Fusão de Informação em Análise de Dados

Tiago Emanuel Pacheco Caldeira Conceição nº 2021167993

1. Considerações iniciais

Para resolver o problema proposto para o projeto dois, foi utilizado o método aprendido nas aulas, Bayes Theory e este foi utilizado para fundir as diversas informações.

Iniciando o projeto, e de acordo com o enunciado foi realizado a fusão dos seguintes sensores σHRBP = 2 e σHRECG = 0,5 utilizando a seguinte formula:

$$\hat{x} = \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right) z_1 + \left(\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right) z_2$$

Após ser efetuado o cálculo, vai resultar numa nova coluna que é o **Heart Rate** (HR), que vai ajudar a determinar resultados mais conclusivos.

Uma vez já efetuada a fusão dos dados, foi realizada as *clinical guidelines* que de acordo com o que estava proposto, caso o paciente tivesse a creatinina superior a 1.3, o segmento st for igual ao valor 1 e o killip for maior ou igual a 2 vai resultar num risco igual a 1, e emerge numa nova feature denominada por *guidelines.*

2. Metodologia

Os dados neste projeto dividem-se em dois tipos de dados:

- Discretos;
- Contínuos;

Desta forma, os dados discretos são os dados que surgem são um número finito, assim sendo as seguintes colunas fazem parte deste tipo de dados:

- Gender;
- Risk Factors;
- ST;
- Killip;
- Guidelines;

Para os dados contínuos, os dados assumem valores infinitos ou seja assumem um intervalo de valores possíveis sem interrupções ou saltos, e no projeto estas são as colunas que seguem este tipo de dados:

- Age;
- Systolic;
- HR;

- Creatinine;

Cada tipo de dados, tem um método distinto de ser calculado, desta forma para calcular os dados discretos temos de realizar as seguintes contas.

$$P(Xi|A)$$

$$P(Xi|\tilde{A})$$

Após ser calculada a probabilidade de o ocorrer o *event* ou não, multiplica-se as probabilidades para conseguir perceber a que classe pertence cada medição.

$$P(Xi|A) > P(Xi|\tilde{A})$$

Caso o exemplo anterior se confirme, a medida vai pertencer a classe A que no caso do nosso problema corresponde ao paciente ter cancro.

Para os dados contínuos, para cada *feature* tem de se calcular a sua média e o seu desvio padrão e após ser efetuado cada calculo para cada uma das condições calculamos a sua probabilidade da seguinte forma.

$$prob(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Desta forma, e após ter sido calculada cada probabilidade, faz-se como nos dados discretos, multiplica-se todas as probabilidades obtidas e vê-se a que classe pertence.

3. Resultados

```
Sensibility --> 0.6145251396648045
Specifity --> 0.920863309352518
Accuracy --> 0.8008752735229759
Precision --> 0.8333333333333334
```

Ao ser analisados os resultados obtidos, conseguimos perceber que de acordo o significado de cada métrica, temos um valor elevado na *specifity* que de acordo a *confusion matrix* corresponde à probabilidade de um paciente não ter cancro.

Apesar de tudo e de ter uma accuracy e precision aceitável, era expectável obter valores superiores, isso pode não estar a ocorrer, por alguma má implementação de métodos, ou por erros do dataset.

No que toca a utilização de todos as variáveis, por vezes a utilização de todas pode-se obter melhores resultados, mas neste caso, após testagem e discussão com alguns colegas, foi notado que em alguns casos neste projeto, após a não utilização de todas as variáveis obteve-se resultados nas métricas superiores.