FACULDADES METROPOLITANAS UNIDAS - FMU

WANDERLEY JOSÉ ZOTELLI DOS SANTOS

# ANÁLISE COMPARATIVA ENTRE ALGORITMOS DE CLASSIFICAÇÃO

SÃO PAULO

2017

SUMÁRIO

1. OBJETIVO…........…………………………………………………………………………..3

2. DESENVOLVIMENTO…...….…………………………………………………………….3

2.1 ÁRVORES DE DECISÃO.….…………………………………………………………….3

2.2. ALGORITMO KNN (K-NEAREST NEIGHBORS)……....……………………………..3

2.3.ANALISE 1 – J48.................................................................................................................3

2.4.ANALISE 2 - KNN...............................................................................................................6

3. RESULTADOS OBTIDOS.………………………………...………………………………7

4. CONCLUSÃO……………………….……………………………………………………...7

1. OBJETIVOS

Este trabalho tem como finalidade, uma pesquisa sobre o dataset Hipotireoidismo aplicado a dois algoritmos de predição, quais são os fatores de risco, a relevância dos atributos e uma forma de simplificar o entendimento sobre as causas desta doença.

2. DESENVOLVIMENTO

Neste capítulo, abordaremos o problema proposto e descreveremos os algoritmos de predição escolhidos para a aplicação do dataset.

2.1 ALGORITMO DE CLASSIFICAÇÃO

O Algoritmo escolhido para a pesquisa é o J48, um algoritmo de árvores de decisão. O algoritmo J48 é uma implementação em linguagem Java de outro algoritmo, o C4.5 (linguagem C). Este algoritmo utiliza um conjunto de dados de treinamento na construção de árvores de decisão. No exercício proposto, achei pertinente o uso deste algoritmo pois a ilustração das causas e efeitos fica mais compreensível na forma de uma árvore de decisão.

2.2. ALGORITMO KNN (K-NEAREST NEIGHBORS)

Diferente do J48 que é um algoritmo de árvores de decisão, o KNN é um algoritmo de *Lazy Learning*, e é um dos algoritmos mais simples de classificação. Sua função é reconhecer padrões à partir da descoberta do vizinho mais próximo, numa certa distância. Como o nome já diz, o KNN, o algoritmo classifica um dado elemento de acordo com um valor euclidiano K.O valor reflete o número de elementos próximos que serão agrupados, os elementos agrupados são então avaliados e o elemento mais frequente se torna o elemento dominante.

2.3. ANÁLISE 1 – CLASSIFICAÇÃO

O dataset selecionado possui 3772 instâncias e 30 atributos. Para iniciar a análise, foi feito uma análise dos resultados preliminares com o algoritmo de árvore “J48”, seguido da matriz de confusão e curva de erros e sua respectiva árvore, conforme imagens abaixo:

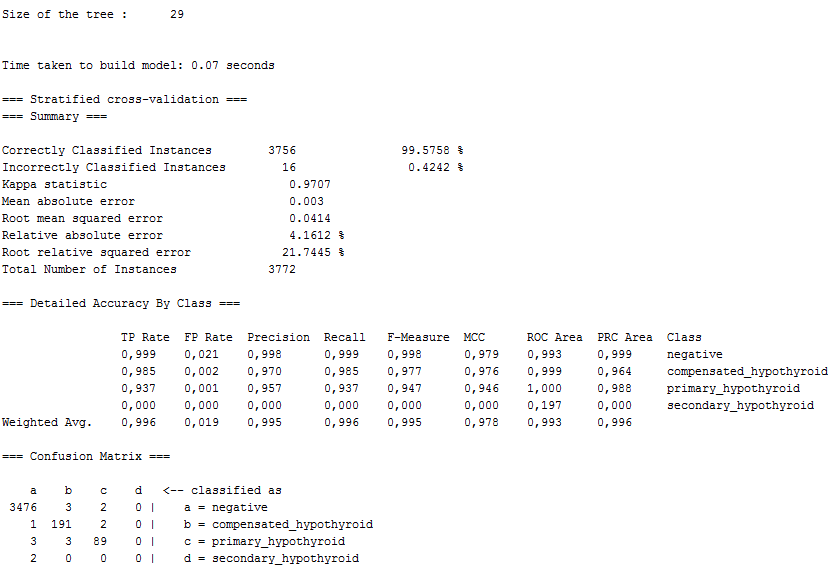


Figura 1 – Matriz de confusão do algoritmo J48

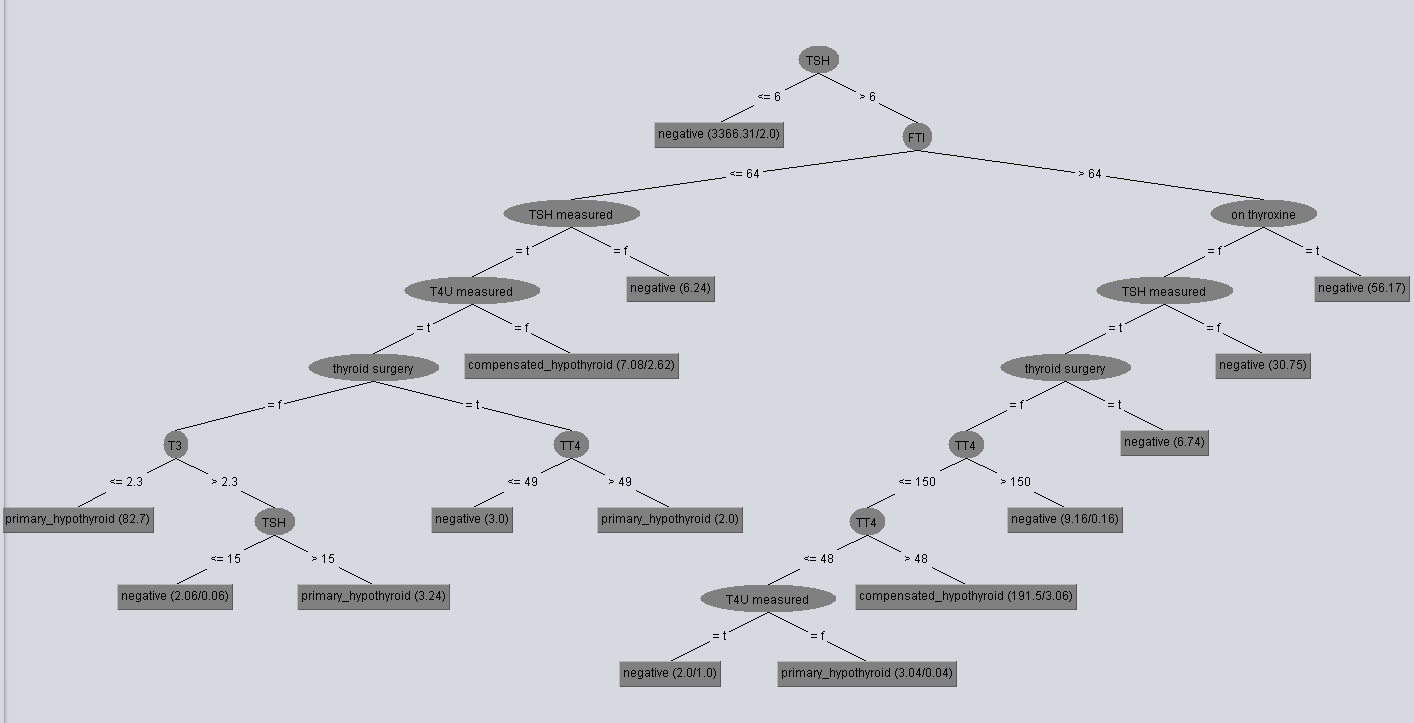


Figura 2 – Árvore de decisão

Após esta análise preliminar, pode-se perceber pela árvore de decisões que muitos dos atributos fornecidos pelo dataset não são relevantes com a ocorrência ou não de hipotireoidismo.

O passo seguinte foi eliminar alguns desses atributos não relevantes para o resultado final, com o intuito de melhorar a acertividade e também simplificar a árvore de decisão.

Após a remoção dos atributos, os resultados foram:

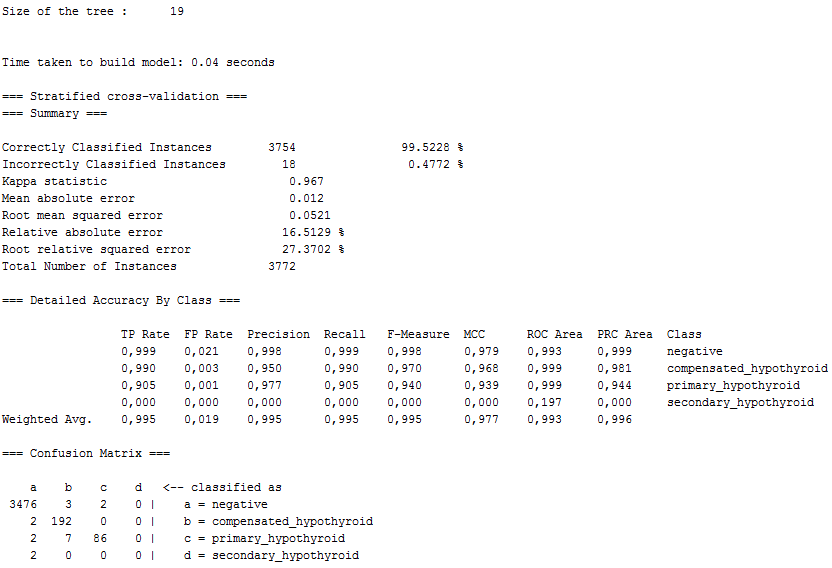


Figura 3 – Matriz de Confusão com algoritmo J48

É possivel notar apenas uma diferença na Area ROC, o restante é praticamente igual. Porém, a árvore foi simplificada de forma muito satisfatória:

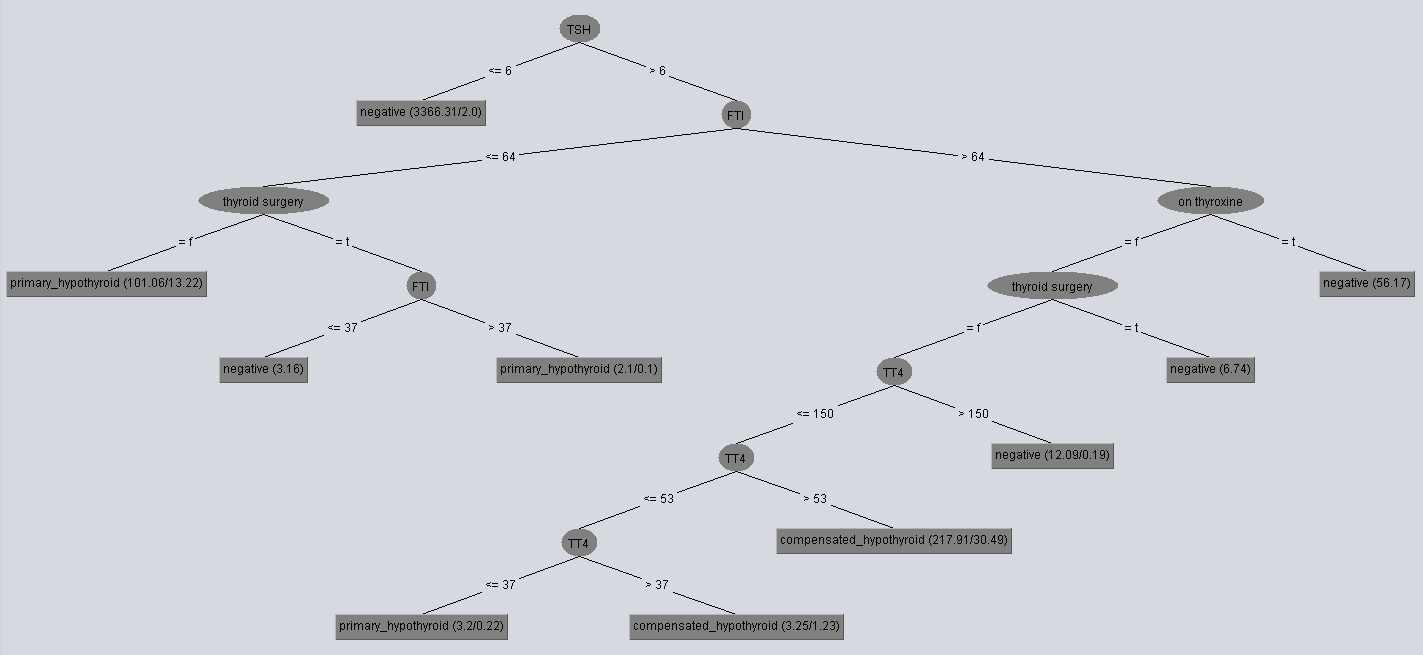
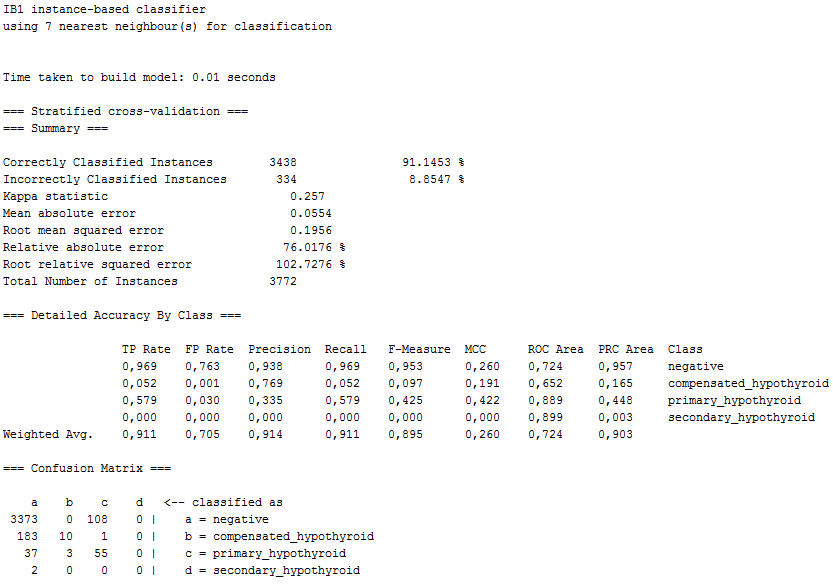


Figura 4 – Árvore de decisão simplificada

2.4. ANALISE 2 - K-NN

O conjunto de dados utilizado para a segunda análise será o mesmo dataset de hipotireoide. Para iniciar a análise, foi feito uma análise dos resultados com o algoritmo “IBK”, seguido da matriz de confusão e curva de erros e sua respectiva árvore, conforme imagens abaixo:

Figura 4 – Matriz de Confusão do IBK

3. RESULTADOS OBTIDOS

Podemos constatar que a taxa de acertividade se mantem alta, porém a área ROC não impressiona muito se comparado ao J48, o valor de K também se mostra bem sensível nesta etapa, uma leve mudança pode trazer enorme impacto nos resultados.

4.CONCLUSÃO

Com ajuda dos algoritmos de classificação, e com a ajuda deste dataset, foi possível validar todas as descobertas feitas até agora sobre o hipotireoidismo e a relação da baixa produção de certos hormônios como a causa da doença.

Claro que, este é um cenário já conhecido, porém, os algoritmos foram capazes de chegar a conclusões precisas quan as causas e fatores de risco da doença.