**AV: Exercício individual - Redução de dimensionalidade: impactos positivos e negativos**

**Enunciado da atividade**

Por vezes o cientista de dados é contratado para extrair conhecimento em bases de dados que não foram planejadas para tal. É comum encontrar bases de dados que não possuem padronização nenhuma entre as tabelas ou mesmo as relações entre estas não são construídas adequadamente. Além disso, de toda a informação disponível no banco de dados, quanto é realmente relevante? Sabemos que atributos como CPF ou qualquer outro tipo de identificador único não contribuem em nada para a análise dos dados, porém, nem sempre as coisas são tão obvias.

Para este exercício você receberá um jupyter-notebook contendo informações e códigos exemplo sobre como utilizar abordagens estatísticas para redução de dimensionalidade: redução por variância e por testes estatísticos univariados (SelectKBest).

Você irá utilizar a base de dados previamente utilizada no exercício anterior, aplicar redução de dimensionalidade e responder as seguintes questões:

* Existe melhora na performance do classificador ao reduzir a dimensionalidade?
* Para o método SelectKBest, qual o valor de K que maximiza a performance?

Realize a entrega da atividade clicando no título no ambiente virtual de aprendizagem.

**Respostas:**

Sim houve melhora significativa com o k=2. Analise abaixo:

def reduceWithScore():

for k in range(2, 7):

x\_new = SelectKBest(chi2, k=k).fit\_transform(X\_, y\_)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x\_new, y\_, test\_size=0.3, random\_state=1)

clf.fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

print('Resultado com redução de dimensionalidade= ', k)

print(classification\_report(y\_test, y\_pred))

reduceWithScore()

Resultado com redução de dimensionalidade= 2

precision recall f1-score support

0.0 0.68 0.62 0.65 205

1.0 0.64 0.69 0.66 197

accuracy 0.66 402

macro avg 0.66 0.66 0.66 402

weighted avg 0.66 0.66 0.66 402

Resultado com redução de dimensionalidade= 3

precision recall f1-score support

0.0 0.57 0.54 0.55 205

1.0 0.55 0.58 0.56 197

accuracy 0.56 402

macro avg 0.56 0.56 0.56 402

weighted avg 0.56 0.56 0.56 402

Resultado com redução de dimensionalidade= 4

precision recall f1-score support

0.0 0.62 0.60 0.61 205

1.0 0.60 0.61 0.60 197

accuracy 0.61 402

macro avg 0.61 0.61 0.61 402

weighted avg 0.61 0.61 0.61 402

Resultado com redução de dimensionalidade= 5

precision recall f1-score support

0.0 0.60 0.56 0.58 205

1.0 0.57 0.61 0.59 197

accuracy 0.58 402

macro avg 0.59 0.59 0.58 402

weighted avg 0.59 0.58 0.58 402

Resultado com redução de dimensionalidade= 6

precision recall f1-score support

0.0 0.61 0.55 0.58 205

1.0 0.57 0.63 0.60 197

accuracy 0.59 402

macro avg 0.59 0.59 0.59 402

weighted avg 0.59 0.59 0.59 402