Relatório Análise de dados Diabetes

Carlos Frederico Carvalheira Mello

Objetivos

Fazer uma análise estatística dos dados de diagnósticos de diabetes visando a construção de um sistema de diagnóstico da doença

Metodologia

O trabalho teve como metodologia uma análise quantitativa da variância dos dados, visando reduzir o máximo o número de variáveis dos dados finais, usando técnicas de remoção visual, observando fatores como a distribuição dos valores das colunas e usando transformações ortogonais dos dados, técnica conhecida como *Principal Component Analysis* (PCA), que visa transformar esses dados que possivelmente estão correlacionados de alguma forma em um conjunto de valores descorrelacionados linearmente e chamados de componentes.

Resultados

Análise Explanatória

Na primeira análise feita, foi importado a tabela de forma crua, ou seja sem nenhuma alteração na tabela. Em seguida foi feito todo um trabalho de limpeza e conversão de alguns valores incorretos.

A primeira tomada de decisão para remoção de algumas colunas do dataset foi a distribuição de seus valores. Como em teoria os valores que estavam no dataset foram obtidos de exames médicos de uma amostra da população, existe uma certa tendência dos valores corretos gerarem uma curva Gaussiana como na Imagem 1.1 e não como a curva representada na imagem 1.2

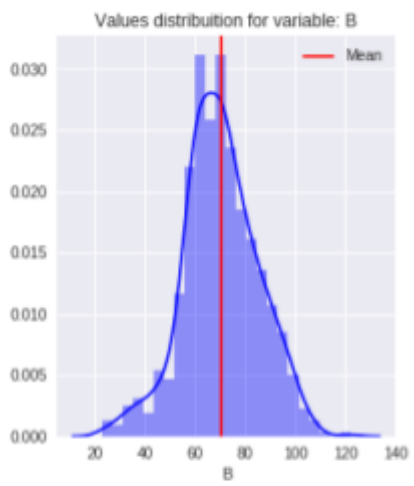

Figura 1.1

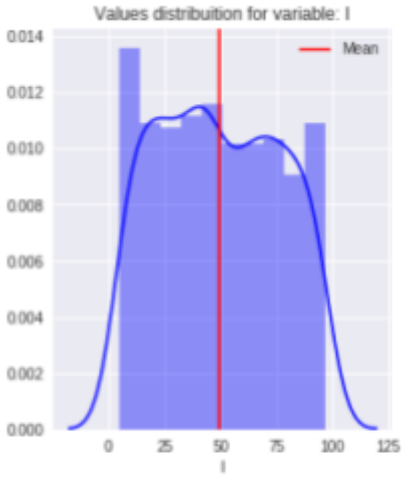

Figura 1.2

Então foram deixadas apenas as colunas que de alguma forma a sua distribuição dos valores lembrar-se o mínimo possível a de uma curva Gaussiana.

Logo após foram feitas as remoções visando remover a redundância de valores entre colunas, foi feita assim uma análise das correlações lineares entre cada uma das das variáveis do dataset e todas estas que tiveram uma correlação entre si próximo a 1 foram removidas. Como pode ser visto na figura 1.3 um mapa de calor foi feito para ilustrar as correlações entre cada variável do dataset, note que quanto mais escura a cor, maior a correlação da variável.

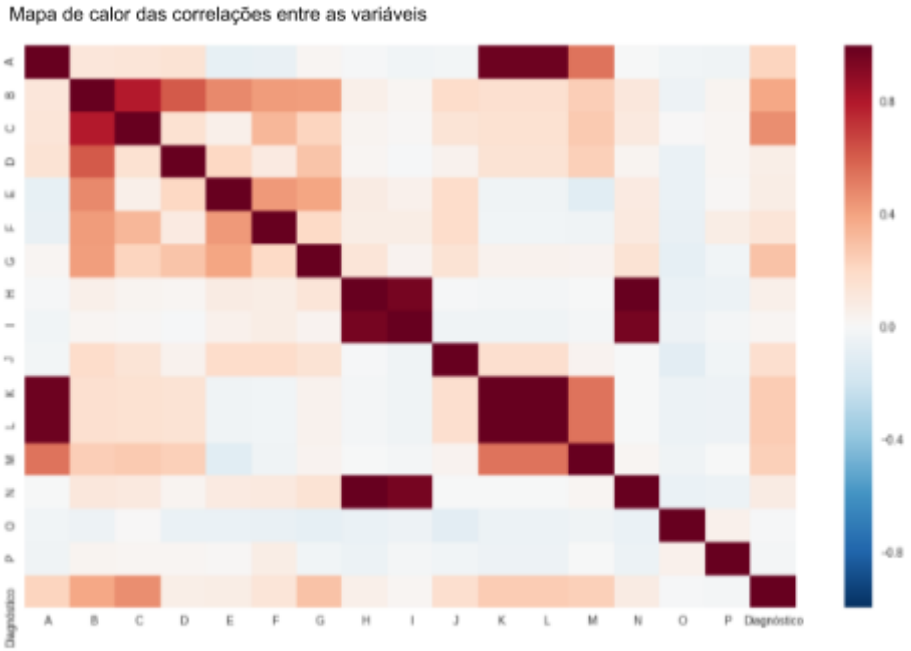

Figura 1.3

Ao final foram removidas as seguintes colunas H, I, K, L, N, O e P. Totalizando sete colunas a menos do dataset, o estado desse conjunto de dados foi salvo para ser usado futuramente na análise de cada componente.

Analise principal de componente(PCA)

Utilizando o dataset resultado da última análise, já com menos features o pca foi usado para reduzir ainda mais a dimensão do dataset em apenas alguns componentes.

O Número final de componentes foi escolhido com base na verificação explanatória de um somatório da representatividade de variância de cada componente. Podemos observar na imagem 1.4 que com apenas 4 componentes finais usando a transformação ortogonal podemos praticamente representar cerca de 99% dos dados do dataset.

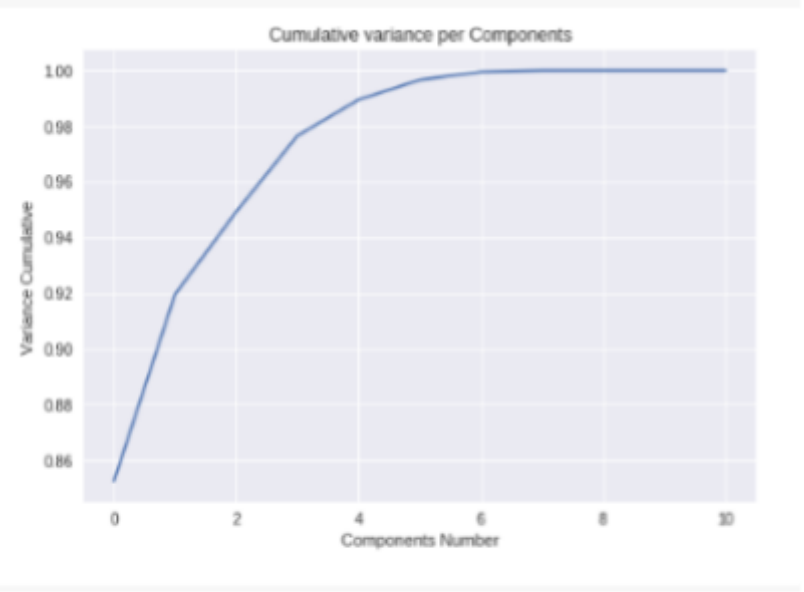

Figura 1.4

Então ao final da transformação feita usando o pca com o número de componentes igual a 4 acabei com um dataset com apenas 4 features que representam cerca de 99% de representatividade dos dados finais, mais ainda a coluna de diagnóstico, totalizando cerca de 5 colunas finais, no qual o mesmo foi salvo para um futuro uso em algum modelo de aprendizado

Conclusão

O Dataset final gerado pela aplicação da remoção das colunas por meio de explanação dos dados, mais a transformação gerada pelo PCA, é consideravelmente menor do que o original, contém ainda quase 100% da informação da tabela inicial e possui já

valores padronizados por meio da função z-score. O Dataset encontra-se pronto para ser dividido nos conjuntos de treino, teste e validação, todos disjuntos e igualmente representativos de cada classe.