Risco de Diabetes

Rene S. Freire

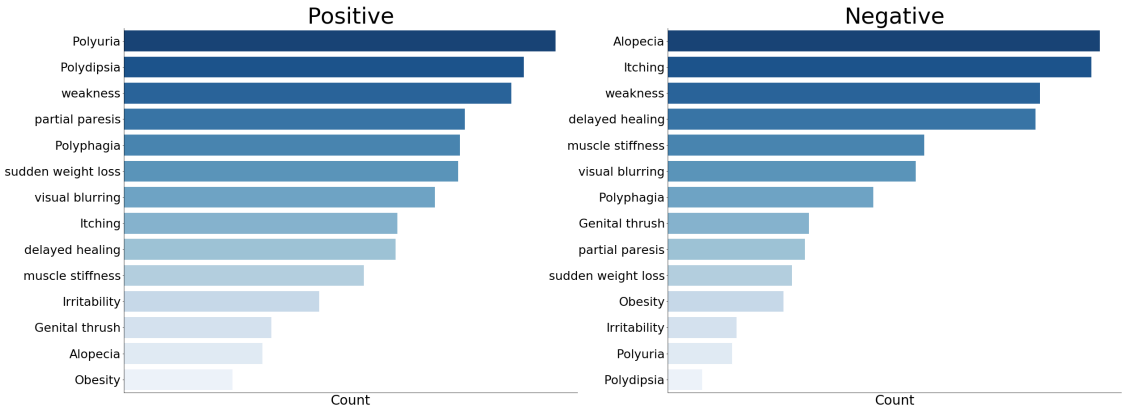
Fazemos um estudo do risco de diabetes precoce, baseado em dados de 520 pessoas, das quais 328 são homens e 192 são mulheres, entre 16 e 90 anos (média de 48 anos).

Esses dados reúnem, além das informações sobre o sexo, a idade e o diagnóstico de diabetes, alguns sintomas: poliúria, polidipsia, perda de peso brusca, fraqueza, polifagia, candidíase genital, visão turva, coceira, irritabilidade, cura retardada, paresia parcial, rigidez muscular, alopecia e obesidade.

Sintomas

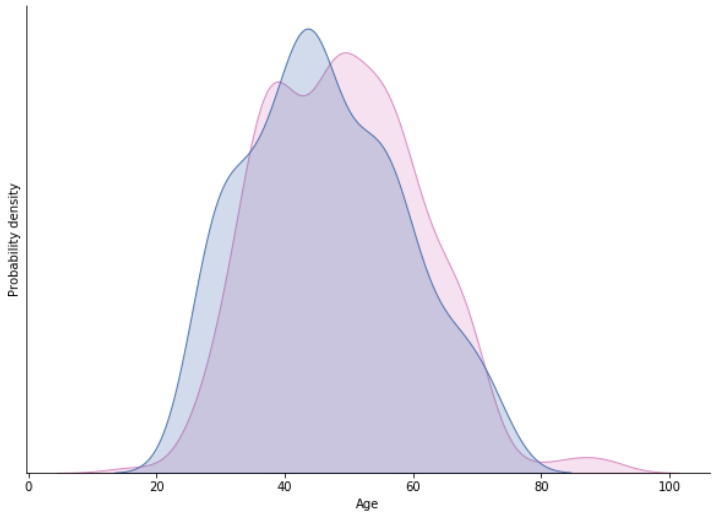
Desses sintomas, os que têm maior correlação com o diagnóstico positivo são poliúria e polidipsia (urina excessiva e sede excessiva, respectivamente). Os que têm menor correlação são coceira e cura retardada.

O gráfico abaixo (que é um gráfico de contagem de casos, não de correlação) ilustra mais ou menos isso:



Idade

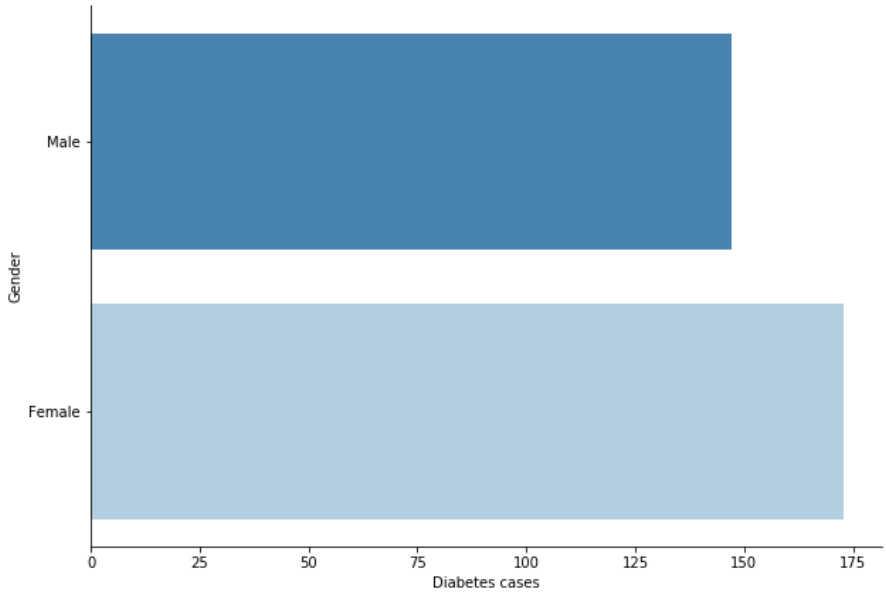
A idade não possui correlação alta com o diagnóstico, mas o nosso modelo a utiliza como principal fator para classificação. Isso ocorre por que há predominância de diagnóstico negativo nas idades menores e predominância de diagnóstico positivo nas idades mais avançadas, conforme o gráfico abaixo.



No gráfico acima, a área azul é a densidade de probabilidade variando com a idade do diagnóstico ser negativo, a área vermelha, por outro lado, é a densidade de probabilidade, variando com a idade, do diagnóstico ser positivo.
Ou seja, apesar de haver uma grande intersecção (e por isso a correlação com o diagnóstico é baixa), nos extremos e na média há predominância de um diagnóstico versus o outro.

Sexo

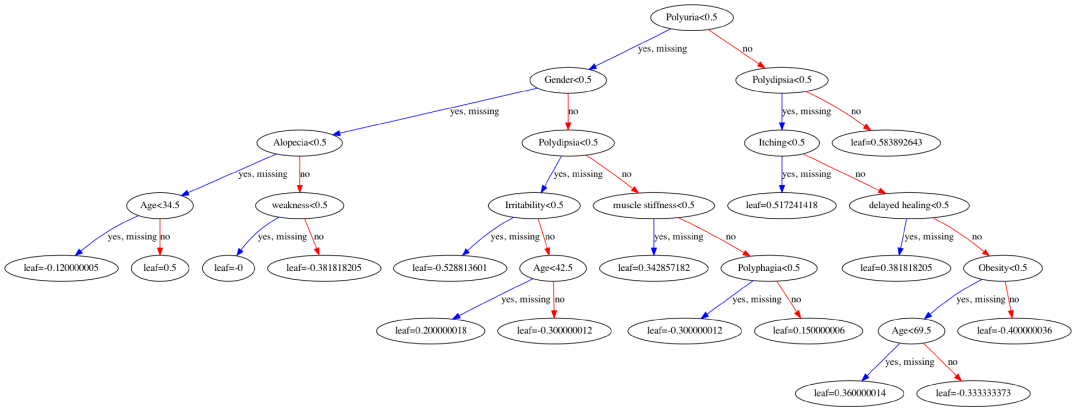
Conforme o gráfico abaixo, também há uma relação entre o gênero e o diagnóstico:



No eixo vertical, os sexos, no eixo horizontal, a contagem de casos positivos. A princípio não parece que a diferença é muito grande, mas temos que lembrar que nesses dados há mais homens que mulheres. Então, o sexo acaba sendo um fator importante no nosso modelo (de fato, como veremos abaixo, é o terceiro fator mais importante para maximizar a precisão do nosso modelo).

O Modelo

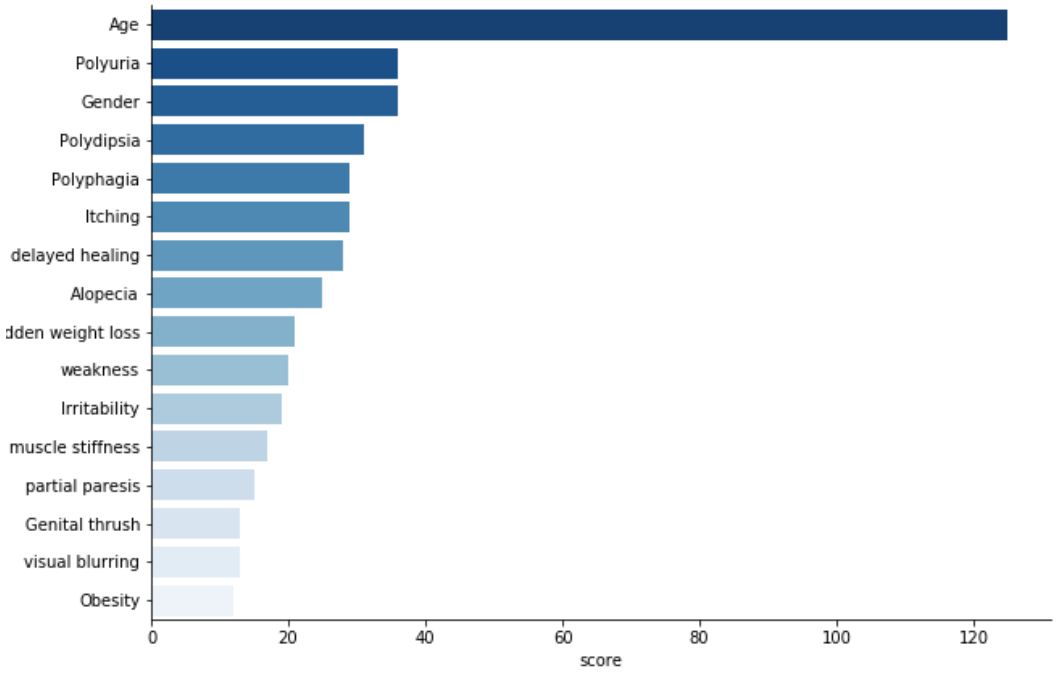
Para classificação usamos o *XGBoost*, que é um modelo sofisticado do tipo árvore (uma espécie de fluxograma). Um exemplo de uma dessas árvores (o algoritmo gera várias, e depois faz uma compilação de todas para construção do modelo final) é mostrado na figura abaixo:



As últimas cédulas da imagem acima são as folhas (*leaf*) da árvore. Como essa é uma árvore particular, e o modelo final é uma combinação não trivial de várias árvores, as folhas contém parâmetros, ao invés do resultado final (viz., positivo ou negativo).

Na construção do nosso modelo queremos selecionar as propriedades mais importantes para a classificação do diagnóstico. Mas isso não pode ser feito de forma heurística: e.g., não podemos olhar para os gráficos acima e discernir uma combinação de idade, sexo e sintomas. Precisamos ver quais propriedades o algoritmo classifica como as mais importantes.

O gráfico abaixo ordena as propriedades de acordo com o peso que o algoritmo dá para cada uma.
O eixo horizontal é o peso que o modelo atribui para cada propriedade.



A idade é a propriedade com maior peso, apesar de não ter correlação forte com o diagnóstico. Isso se deve ao fato que comentamos acima: nos extremos há predominância de um diagnóstico sobre o outro. Como o nosso modelo é do tipo árvore, ele utiliza a idade muitas vezes como critério de classificação.

Um outro fato interessante é que a obesidade é o fator menos decisivo, o que acaba com um preconceito sobre diabetes.

Treinamos novamente o modelo usando as melhores propriedades (excluindo as cinco piores). Usamos 75% dos dados para treinar o modelo e 25% para validar. Essa divisão foi feita de forma aleatória, mas preservando a proporção positivos para negativos em cada uma.

O modelo final nos dá probabilidades de classificação. Após fazer uma otimização da linha que diz o que é uma probabilidade alta o suficiente para classificar uma pessoa como positiva ou negativa, obtivemos uma precisão de aproximadamente 98.5% na classificação.