Estudo de Caso - Automação de diagnóstico de Câncer de Mama.

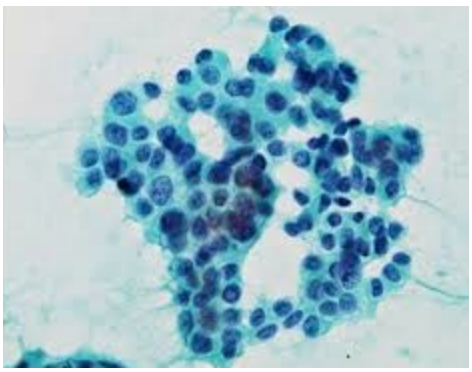
Data de entrega: 7 de dezembro de 2015.

Logística

Grupos de, no máximo, cinco estudantes.

O Cenário

Anualmente pouco mais de 500 mil mulheres morrem vítimas do Câncer de Mama, e cerca de 1,7 milhões são diagnosticadas com o tumor. As chances de cura para aquelas que têm o diagnóstico precoce são até 90% maiores. Por isso é importante que sejam realizados exames que apontam a presença do tumor com frequência.
A mamografia é o exame para a detecção precoce do câncer de mama. Esse exame de imagem tem o papel de identificar lesões (tumores) nas mamas com potencial malignidade. Uma vez identificado risco no laudo mamográfico, um exame complementar, chamado Punção Aspirativa por Agulha Fina (PAAF), é realizado para confirmação de malignidade da neoplasia (câncer).
A massa de células aspiradas a partir da PAAF é então submetida a análise microscópica de nível celular. O médico espacialista em diagnóstico de células neoplásicas observa a massa aspirada sobre uma lâmina de microscópio e as classifica em malignas ou benignas.



O Problema

O diagnóstico de classificação da neoplasia é custoso e pode levar dias até o material coletado ser avaliado pelo médico. Dependendo do local onde o exame foi realizado, essa especialidade médica pode não ter mão-de-obra disponível, implicando na extensão de prazos e em maiores custos para a obtenção do laudo.

O Banco de Dados

Um banco de dados contendo 569 diagnósticos de neoplasias mamárias (Câncer de Mama) está disponível no link.

Cada instância do banco consiste de 10 métricas de uma imagem de células coletadas pela PAAF, contendo as seguinte informações:

1) numero identificador (ID) da imagem
2) diagnóstico (M = maligno, B = benigno)
3-32) Dez métricas computadas para cada núcleo de célula:

a) raio (média das distâncias entre o centro e o pontos do perímetro)
b) textura (desvio-padrão dos valores na escala de cinza)
c) perímetro
d) área
e) suavidade (variação local em tamanhos de raio)
f) compactação (perimetro^2 / área - 1.0)
g) concavidade (severidade das porções concavas do contorno)
h) pontos concavos (número de porções concavas do contorno)
i) simetria
j) dimensão fractal

Para cada uma dessas medidas são cálculadas a média, o desvio padrão, os piores valores (média dos três maiores valores) de cada imagem, resultando em 30 atributos. Assim, por exemplo, o atributo 3 é a média do raio, o atributo 13 é o desvio-padrão do raio e o atributo 23 é o “pior” valor de raio.

Todos os atributos estão com precisão de 4 dígitos.

Não há atributos faltantes.

Distribuição das classes: 357 benignos, 212 malignos.

A Tarefa

A tarefa consiste em criar um *workflow* no Knime com o objetivo de treinar um modelo de classificação eficiente e robusto para o diagnóstico de Câncer de Mama. Para isso você deve realizar um experimento utilizando os seguintes modelos:

- vizinhos mais próximos
- classificador Naive Bayes
- Árvores de Decisão
- regressão logística
- redes neurais artificiais
- máquinas de vetores de suporte

Cada um dos modelos deve ter seus parâmetros ajustados com objetivo de maximizar sua capacidade de generalização: aumentando a acurácia e diminuindo a variância. Portanto, o

workflow a ser construído deve implementar um esquema de validação cruzada, garantindo eficiência e robustez do modelo.

Ainda, parte da tarefa pode exigir que os atributos das imagens sejam tratados a fim de se obter melhores resultados na classificação. Desta forma, você deve pesquisar formas de transformação e redução dimensionalidade nos dados.

Métricas

Deve-se utilizar diversas métricas para análise dos resultados, entretanto são obrigatórias:

1) Tabela de confusão
2) Acurácia
3) Taxa de erro
4) Precisão
5) Cobertura

Questões para análise

1) Quais modelos têm maior acurácia?
2) Quais modelos têm menos variância?
3) Quais modelos erram menos casos graves?
4) Quais casos (diagnósticos) são mais difíceis de classificar?
5) Quais casos (diagnósticos) podem ser vistos como ruído?
6) Quais atributos são mais decisivos para classificar entre Malignos e Benignos?

Artefatos Produzidos

1) *Workflow* completo de todo o estudo de caso implementado no Knime.
2) Relatório em formato de artigo no LaTex, contendo a discussão, gráficos e tabelas dos resultados.
3) Apresentação do trabalho em formato PPT/PDF a ser realizada presencialmente.