今回の総合実験では課題として糖尿病患者の判別とパーキンソン病患者の判別を行った。

課題1 糖尿病患者の判別

　とりあえず散布図行列を見て全体の傾向を掴む。

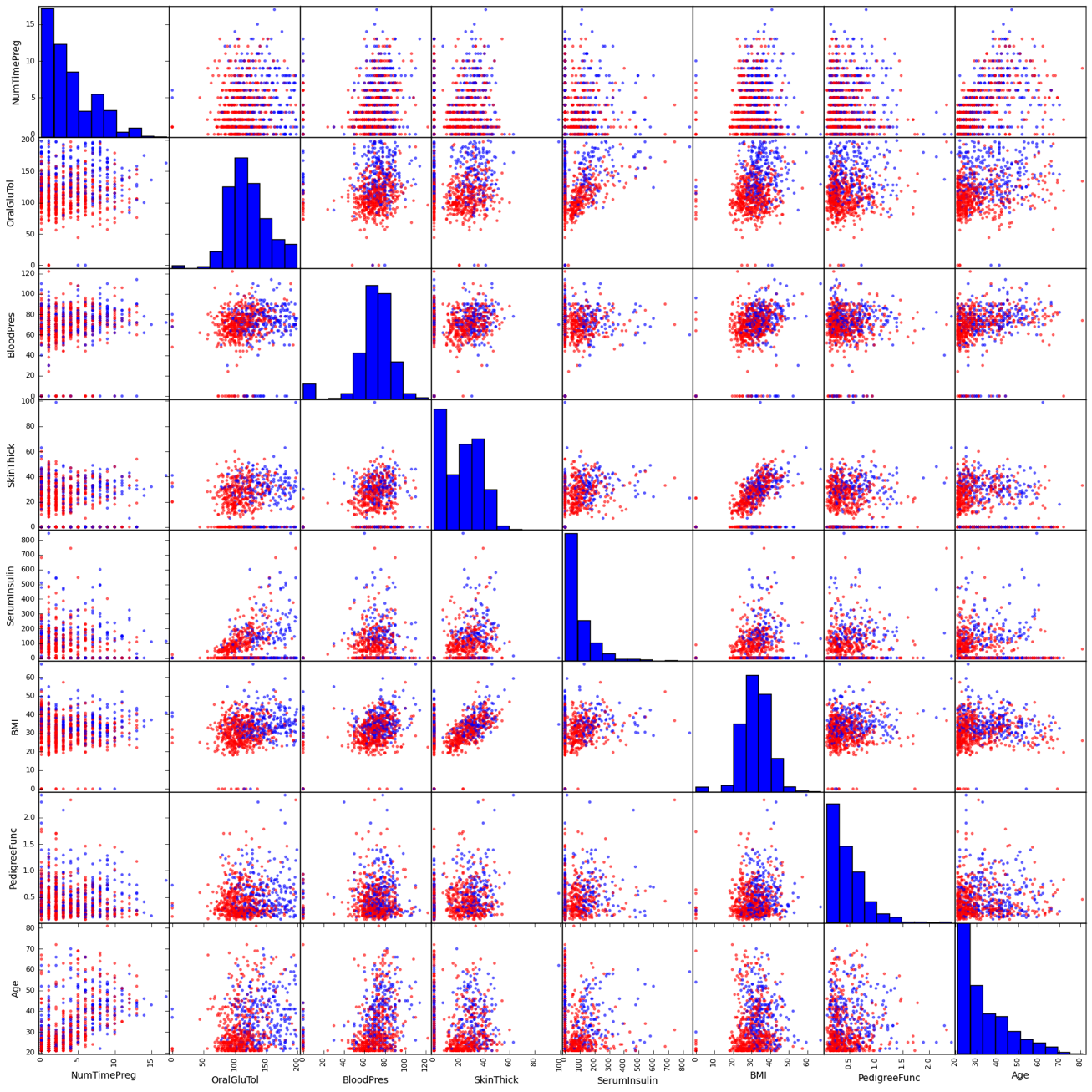


図1-1 糖尿病患者の散布図行列

これを見ると赤と青が少しだけずれているように思える。PCA解析によって分散を高くしてもあまり分かれなかった。さらに累積寄与度を見ても比較的寄与率の高い主成分の数と元々の成分の数が変わらないためあまり有益な情報は得られなさそうに思える。

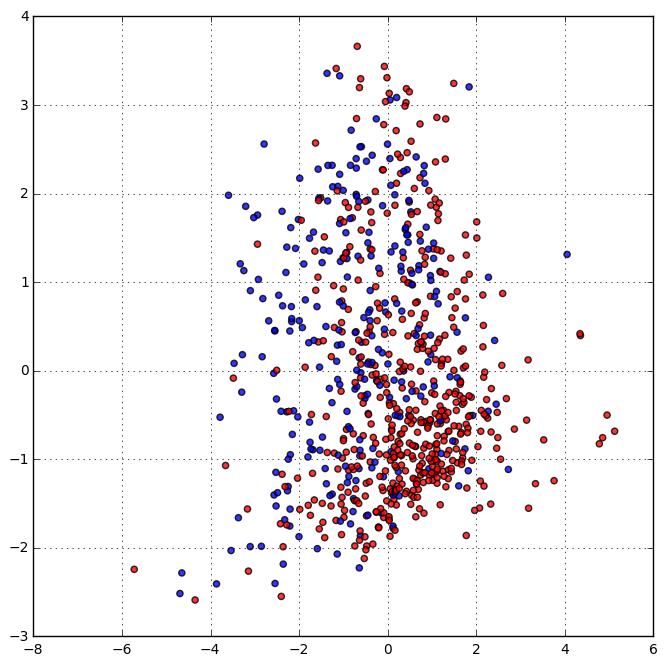


図1-2 PCA解析図(第一成分、第二成分)

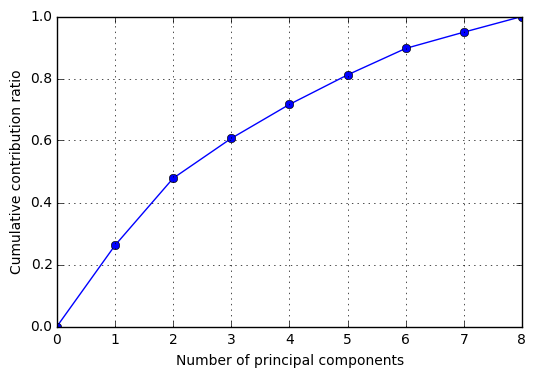


図1-3 累積寄与度の図

SVCによる解析を行うと以下の結果が得られた。

SVC(C=1, cache\_size=200, class\_weight=None, coef0=0.0, degree=3, gamma=0.0,

kernel='linear', max\_iter=-1, probability=False, random\_state=None,

shrinking=True, tol=0.001, verbose=False)

Accuracy\_score:0.7656967840735069

elapsed\_time:113.15214109420776

複数のパラメータで試そうとすると動かなくなるためパラメーターの数を絞った。さらに教師セットを大きくとると非常に時間がかかるため、教師セットを15%に設定した。精度が70%ほどである。

以下に予想と答えの数の表を載せる。

0 1

0 380 53

1. 106 114

表1-4 教師あり機械学習の正解率

糖尿病患者に警告できる確率は50%ほどであり、非糖尿病患者に対し糖尿病と誤診する確率が10%ほどある。今回のデータセットを用いて糖尿病患者を診断するのは難しいと考える。

課題2 パーキンソン病患者の判別

　とりあえず散布図行列を見て全体の傾向を掴む。図2-1

これを見ると赤と青が綺麗に分離できているところとできていないところが分かれている。

　続いてPCA解析を行い分離できているかどうかを確かめる。図2-2

これを見ると、赤と青が少しだけ分離できているように思う。これならば教師あり機械学習を行った時にパーキンソン病患者の判別ができることが予測される。

さらにPCA解析によって成分の数はどの程度減らせたのかを累積寄与度を見て確認する。図2-3

この図によると10個ほどの主成分で累積寄与度が1にかなり近いので、成分の数は半分ほどに減らすことができたと見ることができる。

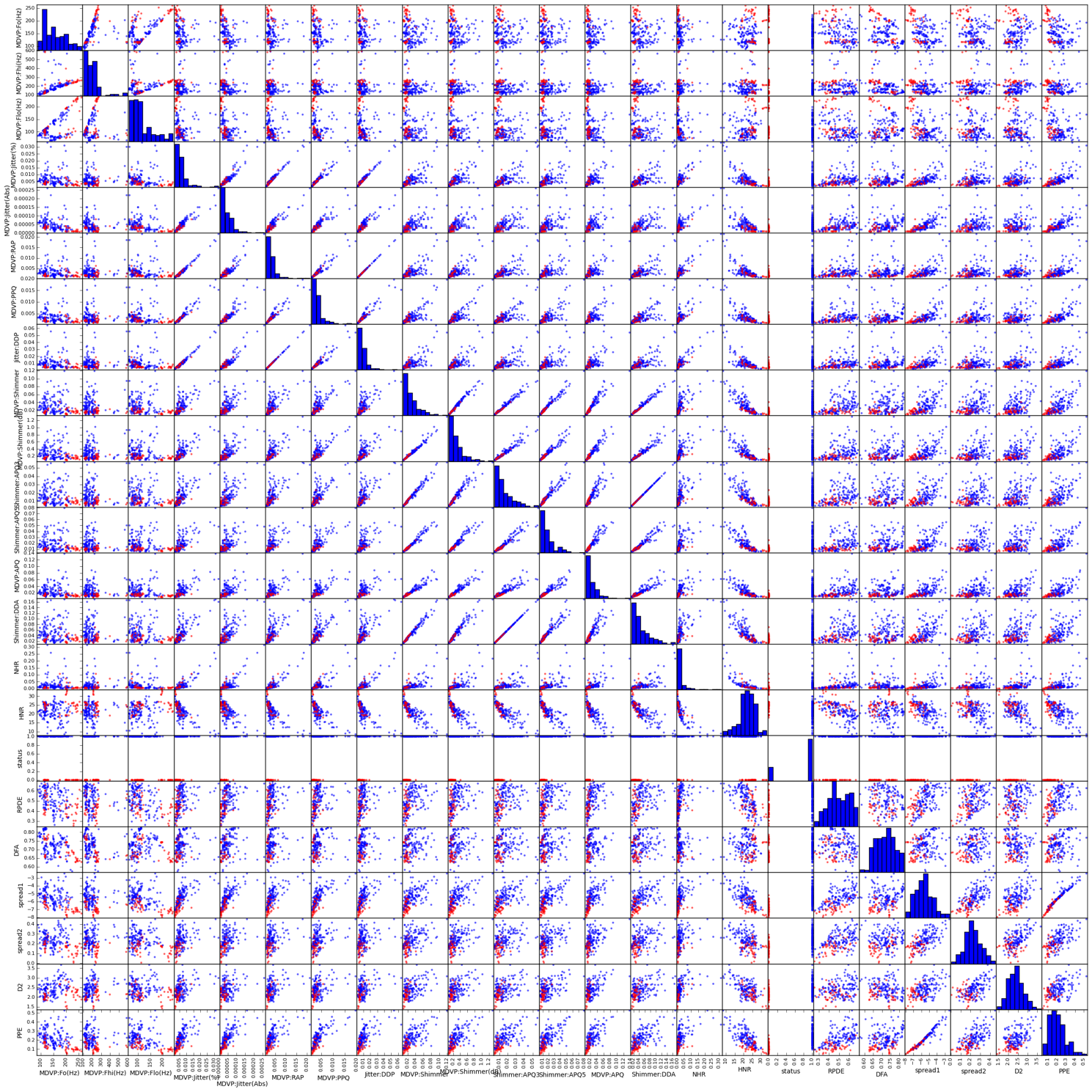


図2-1 パーキンソン病患者の散布図行列(赤が健康者)

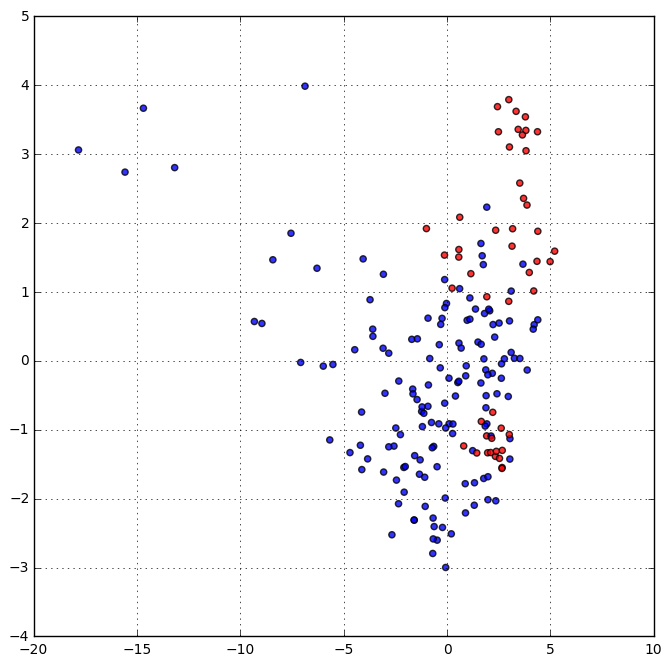


図2-2 PCA解析の第一成分と第二成分の表(赤が健康者)

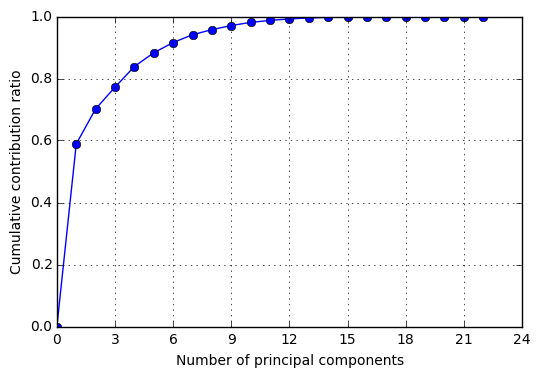


図2-3 累積寄与度の図

　最後に教師あり機械学習(SVC)を行う。

今回は半分のデータを教師として使い、3種類のモデルから最も良かったモデルを使う。結果が以下の通りになった。

SVC(C=1, cache\_size=200, class\_weight=None, coef0=0.0, degree=3, gamma=0.0,

kernel='linear', max\_iter=-1, probability=False, random\_state=None,

shrinking=True, tol=0.001, verbose=False)

Accuracy\_score:0.8673469387755102

elapsed\_time:5.634704828262329

これによると精度が86%で計算に5.6秒かかっていることがわかる。

最後にSVCがどう間違えたのかを確認するため予測データの数と正解数の対応表を作る。

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 予測 | |
|  |  | 0 | 1 |
| 正解 | 0 | 13 | 10 |
| 1 | 3 | 72 |

これを見ると、パーキンソン病を見落とす確率が4%ほどいる。さらに健康な人の50%をパーキンソン病であると誤診してしまう。よって今回のデータセットによるSVCではパーキンソン病の発見に使うことは可能であるが、これだけでパーキンソン病であると診断してはいけないというデータが出た。

感想

今回の実験ではとりあえずコピペして動かなかったらいろんなところを修正する方針で行いました。おそらく学生もそのようにすると思います。

考察に関してどの程度まできちんと考察すればいいのか分かりにくいのでそこを明記してほしい。