EFME 2010 LU Exercise 4

Scientific Report

Gruppe 15:

Fritz Daniel - 0507049 - 935

Hiller Elias - 0525787 - 935

Sonderegger Josef - 0501625 - 935

Practical Approach

Das Datenset

Als Datenset wurde die *“Pima Indians Diabetes Database”* gewählt, welches acht biologische Features von 768 Frauen enthält. Es gilt aufgrund dieser Features herauszufinden, ob die Personen Diabetes haben oder nicht. Die Datenbank enthält auch eine neunte Spalte, welche bereits eine Klasseneinteilung beherbergt.

Die Features:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Dabei haben

- 500 Frauen kein Diabetes
- 268 Frauen Diabetes

Statistische Betrachtung & Auswahl der Features

Feature	Durchschnitt	Standardabweichung
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

Tabelle 1: Durchschnitts- und Standardabweichungswerte der einzelnen Features

	Diabetes		kein Diabetes	
Feature	**Durchschnitt**	**Standard-abweichung**	**Durchschnitt**	**Standardabweichung**
1	4.865	3.741	3.298	3.017
2	141.257	31.940	109.980	26.141
3	70.824	21.491	68.184	18.063
4	22.164	17.680	19.664	14.890
5	100.336	138.689	68.792	98.865
6	35.142	7.263	30.304	7.690
7	0.551	0.372	0.430	0.299
8	37.067	10.968	31.190	11.668

Tabelle 2: Durchschnitts- und Standardabweichungswerte der Features von Diabeteserkrankten im Gegensatz zu Nichterkrankten

Durch die Berechnung (***Tabelle 1***) und Gegenüberstellung der Werte in ***Tabelle 2*** kann abgelesen werden, welche Features sich zwischen Erkrankten und Nichterkrankten deutlicher unterscheiden. So kann der Blutdruck etwa auf den ersten Blick als Feature ausgeschlossen werden.

Feature	Unterschied Durchschnitt	Unterschied Standardabweichung
1	32.219 %	19.353 %
2	22.142 %	18.154 %
3	3.728 %	15.954 %
4	11.280 %	15.779 %
5	31.438 %	28.714 %
6	13.768 %	-5.878 %
7	21.938 %	19.677 %
8	15.855 %	-6.377 %

Tabelle 3: Unterschied des Durschnitts und der Standardabweichung in Prozent im Bezug von Erkrankten zu Nichterkrankten

Aus ***Tabelle 3*** kann direkt abgelesen werden, welche Attribute aussagekräftiger sind und welche nicht. Diese statistischen Verfahren ersparen aber keinen Blick auf die Datensätze, denn laut den hier dargestellten Werten wäre das Attribut 5 das wohl aussagekräftigste. Betrachtet man die Datensätze aber selbst, so fällt auf, dass sich sehr viele nichtssagende 0-Werte sowohl bei den Nicht- als auch Erkrankten befinden. Dasselbe trifft auch auf Attribut 4 zu, was darauf schließen lässt, dass diese Werte nicht für alle Personen erfasst wurden (eine Trizepshautfaltendicke von 0 mm ist nicht möglich und korreliert normal gut mit dem Body mass index = Feature 6. Normalwert ca. 20 mm *Quelle: http://www.kup.at/kup/pdf/752.pdf*)

Performance

Zuerst wurde die Erkennungsrate mit allen Features ermittelt. Dabei stellten sich Mahalanobis und das Backpropagation-Verfahren als effektivste Methoden zur korrekten Erkennung einer Diabeteserkrankung heraus. Auch der kNN Klassifizierer eignet sich noch recht gut und ist sogar besser, als die diagonale Mahalanobis Klassifizierung, welche mit ca. 70% im Bereich des Perceptrons mit 10 000 Zyklen liegt. Mit einer Erkennungsrate von unter 50% eignet sich das Perceptron mit 10 Zyklen wohl kaum für eine zufriedenstellende Klassifizierung. Ein solches Ergebnis wäre auch mittels Münzwerfens zu erreichen. Der Leistungsbezogene Vergleich ist in ***Abbildung 1*** dargestellt.

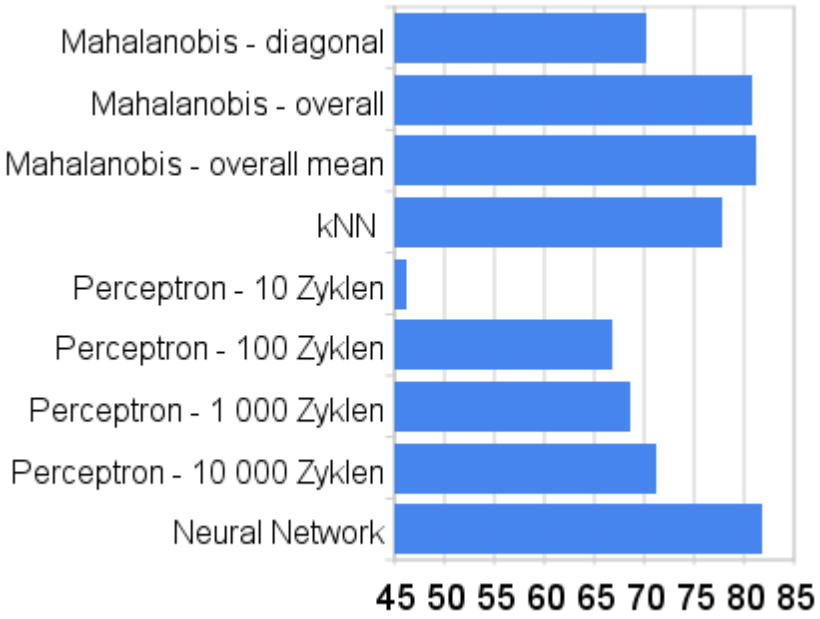

Abbildung 1: Klassifizierungsergebnisse mit allen Features

In ***Abbildung 2*** werden die Klassifikatoren mit Auswahl unterschiedlicher Features miteinander verglichen. Die Leistungsverteilung bleibt dabei, mit Ausnahme des Perceptrons mit 10 Zyklen, in etwa gleich wie bei der Wahl aller Features. Vor allem die Klassifikation ohne die Features 4 und 5 (Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml)), welche viele 0-Werte enthalten, stellt sich dabei als sehr erfolgreich heraus. Lässt man dabei das Feature 3 weg,

welches laut ***Tabelle 3*** eine geringe Aussagekraft besitzt, so lassen sich einige Ergebnisse nochmals verbessern (führt teilweise aber auch zu einer geringen Verschlechterung).
Zur besseren Übersichtlichkeit wurde in ***Abbildung 3*** die Performance mit allen Features und ohne die Features 4 und 5 nochmals gegenübergestellt. Die genauen Werte aller getesteten Verfahren und Features sind in ***Tabelle 4*** ersichtlich. In ***Abbildung 4*** wird dabei der Unterschied der Features beim kNN Klassifikator und in ***Abbildung 5*** der Unterschied beim Backpropagation nochmals dargestellt.

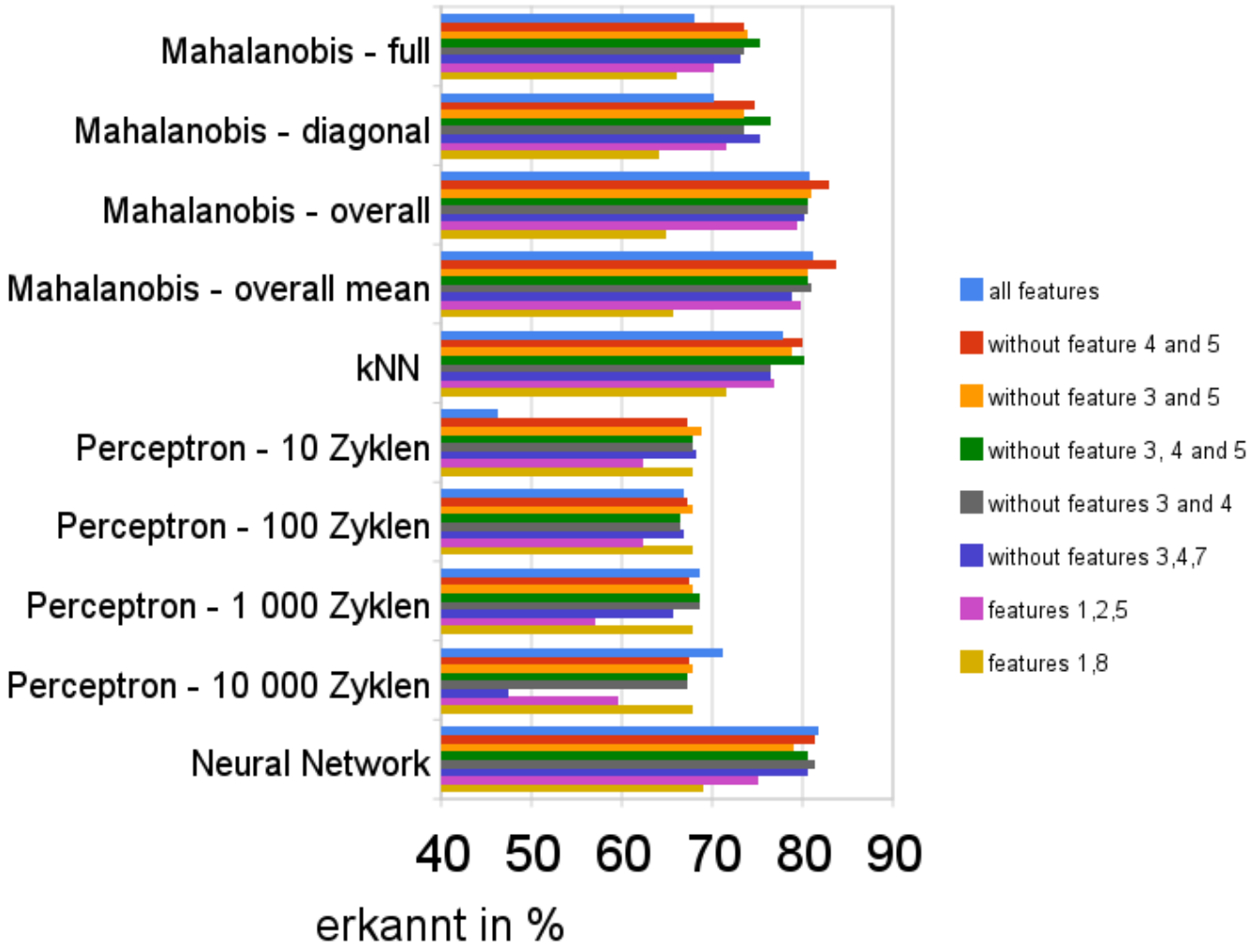

Abbildung 2: Erkennungsrate der unterschiedlichen Klassifizierungsmethoden

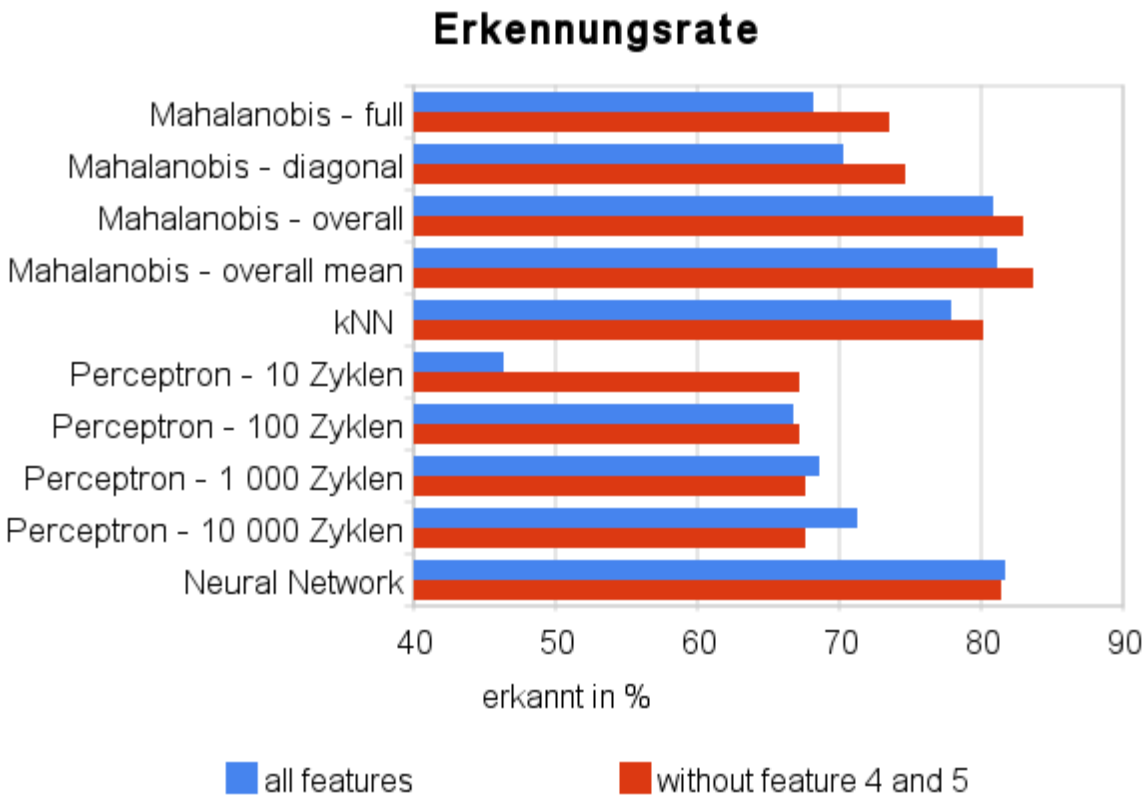

Abbildung 3: Erkennungsrate der unterschiedlichen Klassifizierungsmethoden mittels allen Features und ohne Features 4 und 5

	all features	without feature 4 and 5	without feature 3 and 5	without feature 3, 4 and 5	without features 3 and 4	without features 3,4,7	features 1,2,5	features 1,8
Mahalanobis - full	68.116	73.551	73.881	75.373	73.5075	73.1343	70.1493	66.0448
Mahalanobis - diagonal	70.29	74.638	73.508	76.492	73.5075	75.3731	71.6418	64.1791
Mahalanobis - overall	80.797	82.971	80.97	80.597	80.597	80.2239	79.4776	64.9254
Mahalanobis - overall mean	81.159	83.696	80.597	80.597	80.9701	78.7313	79.8507	65.6716
kNN	77.898	80.072	78.731	80.224	76.493	76.4925	76.8657	71.6418
Perceptron - 10 Zyklen	46.269	67.164	68.731	67.91	67.91	68.284	62.313	67.91
Perceptron - 100 Zyklen	66.791	67.164	67.91	66.418	66.418	66.791	62.313	67.91
Perceptron - 1 000 Zyklen	68.657	67.537	67.91	68.657	68.657	65.672	57.09	67.91
Perceptron - 10 000 Zyklen	71.269	67.537	67.91	67.164	67.164	47.388	59.701	67.91
Neural Network	81.716	81.343	79.104	80.597	81.343	80.597	75	69.03

Tabelle 4: Erkennungsrate in %

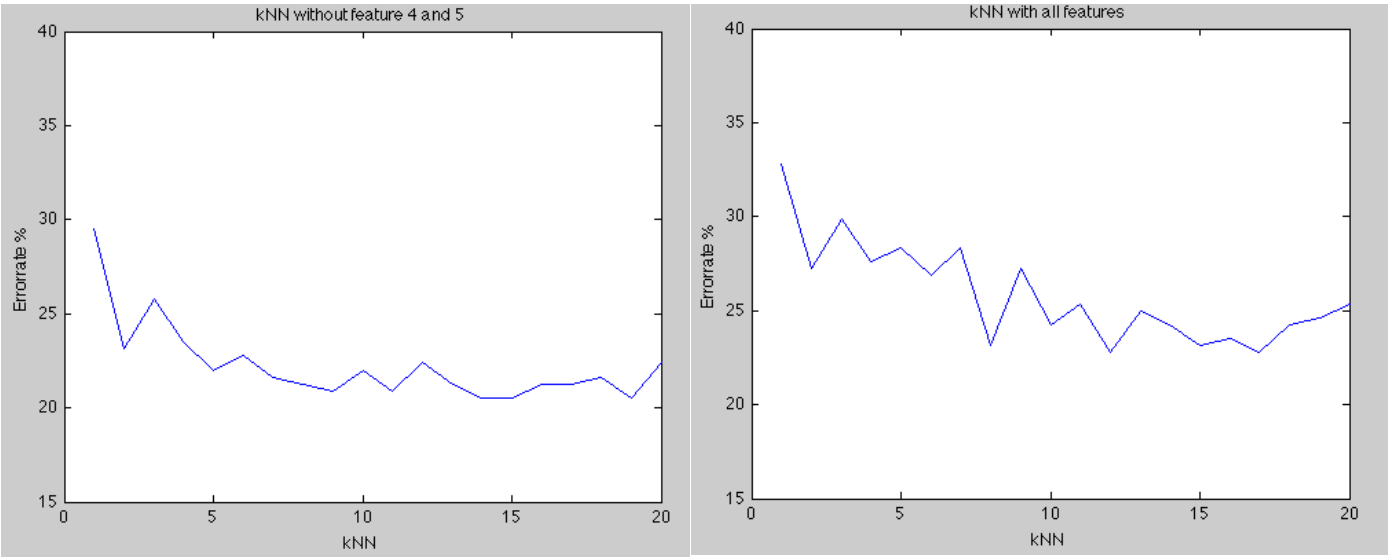

Abbildung 4: *links*: kNN mit allen Features *rechts:* kNN ohne Features 4 und 5

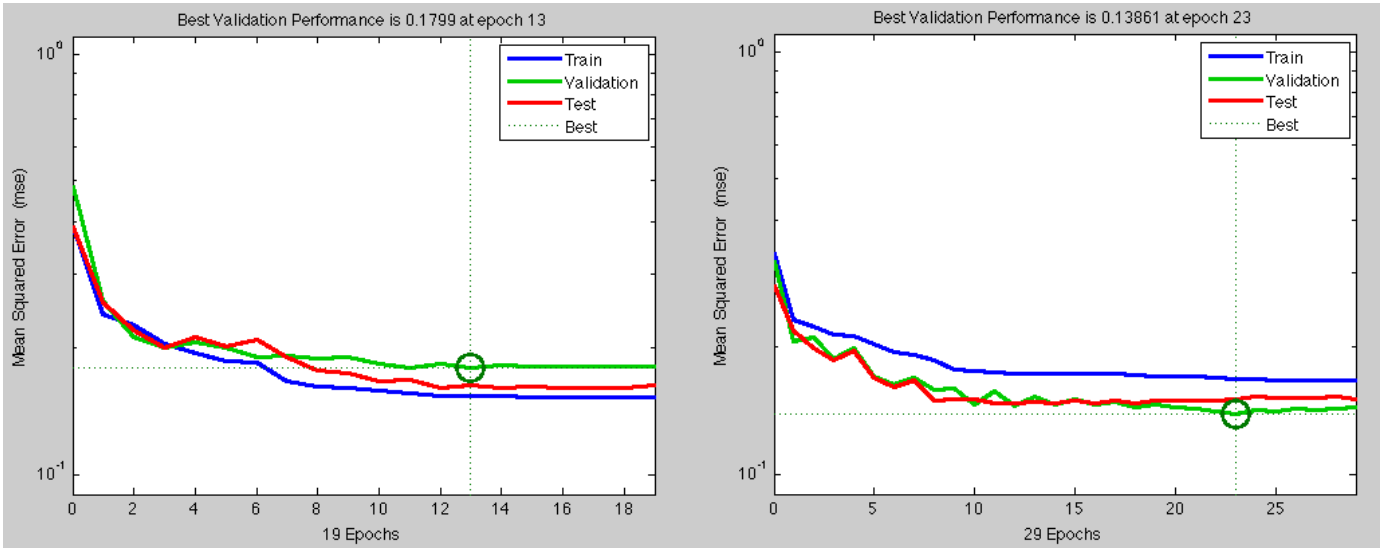

Abbildung 5: Performance von Backpropagation - *links:* alle Features *rechts: ohne Features 4 und 5*

Gründe für die Unterschiedlichen Ergebnisse der unterschiedlichen Klassifikationsmethoden

Die hier verglichenen Klassifizierungsarten unterscheiden sich zum Teil grundlegend und infolgedessen sind sie auch für unterschiedliche Anwendungen geeignet. Der kNN-Algorithmus ist ein nicht-parametrischen Verfahren. Es wird vorab keine Annahme über die Dichtefunktion getroffen, also gut anwendbar, wenn noch keine relevanten Daten vorhanden sind. Da Speicher- und Laufzeitaufwand mit der Größe des Trainingssets und der verwendeten *nearest neighbour* anwachsen lässt die Performance zu wünschen übrig.

Im Gegenzug dazu ist der Mahalanobis-Algorithmus durch die Berechnung der Kovarianz-Matrix bzw. der unterschiedlichen Matrizen sehr leistungsfähig. Dieses Verfahren ist parametrisch, da schon im Vorhinein durch μ und die Kovarianz-Matrix die Dichtefunktion

bekannt ist. Naheliegend ist, dass ein solches Verfahren nicht so flexibel eingesetzt werden kann und performanter ist.
Das Perceptron entspricht einer linearen Diskriminantenfunktion, welches bei entsprechender Datenverteilung sehr gute Ergebnisse liefert, sofern linear separierbar. Ein solcher Algorithmus ist sehr einfach und schnell zu trainieren und auszuwerten, weswegen das Perceptron oft als Probenklassifikator verwendet wird. Dies bestätigt sich auch bei unseren Ergebnissen, da die Erkennungsraten bei den durchgeführten Beispielen 5 bis 15% unter den Ergebnissen von kNN, Mahalanobis und Neural Network liegen.
Backpropagation eignet sich normalerweise sehr gut zum erlernen von künstlichen neuronalen Netzen. Wichtig dabei ist das überwachte Lernverfahren, welches nur möglich ist, wenn der Zielwert beim Trainingsverfahren bereits bekannt ist. Beim Lernverfahren wird dabei versucht die gegebenen Eingabevektoren möglichst genau auf die gegebenen Ausgabevektoren abzubilden. Hierzu wird eine Fehlerfunktion beschrieben, welche minimiert werden soll. In der Regel wird aber nur ein lokales Minimum erreicht. Das einlernen erfolgt durch Änderung der Gewichte. Durch das Erlernen ist das Netzwerk von keinem vorgegebenen Wert sondern lediglich von der Größe und Aussagekraft des Trainingssets abhängig, sodass wenn dieses gut gewählt wurde, die Erkennungsrate in der Regel sehr hoch ist.

Beeinflusst die Wahl weniger Features die Performance?
Die Performance im Sinne der Geschwindigkeit der Algorithmen wird durch die Anzahl der Features nur leicht beeinflusst und fällt durch die lange Rechenzeit der 10000 Zyklen z.B. beim Perceptron oder miteinbeziehen vieler Nachbarn bei kNN kaum ins Gewicht.
Bei der Performance im Sinne der Fehlererkennung beeinflusst die Anzahl der gewählten Features jedoch sehr wohl das Ergebnis. So ist der Algorithmus auch bei der Wahl von schlechten Features bei genügender Anzahl guter Features immer noch sehr effektiv. Werden hingegen wenige und schlechte Features gewählt, so ist ein deutlicher Leistungsabfall spürbar. Das Ergebnis kann aber durch die Auswahl mehrerer Features, wie etwa in ***Abbildung 3*** dargestellt, teilweise noch deutlich verbessert werden.

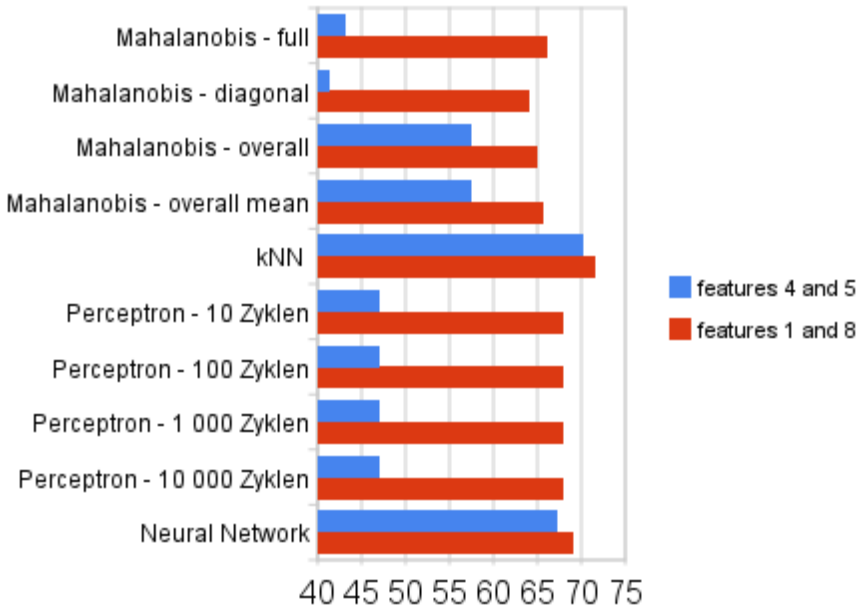

Abbildung 6: Vergleich von wenigen guten (1,8) und wenigen schlechten Features (4,5)

Was ist ein guter Weg um gute Features auszuwählen?
Um sinnvolle Features auszuwählen ist es sicher sinnvoll sich die jeweiligen Daten näher anzusehen und Auffälligkeiten auszumachen. Bei diesem Beispiel haben wir erkannt, dass sowohl in Spalte 4 (Trizepshautfaltendicke) und in Spalte 5 (2h Insulinserum) sehr viele Nullen enthalten sind, die das Ergebnis verfälschen können. Eine genauere Betrachtung des Datensets wurde bereits zu Beginn des Kapitels vorgenommen (siehe Datenset). Eine gute Darstellung dieses Umstands wird in ***Abbildung 2*** ersichtlich.

Was wir gelernt haben

Gibt es eine Einteilung in nur 2 Klassen, wie im oben gewählten Beispiel, so eigenen sich auch einfache Klassifikationsmethoden, wie etwa das kNN Verfahren, oder Mahalanobis sehr gut zur Klassifizierung. Bei nicht so deutlichen Gegebenheiten können Neuronale Netzwerke, wie etwa das Backpropagationsverfahren ihre Stärken ausspielen, wenn das Trainingsset groß genug ist.

Wie bereits angedeutet, spielt die Auswahl des Trainings und Testsets eine entscheidende Bedeutung. Je nach Klassifikationsmethode muss das Trainingsset groß genug gewählt werden, um zufriedenstellende Ergebnisse zu erhalten.
Bei Neuronalen Netzwerken kann die Lernkurve eine komplexe Form annehmen und auch schwierige Klassifizierungen meistern.

Bei der Auswahl geeigneter Features reicht es nicht nur aus oberflächliche statistische Vergleiche anzufertigen, sondern es muss auch ein Blick auf die Datensätze geworfen werden, um Unregelmäßigkeiten, wie etwa die vielen 0-Werte bei den Features 4 und 5 zu erkennen.