Pontifícia Universidade Católica de Goiás

Curso: Big Data e Inteligência Artificial

Da Análise Exploratória ao k-Nearest Neighbour (KNN)

Curso: Big Data e Inteligência Artificial

Luiz Paulo T. Gonçalves

Igor Guimarães de Oliveira

João Paulo M. Rosa

Thiago A. da Silva

GOIÂNIA

2023

1. Introdução

Busca-se no presente trabalho uma rápida análise exploratória e modelar um algoritmo k-Nearest Neighbour (doravante, KNN) – algortimo de machine learning supervisionado. Pois bem, dentro desse objetivo específico encontra-se a problemática que vai ser analisada e modelada: a classificação de indivíduos em relação à presença ou ausência de diabetes com base em determinados inputs de entrada (dados de entrada). O KNN é amplamente utilizado em modelagem de classificação, seja binária ou multiclasse.

Para além desta secção, o trabalho será divido em quatro secções. Na primeira secção, apresenta-se os objetivos gerais e específicos; na terceira, a metodologia de pesquisa; na quarta busca-se apresentar os resultados e, por fim, na quinta e última secção, apresenta-se a conclusão.

2. Objetivos

Como mencionado brevemente na introdução, o objetivo geral do trabalho é modelar e aplicar um algoritmo de classificação KNN. Pois bem, para chegar no objetivo específico o trabalho passa pelos objetivos gerais: aplicar análise exploratória nos dados, transformar os dados quando necessário, análise estatística e aplicar validação cruzada no algoritmo[1].

Todo o desenvolvimento está pautado nos dados sobre diabetes da National Institute of Diabetes and Digestive – está disponível no repositório de datasets da Kaggle[2]. O banco de dados conta com 768 registros, no qual todos os dados são de mulheres com pelo menos 21 anos de idade e de ascendência indígena Pima. Os dados ou variáveis presentes nesse dataset são os seguintes:

- Pregnancies: Para expressar o número de gestações;
- Glucose: Para expressar o nível de glicose no sangue;
- BloodPressure: Para expressar a medida da pressão arterial;
- SkinThickness: Para expressar a espessura da pele;
- Insulin: Para expressar o nível de insulina no sangue;
- BMI: Para expressar o índice de massa corporal;
- DiabetesPedigreeFunction: Para expressar a porcentagem de diabetes;

1 Toda codificação desenvolvida pode ser encontrada no seguinte repositório do Github: https://github.com/LuizPaulo23/ProjetosPUC.

2 Pode ser encontrado para download no seguinte link: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download.

- Age: Para expressar a idade;
- Outcome: Para expressar o resultado final, 1 tem diabetes e 0 é não tem diabetes.

A variável outcome é a variável dependente, isto é, a variável alvo no qual o algoritmo vai ser desenvolvido para classificar a relação binária: diabetes, não diabetes. O restante das variáveis são as variáveis explicativas[3]. Assim, estrutura-se a ideia para o desenvolvimento.

3. Metodologia

A metodologia basicamente resume-se em um algoritmo KNN modelado via Distância Euclidiana. A qual pode ser expressa matematicamente como segue:

$$D_E(p,q) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Na qual calcula-se a distância entre as observações *p* e *q*, isto é, entre as diversas observações das variáveis presentes no problema (ver, BRUCE, 2019). Para minimizar a heterogeneidade entre as observações optou-se por aplicar uma transformação de Zscore:

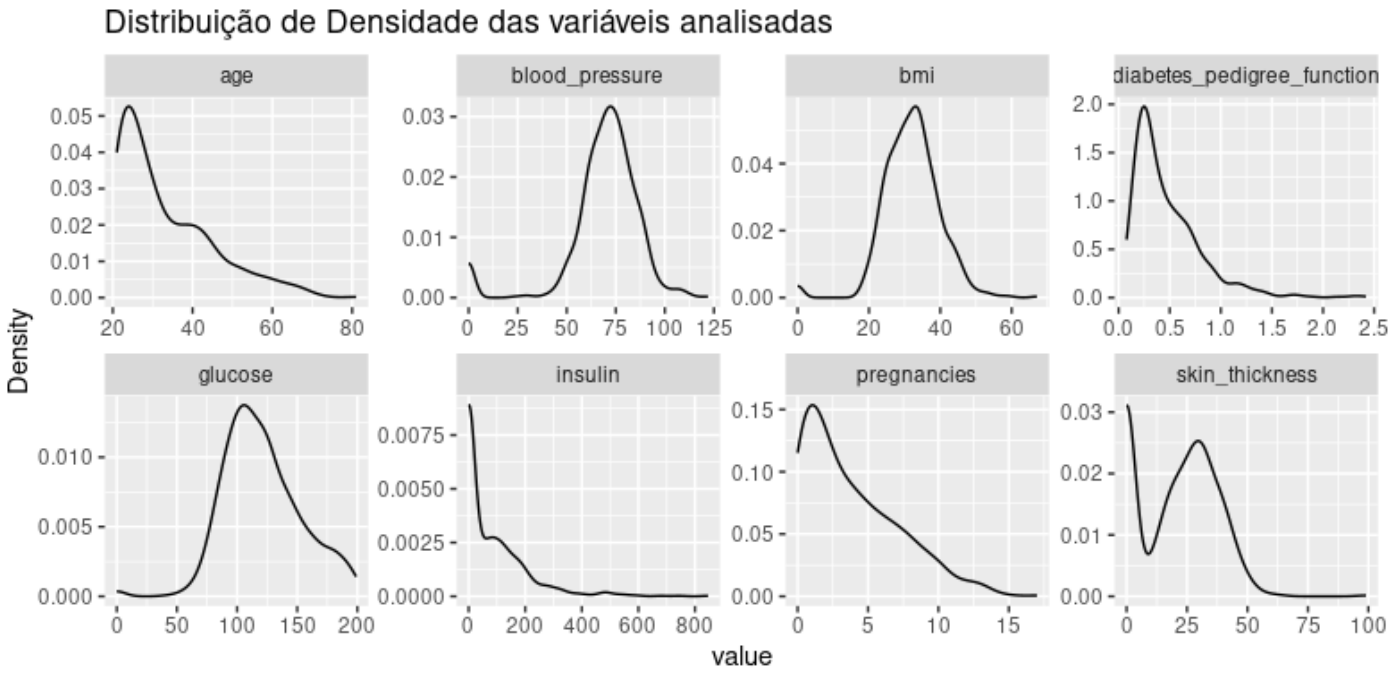
$$Z = \frac{x - \mu}{\sigma}$$

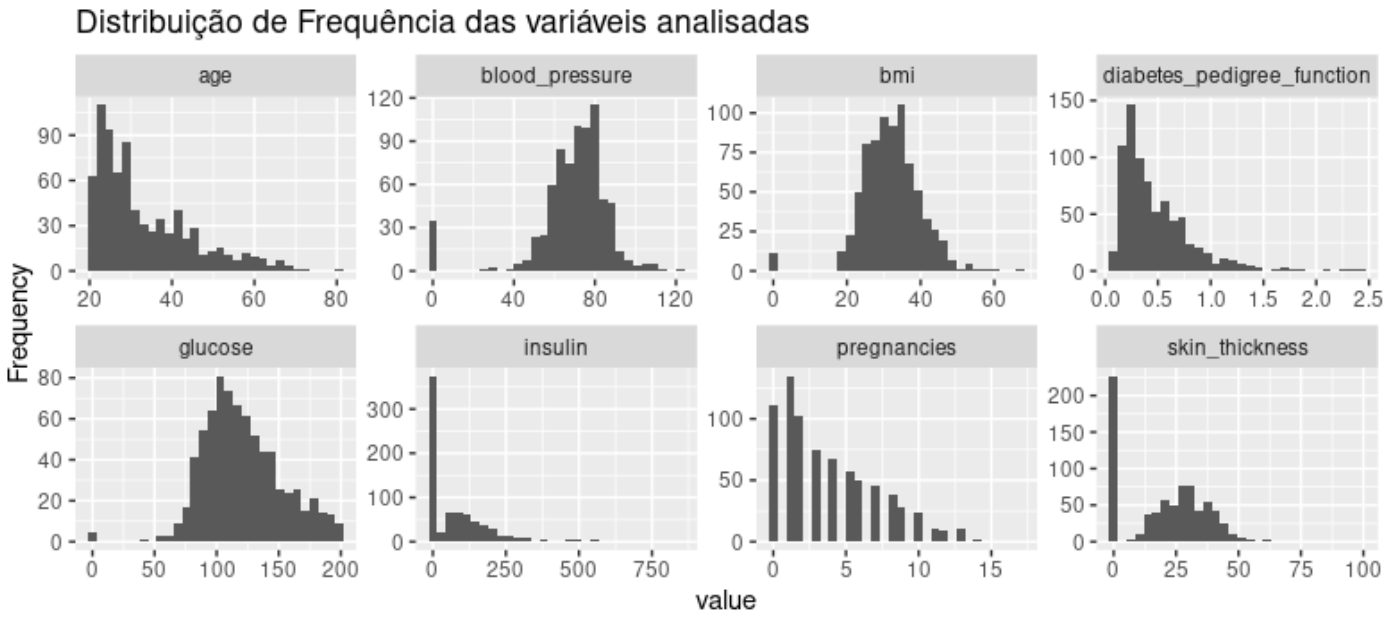
A qual é bem simples, é basicamente uma padronização das observações: μ representa a média e, por sua vez, σ o desvio padrão (MAGALHÃES, 2002). Esses são os pontos fundametais da metodologia desenvolvida.

3 Sobre variáveis explicativas e depedentes, ver: Gujarati & Porter (2011).

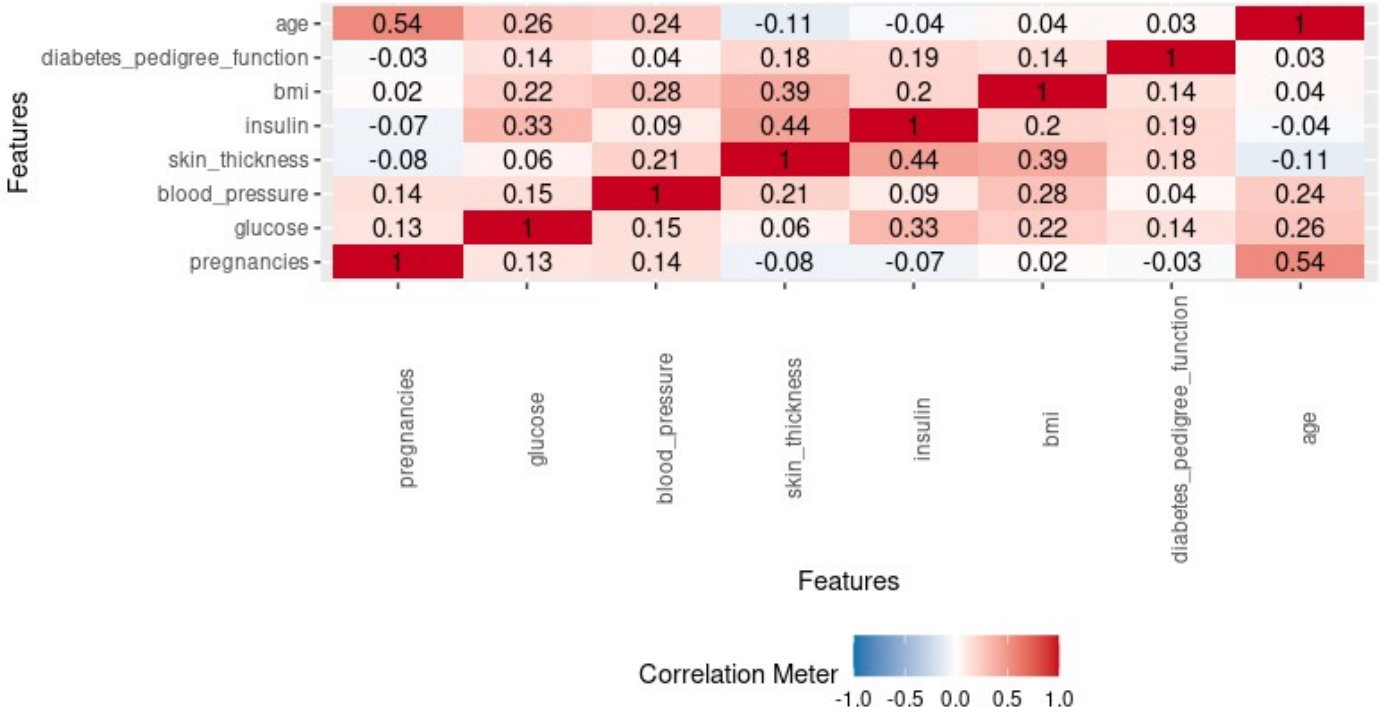
4. Resultados

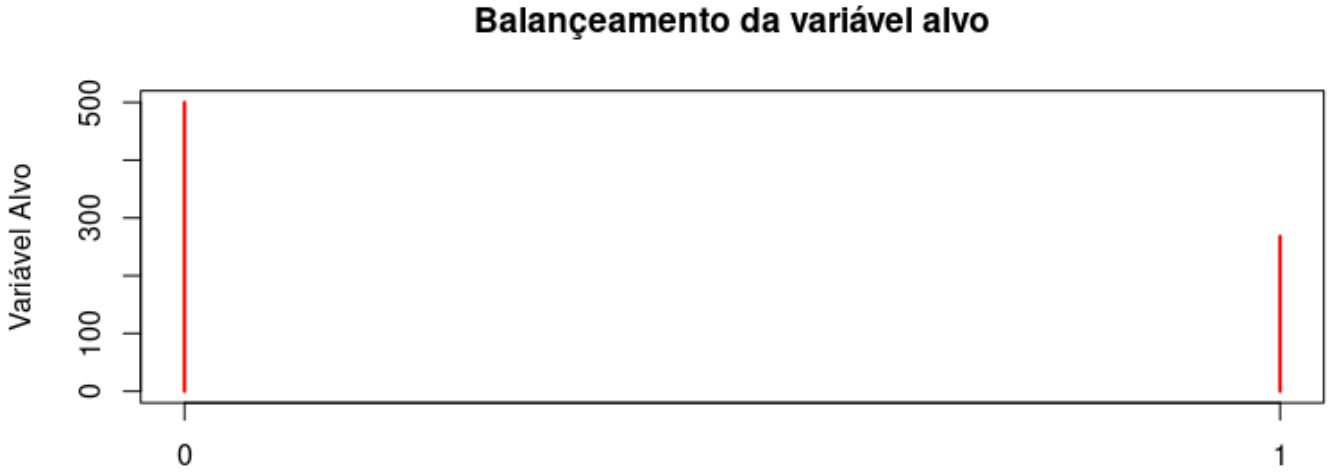
Inicia-se a secção de resultados apresentando alguns dos plotes desenvolvidos na análise exploratória. A seguir, pode-se observar a distribuição de densidade e frequência das variáveis. Nota-se que, com exceção de duas variáveis, a distribuição das variáveis não se enquadra em uma distribuição normal, isto é, Gaussiana.





A correlação entre as variáveis foi outro ponto observado. Na correlação de Pearson a seguir nota-se que basicamente não há uma forte correlação entre as variáveis explicativas. A correlação mais forte é de 0,54 entre número de gestações e idade. Ou seja, evitando o problema de colineridade perfeita entre as variáveis (ver, GUJARATI & PORTER, 2011).





Outro ponto observado, foi sobre o balanceamento da variável alvo. Nota-se que a variável em questão está bem desbalanceada. Com 500 pacientes registrados sem diabetes e 268 com diabetes. Esse desbalaceamento entre 0 e 1 pode ser um problema para o algoritmo. Pois bem, rodando o KNN com 10 K-Folds os resultados na validação cruzada foram os seguintes:

```
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 587, 587, 587, 586, 586, 587,
...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.6967762  0.3273676
   2  0.6821399  0.2986307
   3  0.7213287  0.3795532
   4  0.7168881  0.3653273
   5  0.7196527  0.3707941
   6  0.7222448  0.3748135
   7  0.7232005  0.3763307
   8  0.7219441  0.3720970
   9  0.7308135  0.3901143
  10  0.7285291  0.3818107

Accuracy was used to select the optimal model
 using the largest value.
The final value used for the model was k = 9.
```

Como pode ser observado, a acurácia do algoritmo não está elevada: com 73% de acertos com k-vizinhos igual a 9 usando 85 da base dados para treino. Com certeza isso é efeito do desbalaceamento na variável. Assim, para conseguir obter uma melhor acurácia é necessário aplicar algum método mais avançado para balancear e treinar novamente o algoritmo.

5. Conclusão

Com base nos resultados obtidos, podemos observar que a acurácia do algoritmo não é alta, alcançando apenas 73% de acertos ao usar 9 vizinhos próximos e 85% dos dados para treinamento. Como mencionado anteriormente, essa baixa acurácia pode ser atribuída ao desbalanceamento dos dados na variável alvo.

À guisa de conclusão, o desbalanceamento ocorre quando temos uma proporção significativamente maior de uma classe em relação à outra. No contexto do presente trabalho, pode ser que tenhamos uma quantidade desproporcional de instâncias com diabetes em comparação com as instâncias sem diabetes no conjunto de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

BRUCE, P.; BRUCE, A. **Estatística Prática para Cientista de Dados: 50 conceitos essenciais**. Rio de Janeiro: Alta Books, 2019.

GUJARATI, D.; PORTER, D. **Econometria Básica**. Porto Alegre: AMGH, 2011.

MAGALHÃES, M.; LIMA, A.; **Nocões de Probabilidade e Estatística**. São Paulo: USP, 2002.