Parkinson's disease Mini-project

Steffen Lehmann

Parkinson's disease Mini-project

22 features

Status:

- (one) - Parkinson's
- (zero) - Healthy

A detection problem

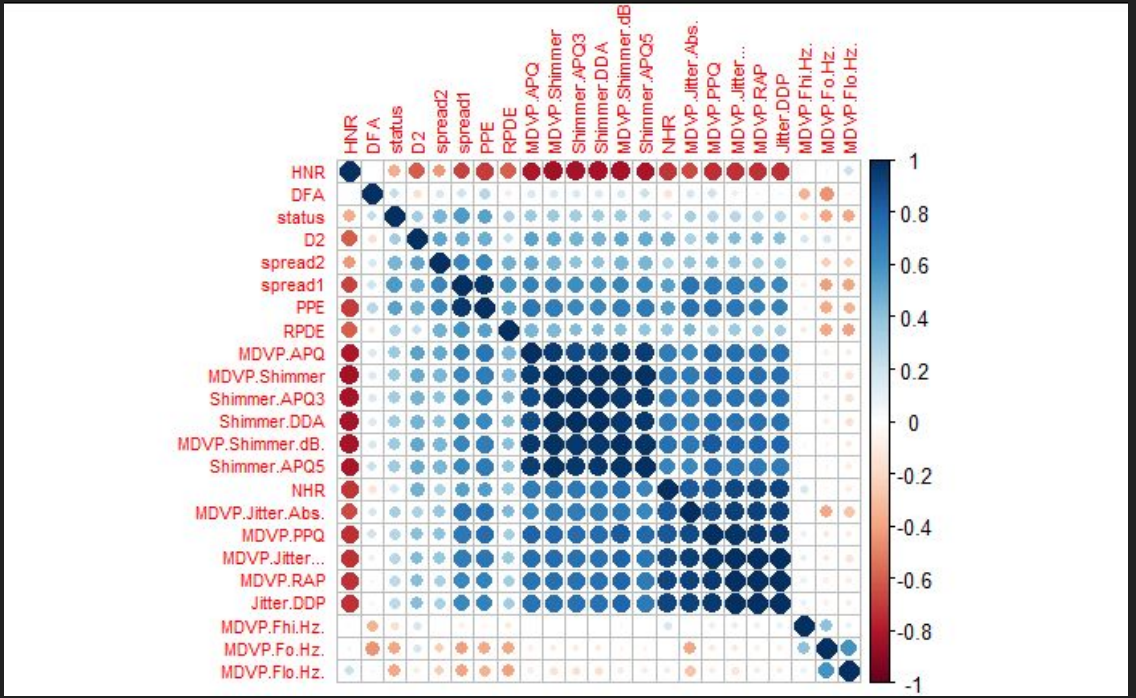
Feature selection - Correlation check

Removed labels + names

Removed > 0.9 correlated features

Top down approach

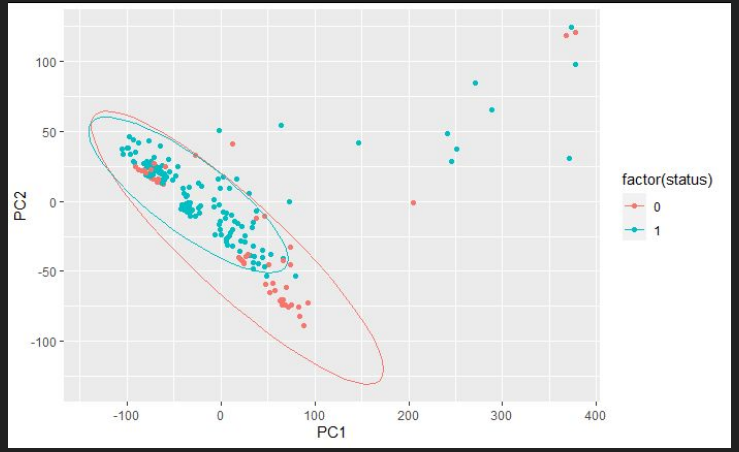
11 features

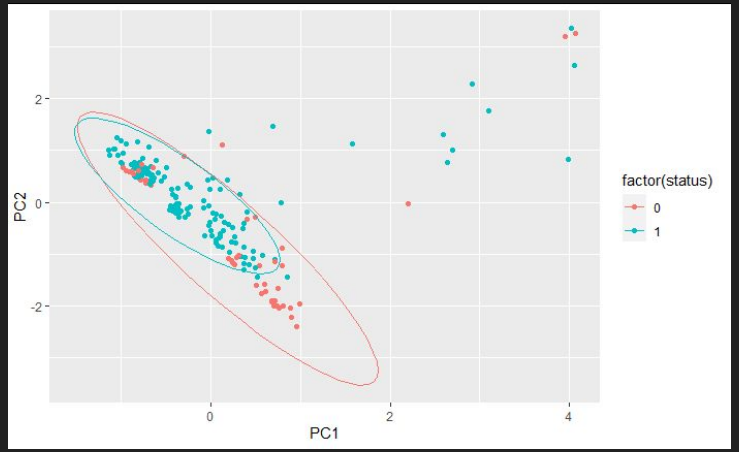


Standardizing the data

z = (x - u) / s

u = mean of the samples
s = SD of the samples



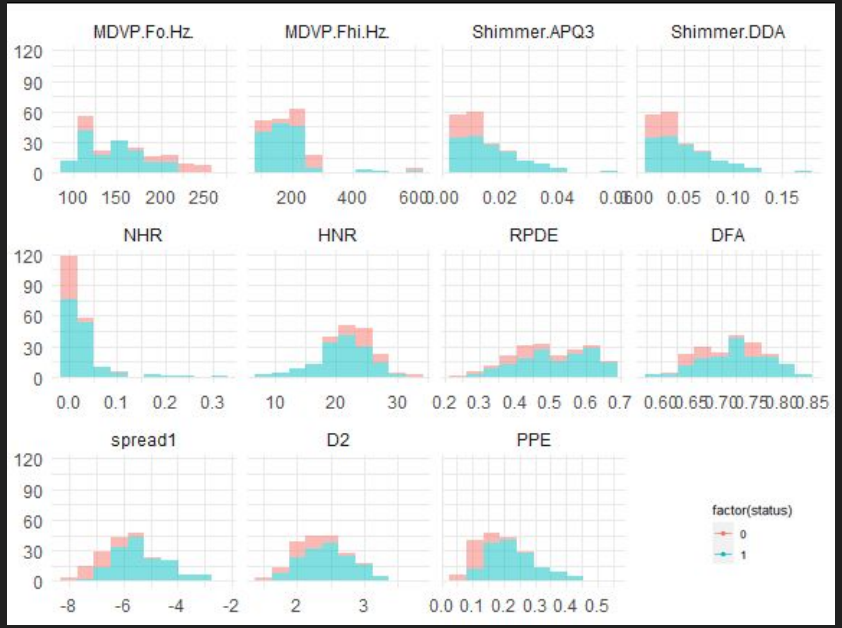


Normality check and distributions

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency has outliers

Shapiro-Wilk test of normality

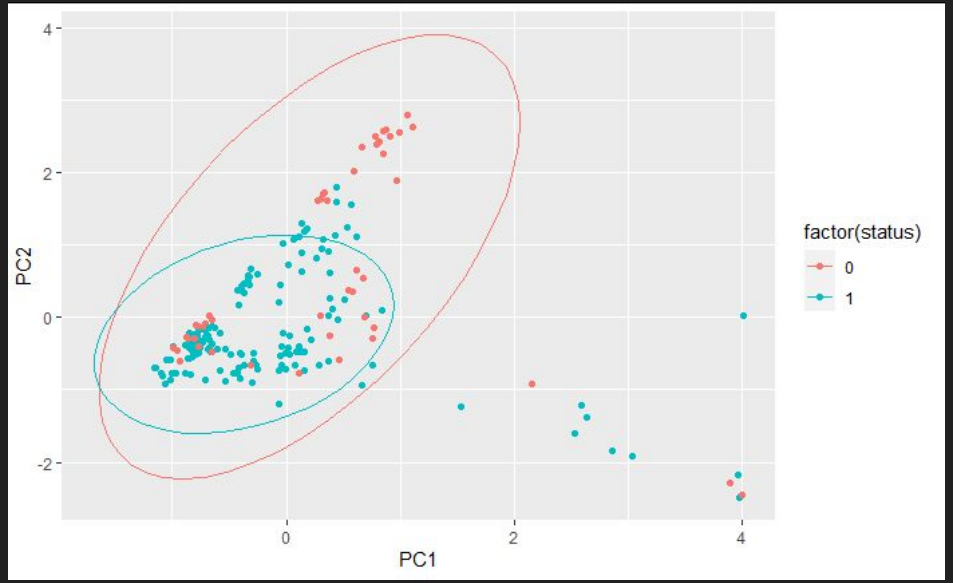
Means and Covariance



Dimensionality reduction: PCA - unsupervised

Before removing > 0.9 correlated features

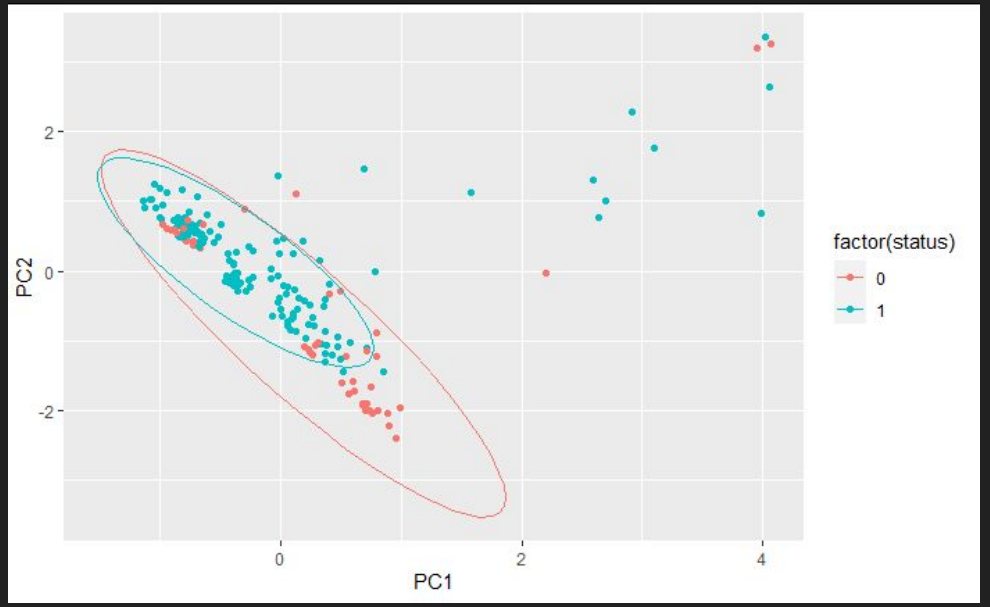
Features = 22



Dimensionality reduction: PCA - unsupervised

After removing > 0.9 correlated features

Features = 11



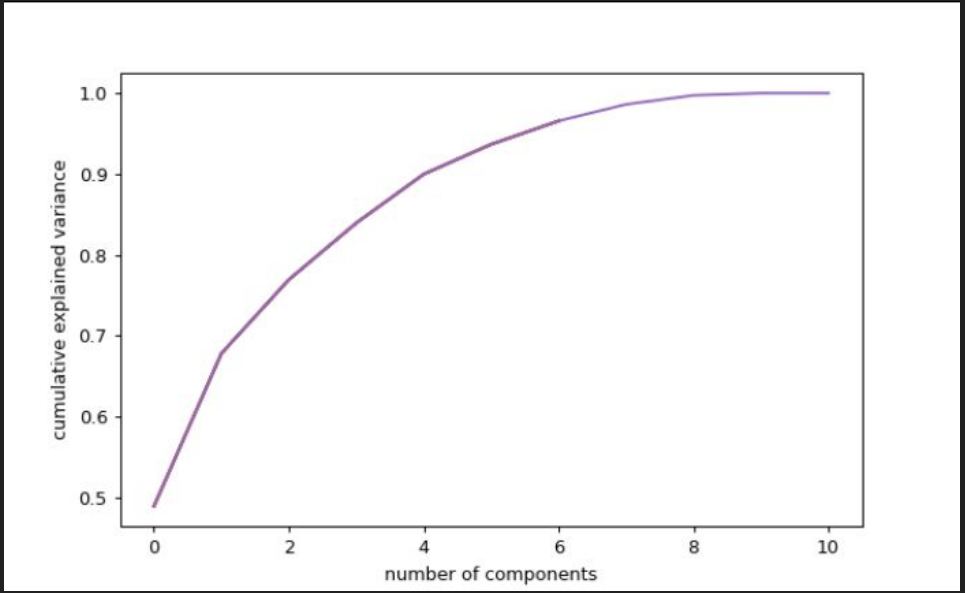
Dimensionality reduction: PCA - unsupervised

After removing > 0.9 correlated features

Features = 11

Retained 0.95 of the variance

7 PCs



Principal Component Analysis - unsupervised

- *N* is the number of scores in each set of data
- *X, Y* are the mean of the *N* in the each data set
- X_i, Y_i are the raw observation in each set
- $x_i y_i$ is the *i*th deviation score in each set

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{x})(Yi - \bar{y})$$

Covariance Matrix

$$Covariance\ Matrix = \begin{bmatrix} \text{Variance} & \text{cov(X, Y)} & \text{cov(X, Z)} \\ \text{cov(Y, X)} & \text{Variance} & \text{cov(Y, Z)} \\ \text{cov(Z, X)} & \text{cov(Z, Y)} & \text{Variance} \end{bmatrix}$$

Principal Component Analysis

Covariance Matrix

$$A \cdot v = \lambda \cdot v$$

$$\det(A - \lambda I) = 0$$

$$\begin{bmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{cov}(Z,Z) \end{bmatrix}$$

$$|A| = a \cdot \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \cdot \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \cdot \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

A = covariance matrix

v = eigenvector

λ = scaler (EigenValues)

I = identity matrix

Principal Component Analysis

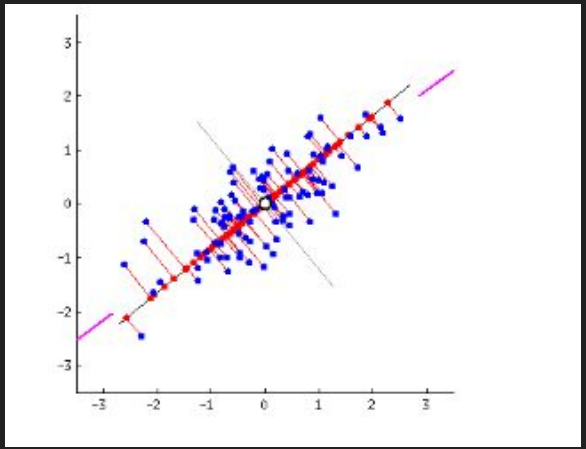
Covariance Matrix

EigenValues

EigenVector

$$A.v = \lambda.v$$

$$(A - \lambda.I).v = 0$$



Principal Component Analysis

The most important features in the covariance matrix

Biggest contributors to eigenvalue (scaler)

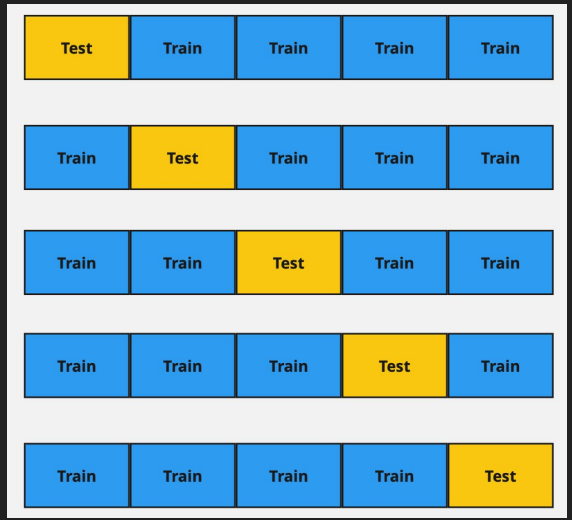
Detection Problem using supervised methods

K fold (interactions): 5

80/20 split

- Gaussian naive bayes
- Support vector Machine
- Random forest
- Neural network

On 11 and 7 features



Gaussian Naive Bayes

Priors

- 0.25 for Healthy
- 0.75 for Parkinson’s

Variance and mean

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
- P(D): the probability of the data (regardless of the hypothesis). This is known as the evidence.
- P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
- P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

If we assume that X's follow a Gaussian or normal distribution, we must substitute the probability density of the normal distribution and name it Gaussian Naïve Bayes. To compute this formula, you need the mean and variance of X.

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

In the above formulae, sigma and mu is the variance and mean of the continuous variable X computed for a given class c of Y.

Support Vector Machine Classifier - Supervised

Cn= Parkinson's and healthy

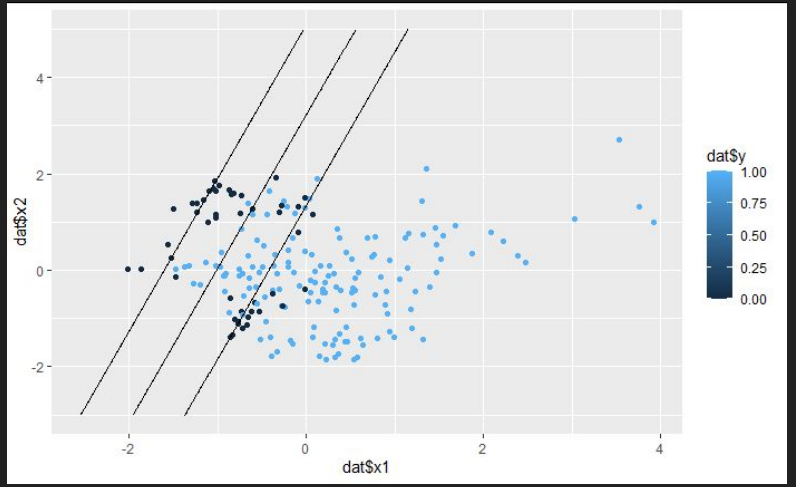
Not the actual model

$$g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} + b$$

$$\mathbf{w}^T\mathbf{z}_n + b \geq 1 \quad \text{if} \quad c_n = +1$$

$$\mathbf{w}^T\mathbf{z}_n + b \leq -1 \quad \text{if} \quad c_n = -1$$

for all n



Support Vector Machine Classifier - Supervised

Cn= Parkinson's and healthy

Not the actual model

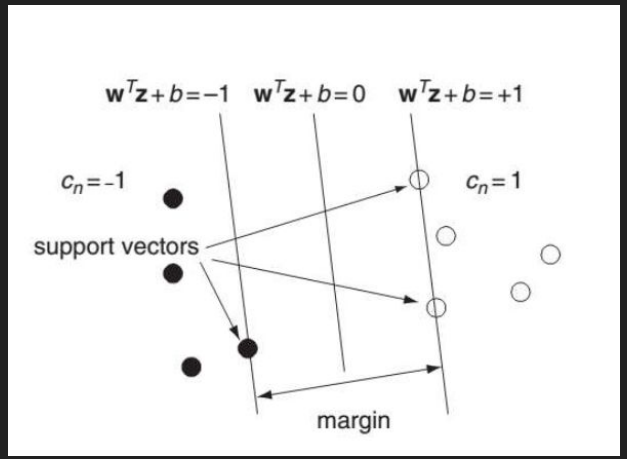
Margins

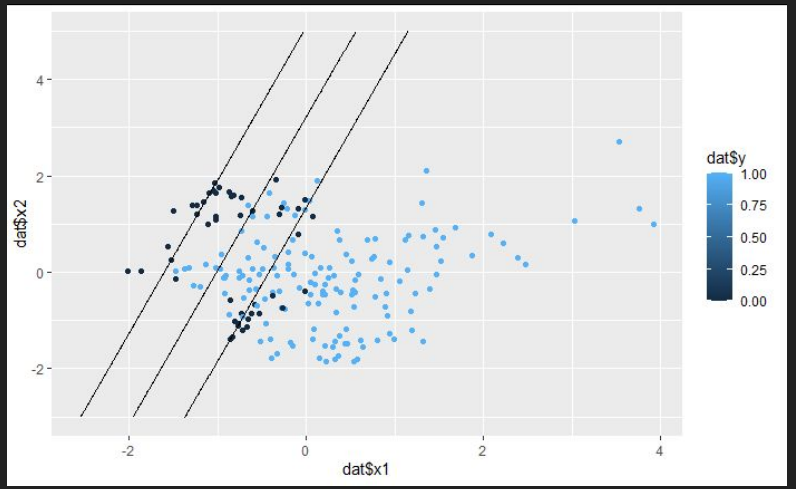
$$g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} + b$$

$$\mathbf{w}^T\mathbf{z}_n + b \geq 1 \quad \text{if} \quad c_n = +1$$

$$\mathbf{w}^T\mathbf{z}_n + b \leq -1 \quad \text{if} \quad c_n = -1$$

for all n



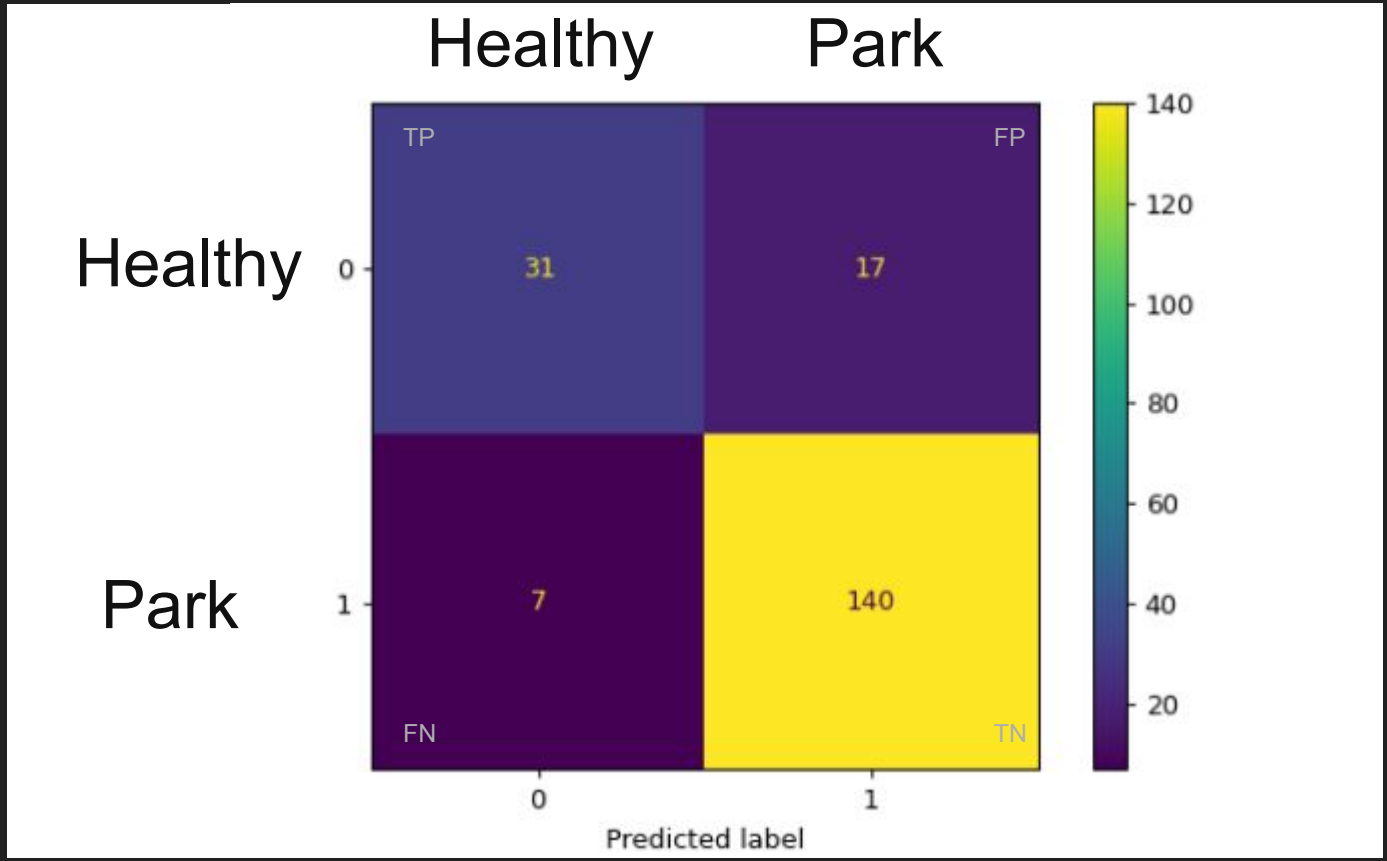


Was the sample Healthy or not?

The SVM model

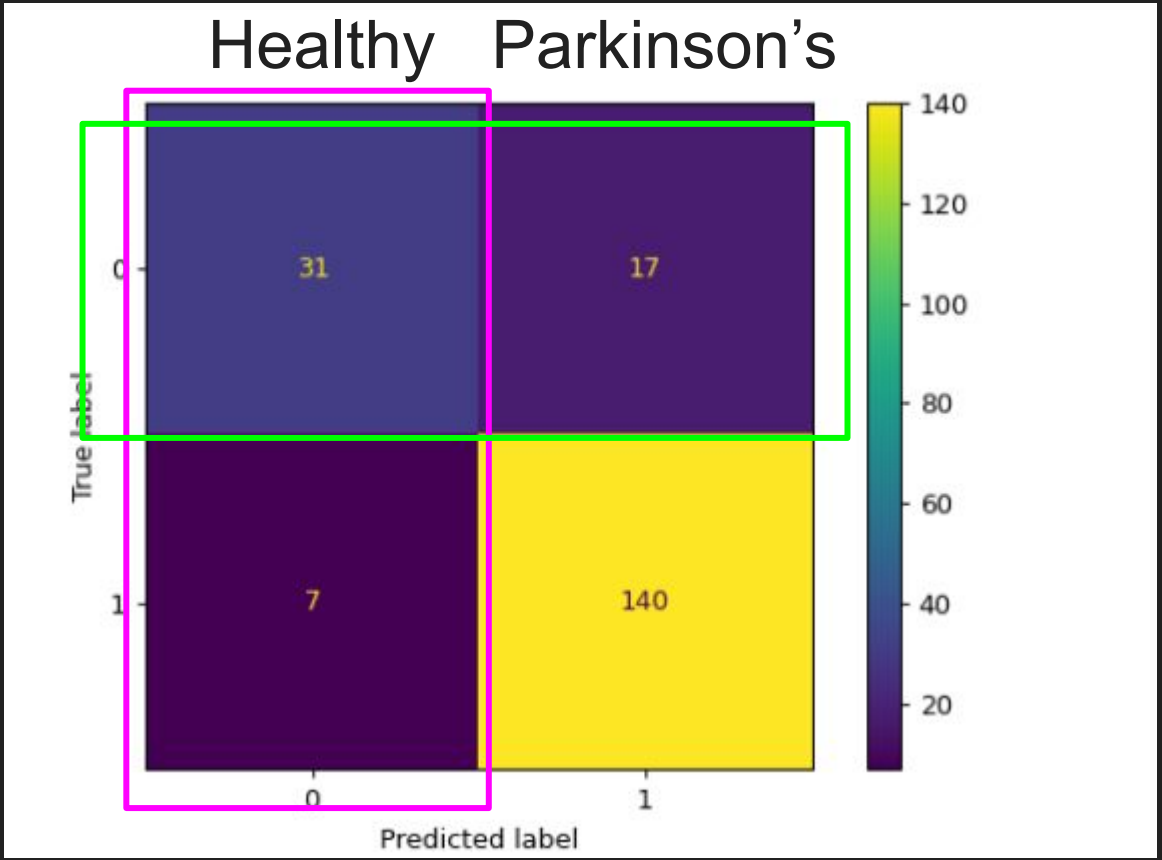
T threshold

FP better than FN?



Was the sample Healthy or not?

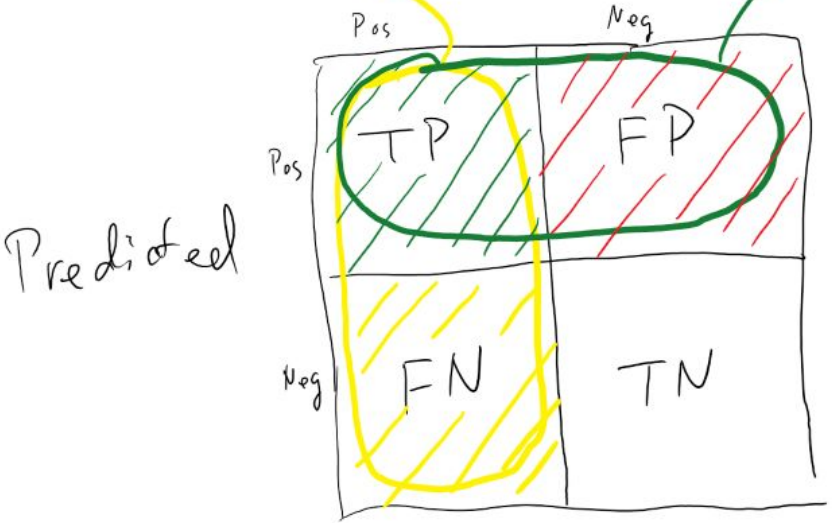
- Recall
- Precision
- Accurarry



$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$



Accuracy scores

Worse accuracy with 22 features

	SVC	RFC	NN	GNB
Standard (11)	**82% (0.06)**	79% (0.07)	79% (0.02)	74 % (0.05)
PCA (7)	80% (0.06)	**83% (0.08)**	**81% (0.04)**	**77% (0.03)**