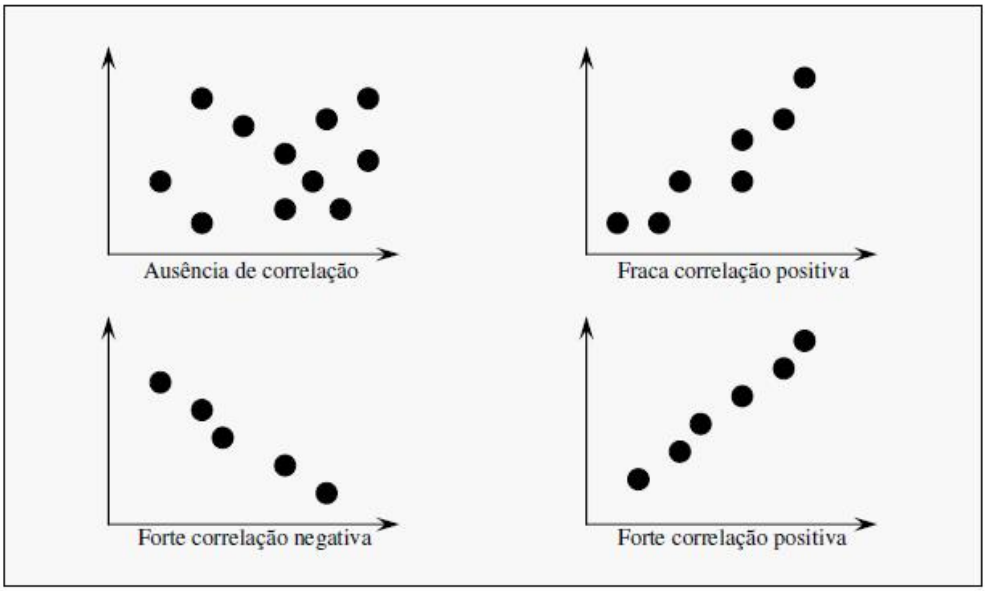

Figura 3 – Interpretação Gráfica de Correlação (REIS et al., 1999).

Essa hipótese consiste em medir a correlação de cada par de features do conjunto de treino e a partir disso realizar uma seleção das Árvores. A ideia consiste em avaliar as árvores cujas features utilizadas para gerá-las tem alta ou baixa correlação entre si.

Nesse método é feito o somatório do módulo da correlação de cada par de features para cada Árvore. Feito isso é possível ordenar as Árvores com base nesses valores de forma crescente ou decrescente. Então com base no hiper-parâmetro porcentagem de árvores seleciona-se as primeiras dessa lista para a predição.

Essa hipótese tem como diferencial, em relação a anterior,a seleção das árvores como parte do modelo.

2.3 3ª Hipótese de Heurística: com base na correlação das features com o valor predito

Essa hipótese é semelhante a anterior, entretanto a diferença consiste em analisar a correlação das features da árvore com a coluna de valores de classificação ou regressão, ou de forma mais explicativa o Y do conjunto de dados.

Análogo a anterior é feito o somatório do módulo da correlação das features utilizadas na Árvore com o Y, em seguida é realizada uma ordenação no vetor de forma crescente ou decrescente. Então com base no hiper-parâmetro porcentagem de árvores seleciona-se as primeiras dessa lista para a predição.

3 Metodologia Experimental

As Hipóteses de Heurísticas foram testadas com dois Datasets de classificação e dois de regressão, disponibilizados pela biblioteca de função Scikit-Learn, através das funções: load_digits e load_diabetes para Classificação e load_breast_cancer e load_boston para Regressão.

As medidas de Acurácia utilizadas foram F1 (Scikit-Learn), que calcula a média harmônica entre Precision e Recall para os testes com classificação, e RMSE - Root Mean Square Error (também do Scikit-Learn) para as coleções de Regressão.

A implementação de Floresta Aleatória utilizada como macro modelo foi a implementação do Scikit-Learn, e para fins de Análise as Heurísticas foram comparadas com essa implementação.

A quantidade de Árvores geradas como teste foi 100, 200, 300, 500, 700 e 1000 e a porcentagem de Árvores utilizadas, que é hiper-parâmetros das três funções foi 25%, 50% e 75%. Uma vez que 100 % consiste em utilizar todas as Árvores do modelo não se avaliou essa opção, uma vez que teria o mesmo resultado da implementação do Scikit-Learn.

4 Resultados e Análise

Os resultados dos testes realizados podem ser visualizados nos gráficos a seguir, é notável que os métodos obtiveram resultados pouco relacionados:

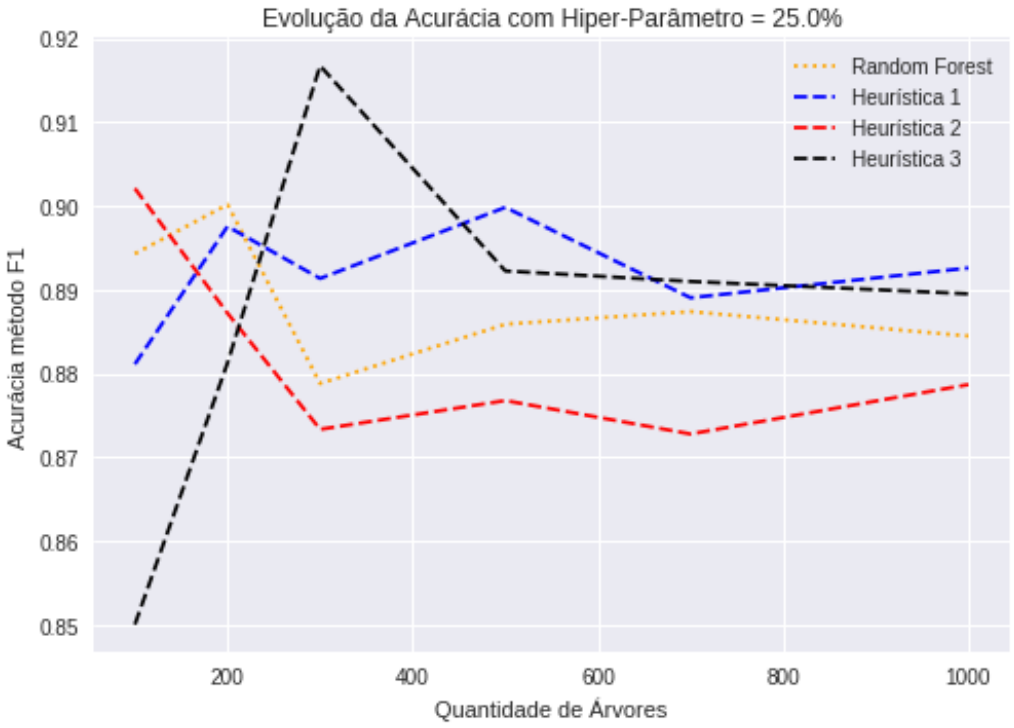

Figura 4 – Evolução do F1 para as Heurísticas com Coleção: Dígitos, e hiper-parâmetro 25%

4.1 Avaliação Geral

Conforme se vê nos gráficos embora as Hipóteses de Heurísticas obtenham bons resultados em alguns testes, esse comportamento é completamente caótico, quando se compara os resultados de uma forma geral. Entretanto algumas afirmações são possíveis: No geral as propostas de Heurísticas obtiveram resultados piores quando a quantidade de árvores é pequena (100), entretanto conforme a quantidade de Árvore cresce essa predição tende a se igualar em relação aos três métodos. Isso ocorre visto que uma grande quantidade de Árvores torna-as repetitivas e portanto a diferença dos resultados dos métodos é pequena. Essa efeito caótico pode estar relacionado às características aleatórias intrínsecas do método Florestas Aleatórias.

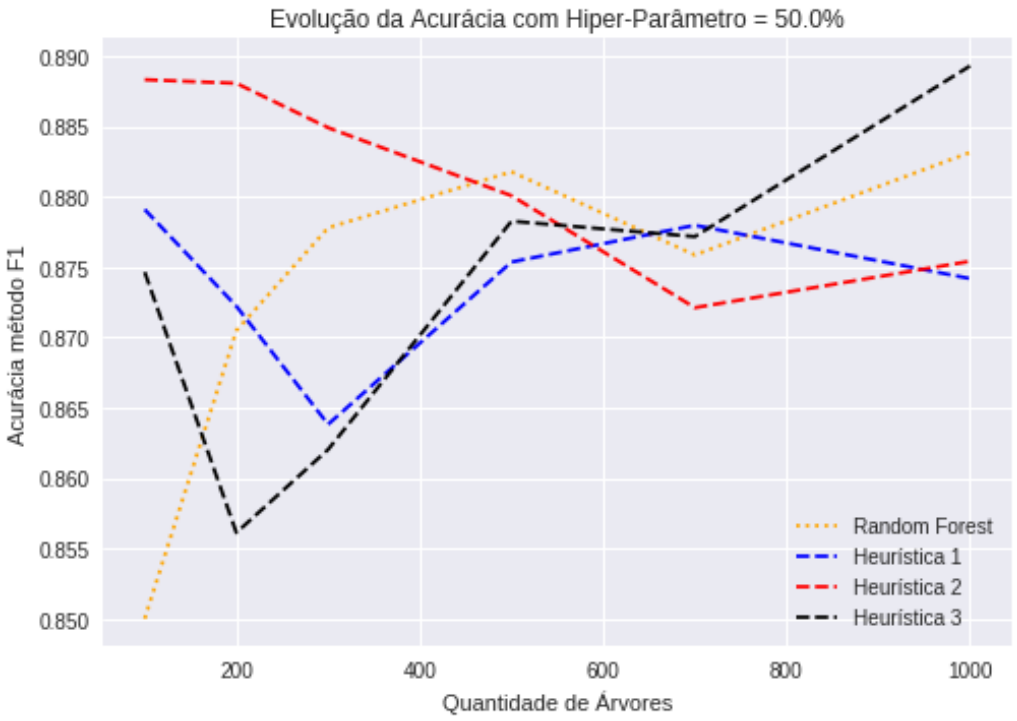

Figura 5 – Evolução do F1 para as Heurísticas com Coleção: Dígitos, e hiper-parâmetro 50%

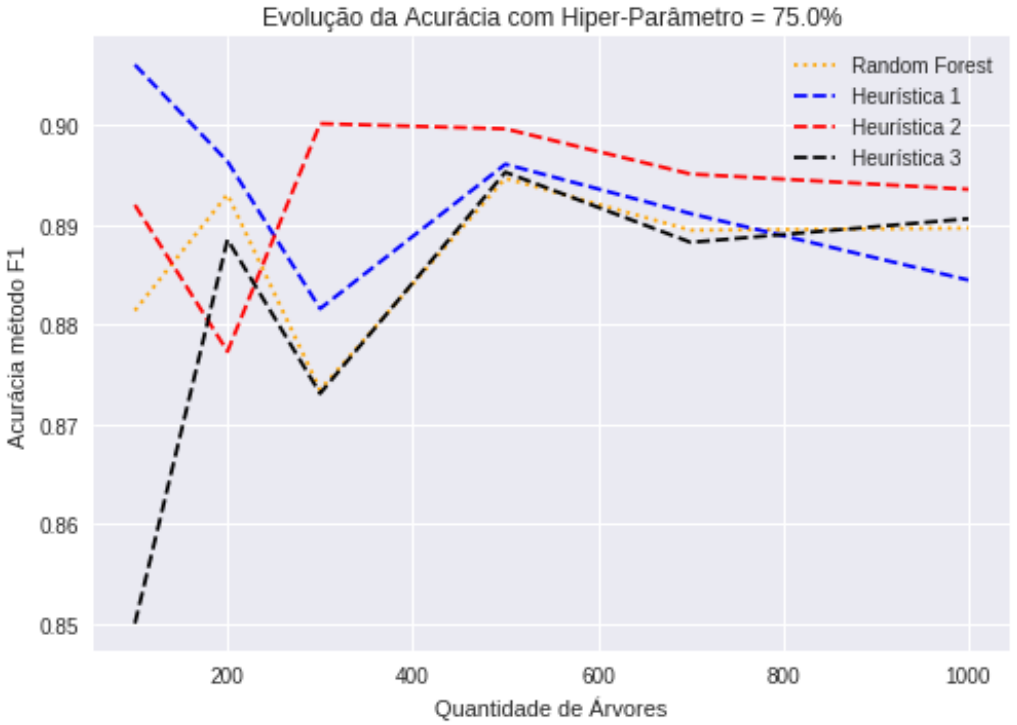

Figura 6 – Evolução do F1 para as Heurísticas com Coleção: Dígitos, e hiper-parâmetro 75%

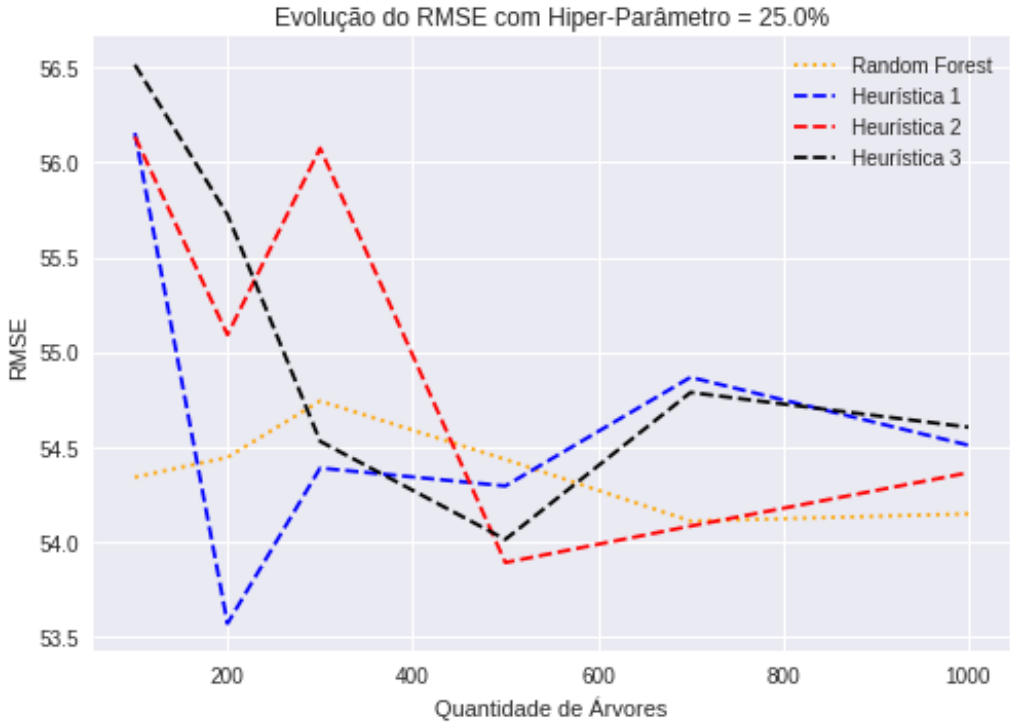

Figura 7 – Evolução do F1 para as Heurísticas com Coleção: Diabetes, e hiper-parâmetro 25%

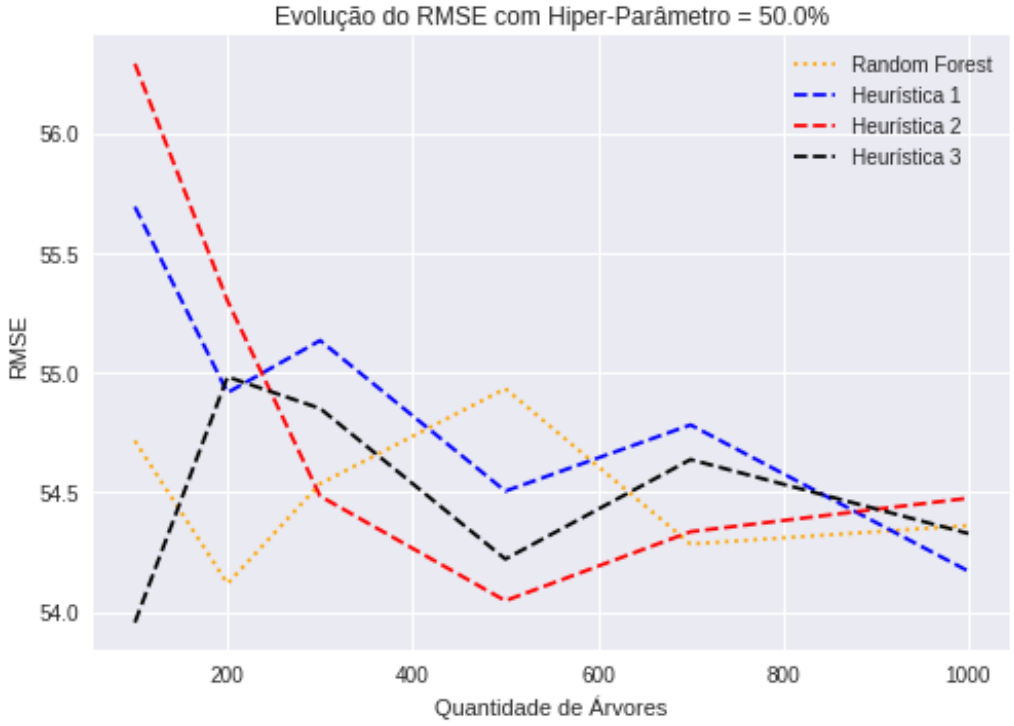

Figura 8 – Evolução do F1 para as Heurísticas com Coleção: Diabetes, e hiper-parâmetro 50%

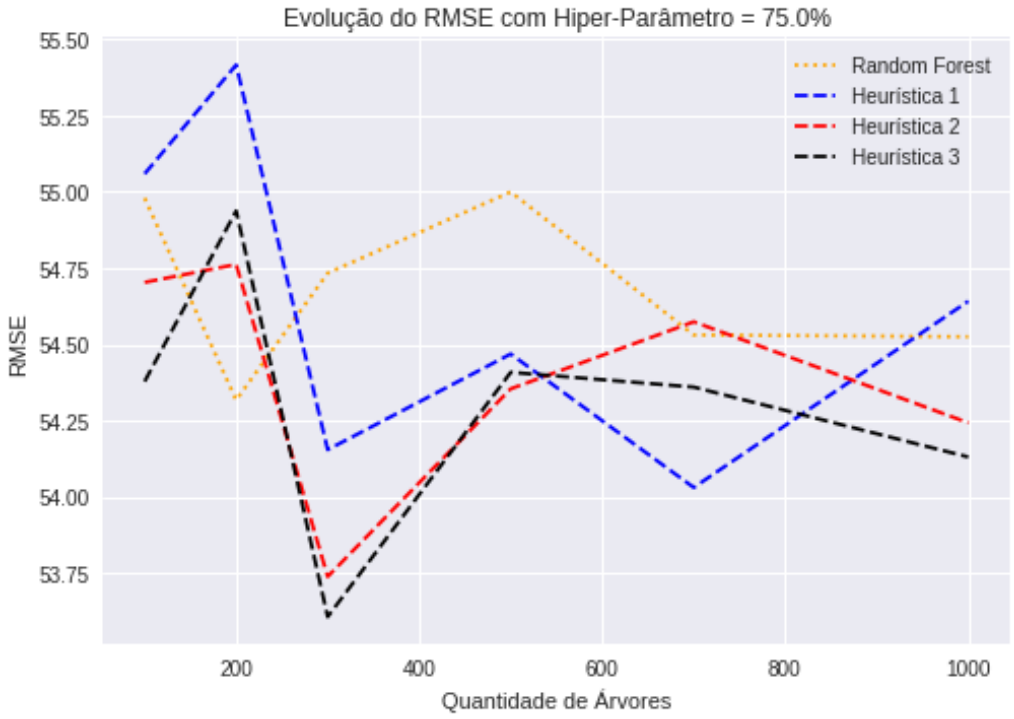

Figura 9 – Evolução do F1 para as Heurísticas com Coleção: Diabetes, e hiper-parâmetro 75%

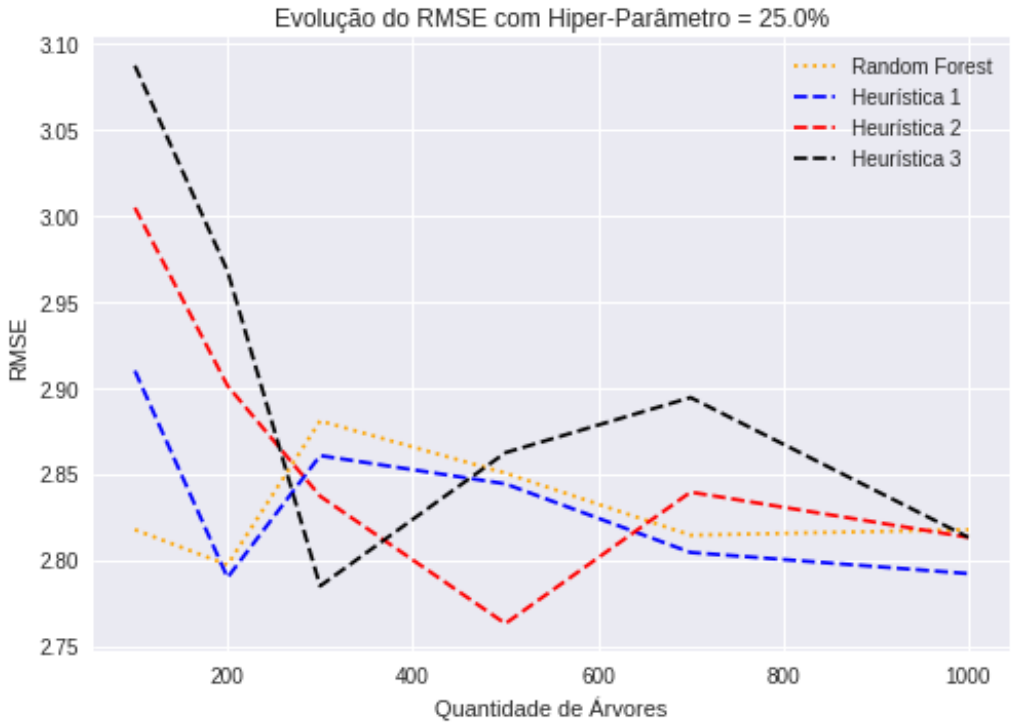

Figura 10 – Evolução do RMSE para as Heurísticas com Coleção: Boston, e hiper-parâmetro 25%

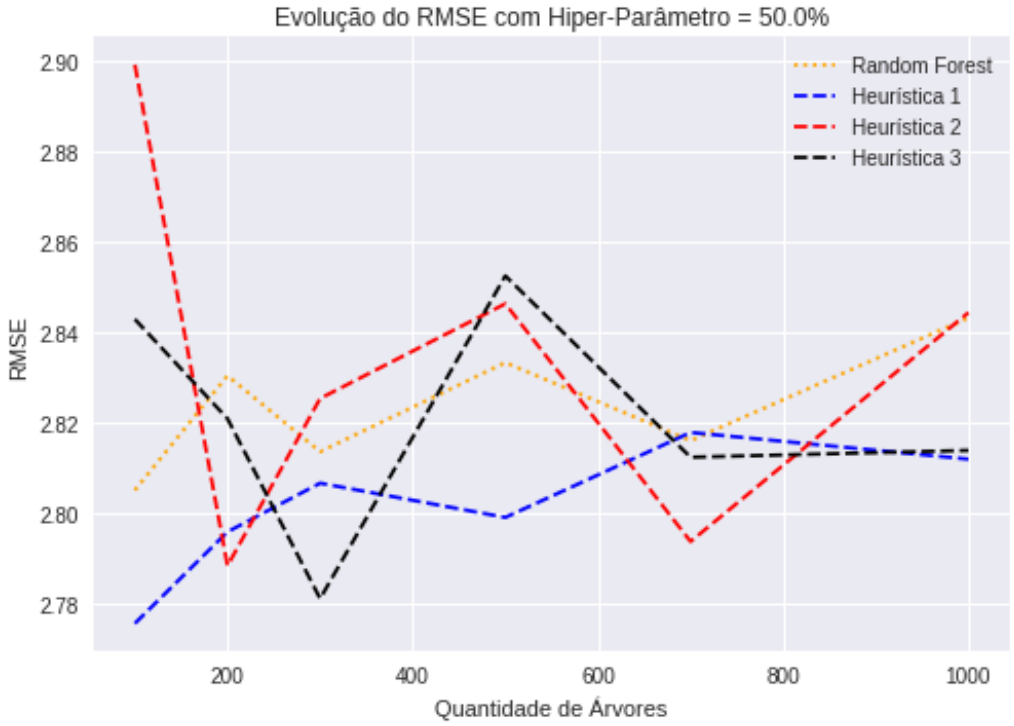

Figura 11 – Evolução do RMSE para as Heurísticas com Coleção: Boston, e hiper-parâmetro 50%

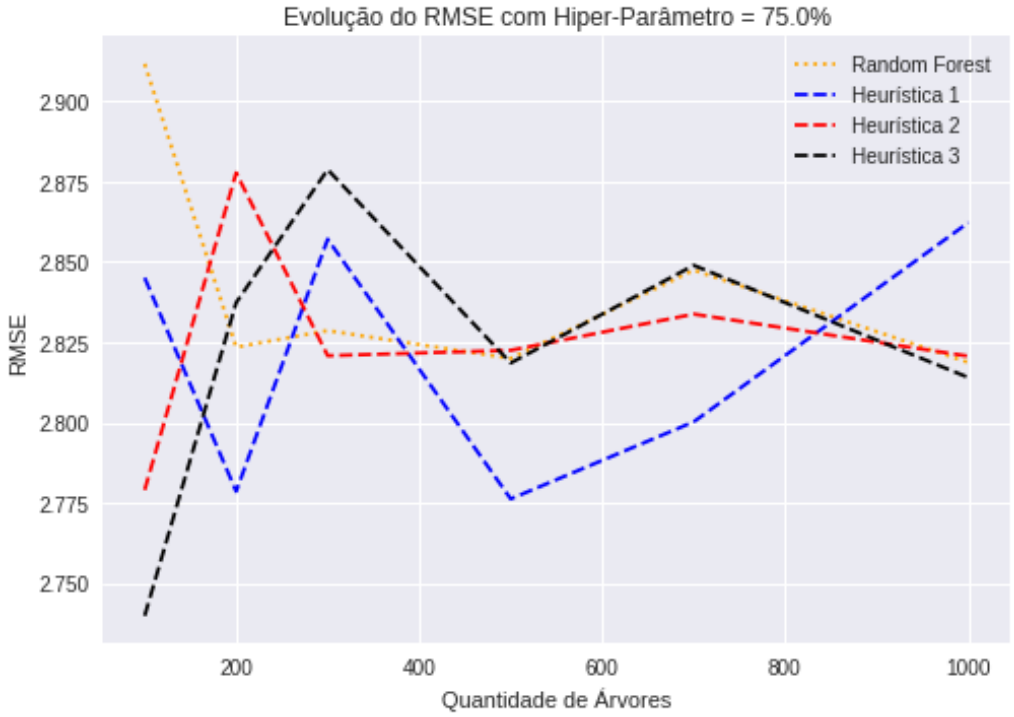

Figura 12 – Evolução do RMSE para as Heurísticas com Coleção: Boston, e hiper-parâmetro 75%

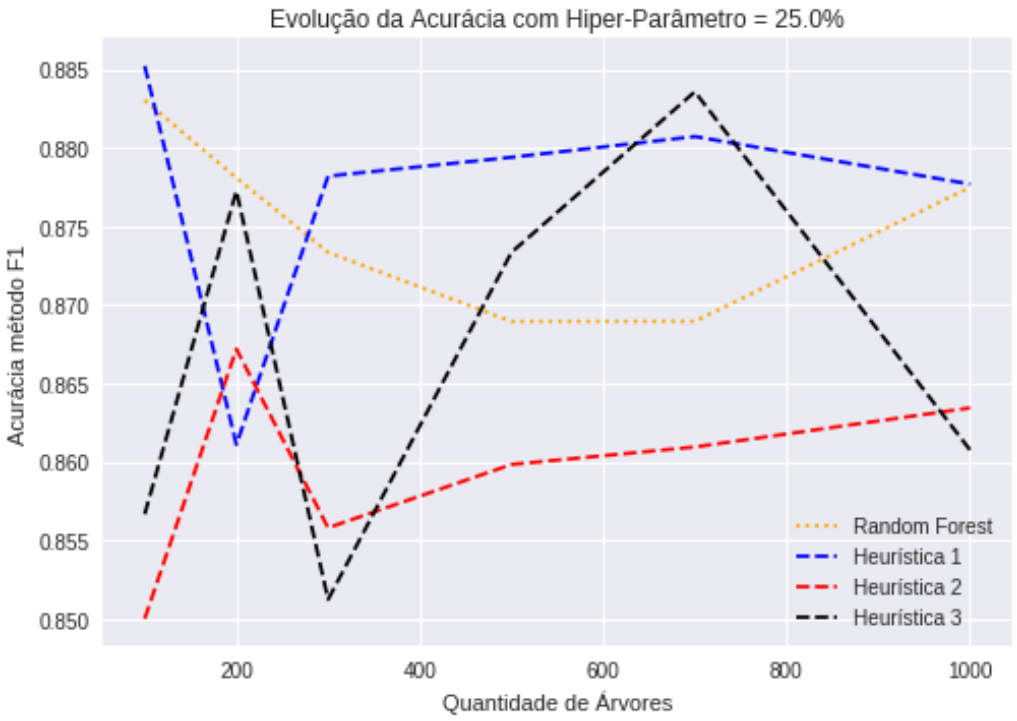

Figura 13 – Evolução do RMSE para as Heurísticas com Coleção: Breast-Cancer, e hiper-parâmetro 25%

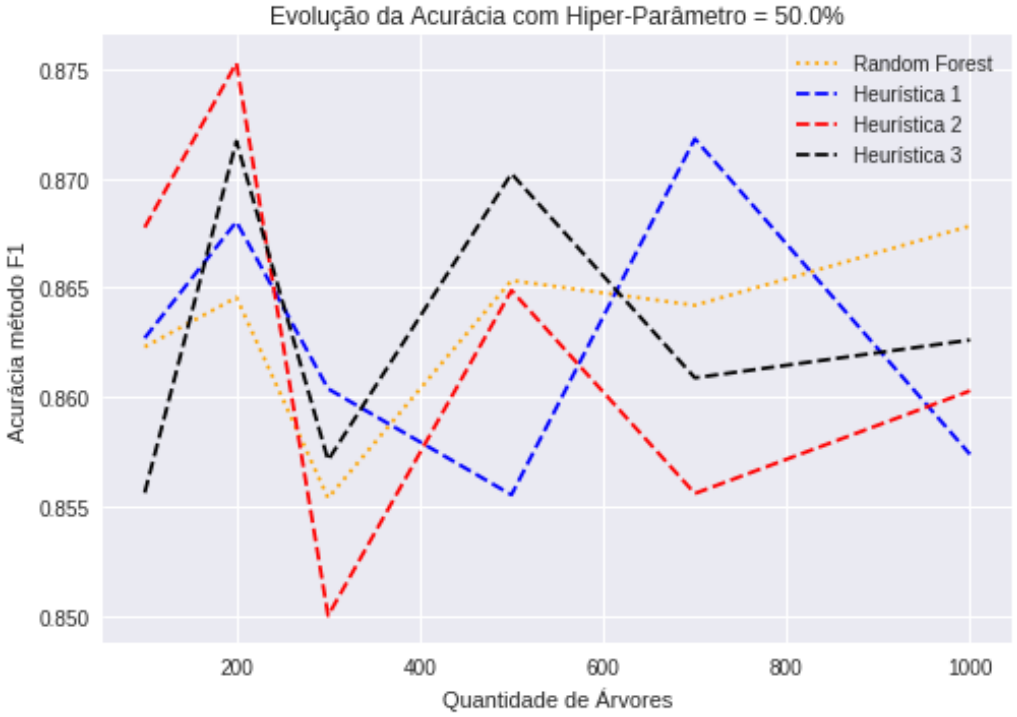

Figura 14 – Evolução do RMSE para as Heurísticas com Coleção: Breast-Cancer, e hiper-parâmetro 50%

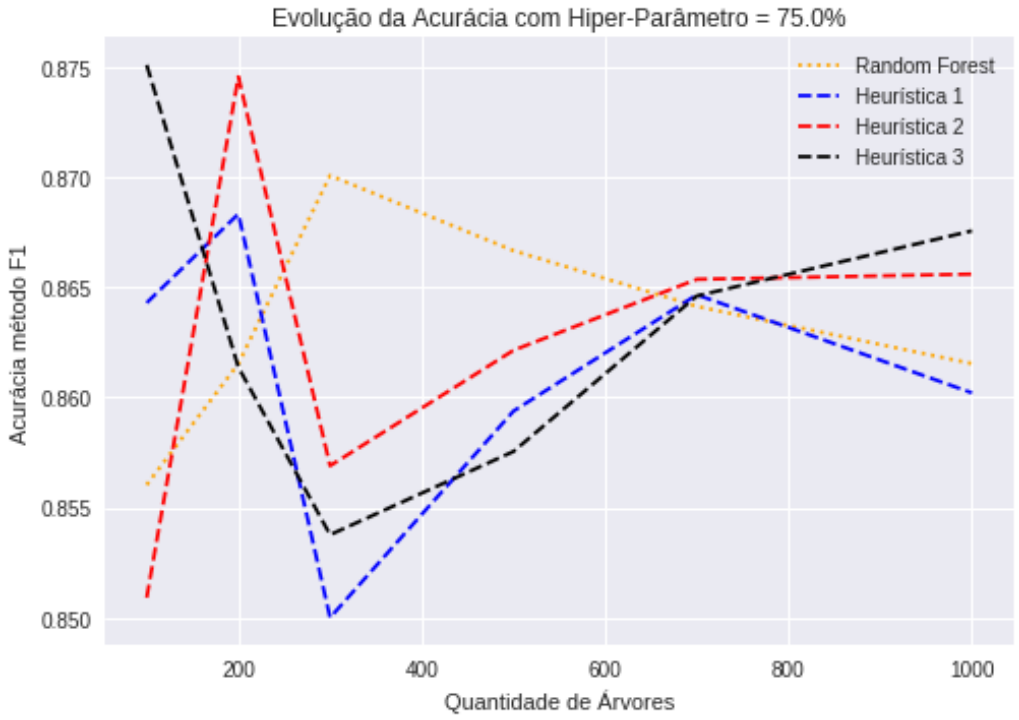

Figura 15 – Evolução do RMSE para as Heurísticas com Coleção: Breast-Cancer, e hiper-parâmetro 75%

5 Conclusão

Conforme visto nos resultados o algoritmo de Random Forest implementado pelo Scikit-Learn sem a aplicação de Heurísticas foi melhor para os conjuntos de dados avaliados em relação às propostas de Heurísticas. Em futuros testes poderão ser comparados esses métodos utilizando bases de dados com modelos já bem determinados com o intuito de verificar e tentar determinar novas propostas de Heurísticas.

Foi possível com a realização desses testes comparar o impacto da seleção de Árvores em algoritmos de Random Forest para Classificação e Regressão. Embora essas hipóteses de Heurísticas não tenham tidos resultados melhores do que a implementação sem elas, isso não significa que a Seleção de Árvores é um método falho, ao contrário, alguns resultados mostraram que não é necessário todas ás Árvores para se ter uma predição com altos valoers de acurácia. Dessa forma para futuros estudos pode-se considerar outras porpostas de Heurísticas para seleção de Árvores, ou ainda a aplicação desses com reformulações.

Referências

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 6.

DEB, S. K.; JAIN, R.; DEB, V. Artificial intelligence—creating automated insights for customer relationship management. In: IEEE. *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. [S.l.], 2018. p. 758–764. Citado na página 5.

DIMITRIADIS, S. I. et al. How random is the random forest? random forest algorithm on the service of structural imaging biomarkers for alzheimer's disease: from alzheimer's disease neuroimaging initiative (adni) database. *Neural regeneration research*, Wolters Kluwer–Medknow Publications, v. 13, n. 6, p. 962, 2018. Citado 2 vezes nas páginas 2 e 7.

REIS, E. et al. Estatística aplicada. *Lisboa: Edições Sílabo*, 1999. Citado 3 vezes nas páginas 2, 8 e 9.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Malaysia; Pearson Education Limited,, 2016. Citado 2 vezes nas páginas 5 e 6.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, n. 3, p. 210–229, 1959. Citado na página 6.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding machine learning: From theory to algorithms*. [S.l.]: Cambridge university press, 2014. Citado na página 5.