Descoberta de Conhecimento - PL01

Vasco Ramos, PG42852

Um problema que pode tirar partido da aplicação de Data Mining é a conceção de um modelo para conseguir prever se um paciente tem ou não diabetes, sendo este processo de decisão baseado em medidas de diagnóstico dos pacientes.

Business Understanding

Tendo em conta o baixo desenvolvimento de populações indígenas como o povo Pima (um grupo de nativo-americanos localizado na região entre o estado do Arizona e o México), especialmente, no que diz respeito à saúde é essencial tanto para ajuda humanitária como para controlo e cura de doenças ser capaz de prever e determinar, consoante dados e informação (estatística), a presença ou não de uma dada doença/problema de saúde numa pessoa deste tipo de povos, neste caso específico no povo Pima. Assim sendo, a principal necessidade do negócio é ser capaz de facilmente identificar, de forma analítica/preditiva, se um dado paciente tem a condição de saúde, neste caso específico, se tem diabetes, reduzindo assim o número de casos de diabetes não acompanhados/tratados.

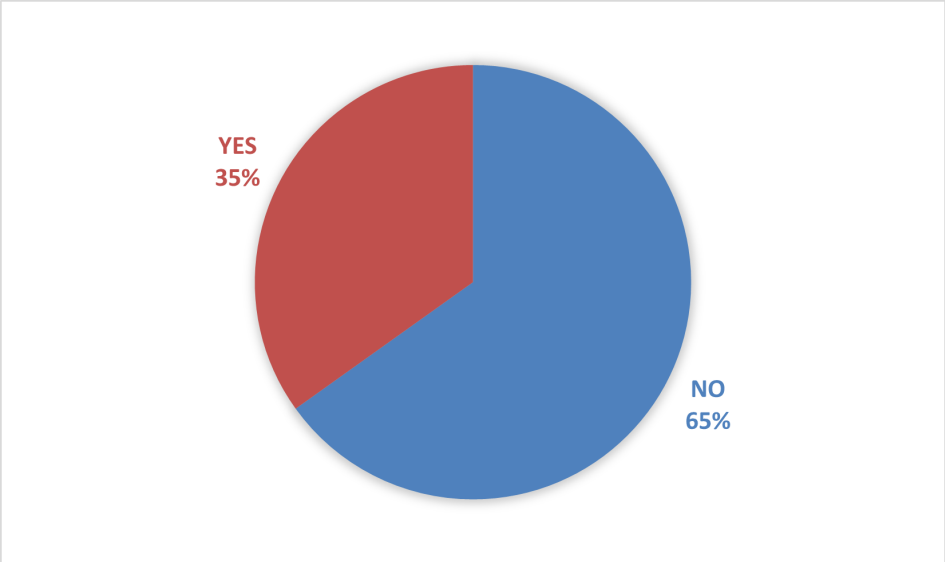
O objetivo da aplicação de *Data Mining* associado a este problema é a capacidade de prever se um paciente tem diabetes, consoante um conjunto de condições de saúde do paciente (como o número de vezes que já esteve grávido, o índice de massa corporal, a idade, entre outros). Para se poder avaliar se a aplicação tem sucesso na correta identificação de casos positivos de diabetes é necessário estabelecer um patamar mínimo do que será aceitável para um modelo desta natureza, pelo que, considera-se aceitável uma precisão de, pelo menos, 70-75%, o que significa que o modelo deve ser capaz de identificar com sucesso ¾ dos casos existentes de diabetes.

Data Understanding

O conjunto de dados utilizado para este problema vem de acompanhamento ao longo realizado por equipas de estudo de saúde dos USA, disponibilizado no Kaggle. Este conjunto de dados consiste em 768 entradas e representa apenas pacientes do sexo femino com idade igual ou superior a 21 anos.

- #pregnancies: o número de vezes que a paciente já esteve grávida
- plasmaGlucose: concentração da glucose plasma
- diastolicBP: pressão arterial diastólica (mm/Hg)
- skinThickness: a grossura dos tecidos de pele dos tríceps (mm)
- serumInsulin: concentração da insulina serum a 2 horas (muU/ml)
- BMI: indíce de massa corporal
- DPF: função do pedigree da diabetes
- age: idades (anos)
- outcome: a classificação, ou seja, se tem ou não diabetes (NO/YES)

Deste modo, como se pode ver, este é um problema de classificação a uma variável, sendo que o NO/YES pode facilmente ser transformado em 0/1. Exceto o outcome (que é uma variável nominal), todas as variáveis são valores numéricos. Como se pode ver na imagem abaixo, dos dados existentes, 65% (500 pessoas) não têm diabetes e 35% das pacientes (268 pessoas) têm diabetes.



Para além disso, tendo em conta que, segundo os estudos efetuados, um dos principais fatores que pode mais facilmente influenciar a presença da diabetes é o número de filhos (um dos estudos iniciais destes dados foi precisamente para tentar encontrar uma correlação entre estas duas variáveis). A tabela abaixo mostra-nos de forma resumida a distribuição do número de vezes que as pacientes da amostra engravidaram, através dos valores de ocorrência máxima, mínima e média.

Max of #pregnancies	Min of #pregnancies	Average of #pregnancies
17	0	4

Benefícios da Aplicação de Data Mining

Por regra, a aplicação de data mining permite: otimizar o valor dos dados existentes sobre um dado problema extraindo conhecimento dos mesmos, tornar mais percetível e analisável o estado atual do problema e apontar *trends* ou então acrescer a possibilidade de facilmente prever ocorrências de eventos, tendo por base todos os dados históricos existentes.

Na situação específica do problema em análise a aplicação de data mining traz a grande vantagem de se conseguir pegar nos dados de que se dispõe e ser capaz de mais facilmente os analisar para criar um modelo capaz de uma forma algo eficiente identificar pacientes (do sexo feminio) do povo Pima com diabetes, dado um conjunto de caracteristicas sobre cada paciente.