BIOMEDICAL DATA SET

Introduction

Over the last decade, we have seen an explosion of machine learning research and applications. One application of machine learning techniques includes analysis of medical data – in particular, medical diagnosis for specialised diagnostic problems. Through application of statistical methods to data, we can develop screening procedures which can accurately diagnose patients for a rare genetic disorder.

In this report we aim to develop such a screening procedure to detect carriers more accurately from four blood sample measurements given that the current industry standard is to use m1 (only). To achieve this aim we will use exploratory analysis tools, linear discriminant analysis and quadratic discriminant analysis as classifier algorithms and Naïve Bayes on data across two files – one for healthy patients and one for carriers of the disease.

Data and Methods

The Data

The data is split across two data files – one for disease carriers (*carriers.txt*) and the other for healthy candidates (*normals.txt*) containing 67 and 127 samples, respectively. Both data files contain the age of the patient, date that blood sample was taken (*mmddyy*) and measurements *m1*, *m2*, *m3* and *m4*. Before carrying out analysis on this data, we combined the two data files into one data frame, *alldata,* and added an additional column, labelled *diagnosis*, to be able to differentiate between disease carriers and healthy candidates. All entries in the date column within our dataset provided no information on the day of the month the blood sample were taken, with all dd entries being set to '00'. To ensure that the column could be meaningfully interpreted in further analysis and improve readability we changed the day of each entry to '15' and converted the column to number of days since 31-12-1977.

The variables are summarised in the table below.

Variable	Description
age	age of patient in years
date	number of days blood sample was taken since 31-12-1977
m1	m1 reading in blood sample
m2	m2 reading in blood sample
m3	m3 reading in blood sample
m4	m4 reading in blood sample
diagnosis	diagnosis of patient for rare genetic disorder

Methods

Exploratory Data Analysis (EDA) – To begin our analysis we used EDA to maximise our insight into a data set and look into the underlying structure of this data. We used a tool known as principal component analysis (PCA), an unsupervised method helping us better understand the variables in the data set. PCA reduces the dimensionality of a dataset whilst preserving as much variability as possible allowing us to visualise which variables are more important than others and which are correlated with one another. This is particularly useful, as we can observe which of our variables have the greatest influence on our components which capture the greatest amount of variance in our dataset. From this we can begin to understand which measurements may be better at identifying disease carriers from healthy patients.

Discriminant Analysis – Our aim is to develop a screening method which accurately differentiates between disease carriers and healthy individuals using our variables. We will be exploring whether these two groups are different and on which measurements are these groups most different. We would also like to predict which group a patient belongs to using our measurement variables. Discriminant analysis is helpful here in making diagnosis predictions using our variables. In our investigation we will carry out both Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Both make the assumption that the predictor variables are drawn from a multivariate Gaussian distribution. LDA also assumes a covariance matrix common to all classes in the data set, unlike QDA which assumes that each class has its own covariance matrix. With large training data, we typically find that QDA is preferred as LDA can overfit the data.

Naïve Bayes Classifier – Another classifier we will use is Naïve Bayes which is based on Bayes theorem of probability to predict the class of our data set. It assumes independence among predictors i.e., presence of a particular feature in a class is unrelated to the presence of any other feature. Despite this over-simplifying assumption, Naïve Bayes classifiers have worked well in many real-world situations and only requires a small amount of training data to estimate necessary parameters.

Comparing classifiers – Given our limited sample of data consisting of 194 entries, we have assigned 70% of the data to the training set and 30% of the data to the test set – we have done this to assess the accuracy of our screening procedure across all classifiers and their ability to work on new data. These classifiers are then compared against each other using ROC curves which visually measure each classifier's accuracy. The accuracy is measured by the area under the ROC curve, where an area of 1 represents a perfect classifier and an area of 0.5 represents a worthless classifier.

Results and Discussion

After performing principal component analysis on our combined dataset, with and without scaling, we observe some separation in the scores plots (Fig 1 and Fig 2 respectively) between the two groups, disease carriers and healthy patients. Scaling the data allows for small peaks in the measurement data, which could be important for any differences between the groups, to have as much influence in the analysis as the larger peaks. We observe that separation appears to be better between positively and negatively diagnosed patients when the data is scaled (Fig 2).

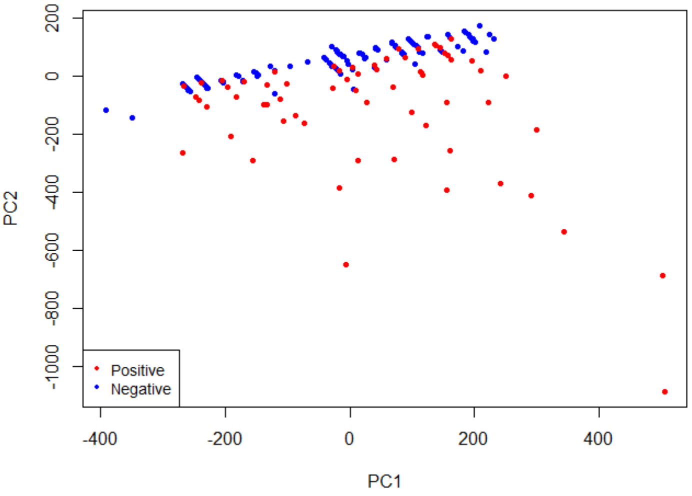

Fig 1: Unscaled scores plot for PC1 against PC2

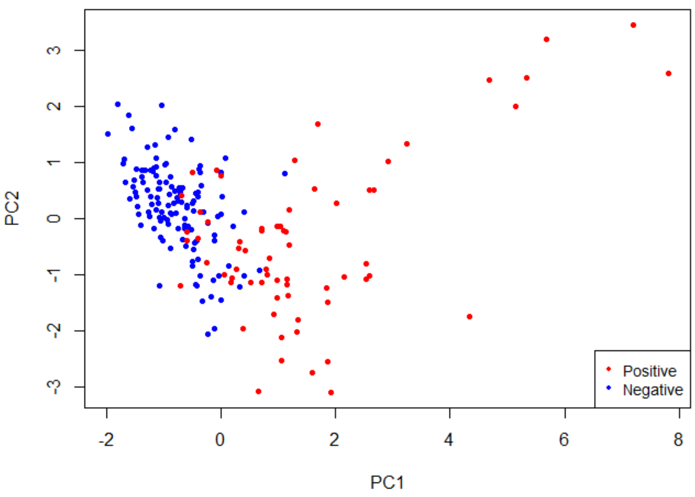

Fig 2: Scaled scores plot for PC1 against PC2

The biplot for the unscaled data (Fig 3) shows that *m1* and *date* dominate the analysis which is typical for variables with large values. After scaling the data to unit variance, the biplot (Fig 4) shows that the analysis is dominated more equally by all our four measurements. We see that *m1, m2 and m3* appear contribute most to PC1 (which accounts for 40.4% of the variation in the data) and *m2* appears to contribute most to PC2 (which accounts for 20.0% of the variation in the data).

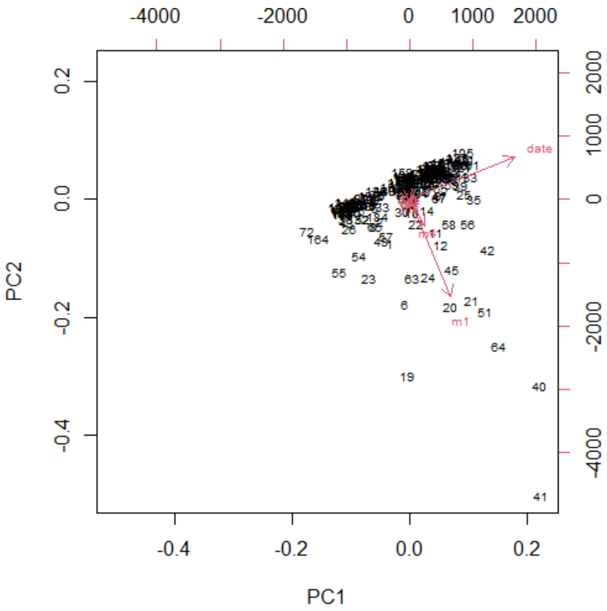

Fig 3: Unscaled biplot of PC1 against PC2

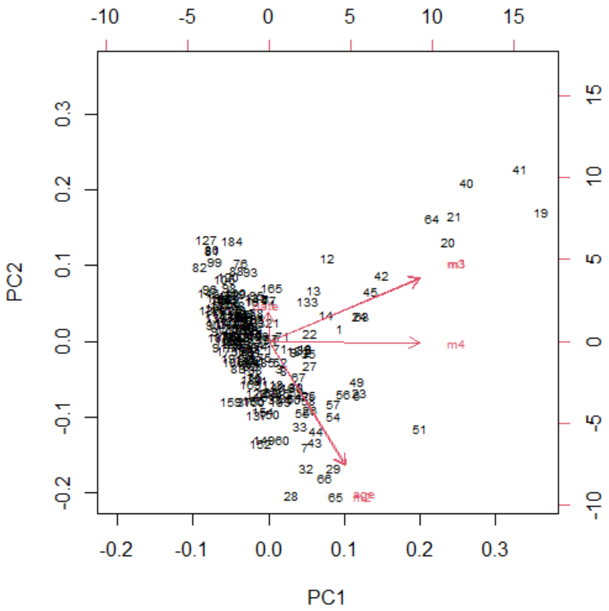

Fig 4: Scaled biplot of PC1 against PC2

Experts in the field have noted that young people tend to have higher measurements. We examine this belief by plotting the age of the patient against each measurement (Fig 5) and see that it appears this belief is incorrect in our dataset. In fact, the data shows that the opposite is true – there is a positive correlation between age and measurement value meaning that older people tend to have higher measurements.

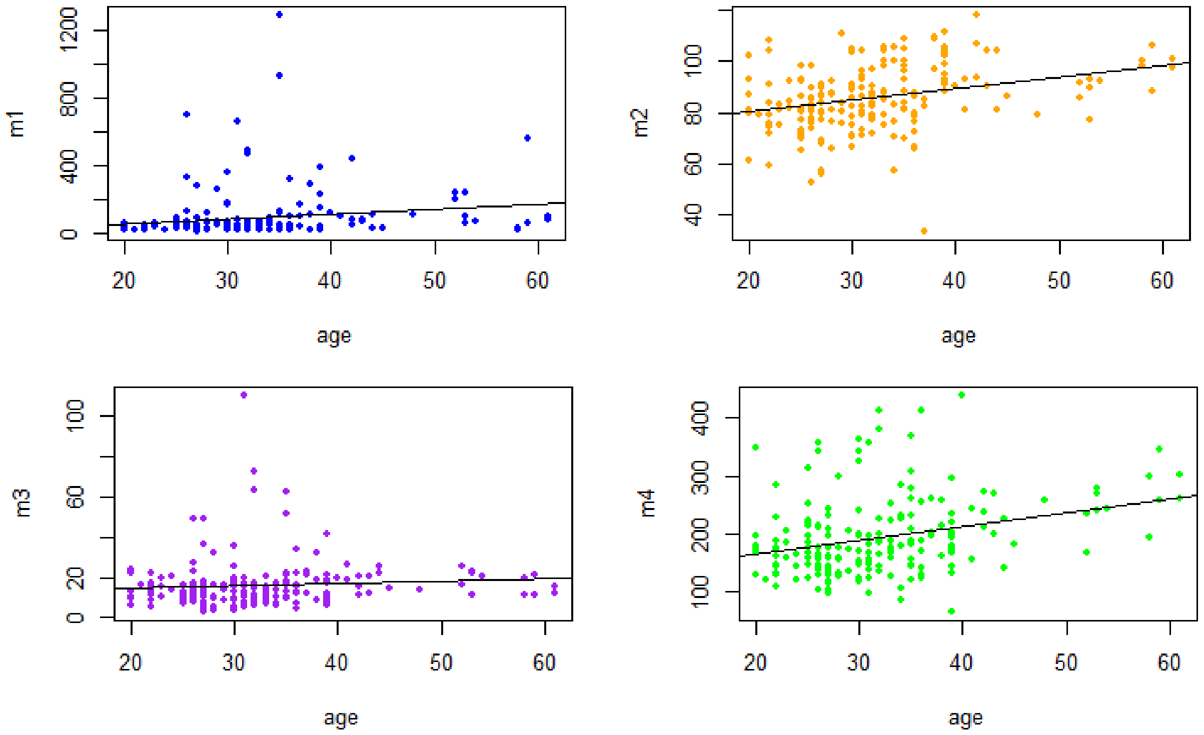

Fig 5: Relationship between age of patient and each measurement

Despite contradicting the common belief held amongst experts that young people tend to have higher measurements, there may still be some truth to this belief. Closer examination of each graph in Fig 5 shows that the high measurement values do come from the younger patients (as well as the low measurement values). Ultimately, it is difficult to conclude whether or not younger people tend to have higher measurements because our data is limited to 197 observations and from this limited sample, the majority of observations appear to be from the younger demographic.

We also examined the possibility for a systematic drift over time in the measurement process by plotting the number of days after 31/12/1997 the blood sample was taken against each measurement (Fig 6). There does not appear to be a single clear correlation between the number of days and the measurement values suggesting there is no systematics drift in the measurement process.

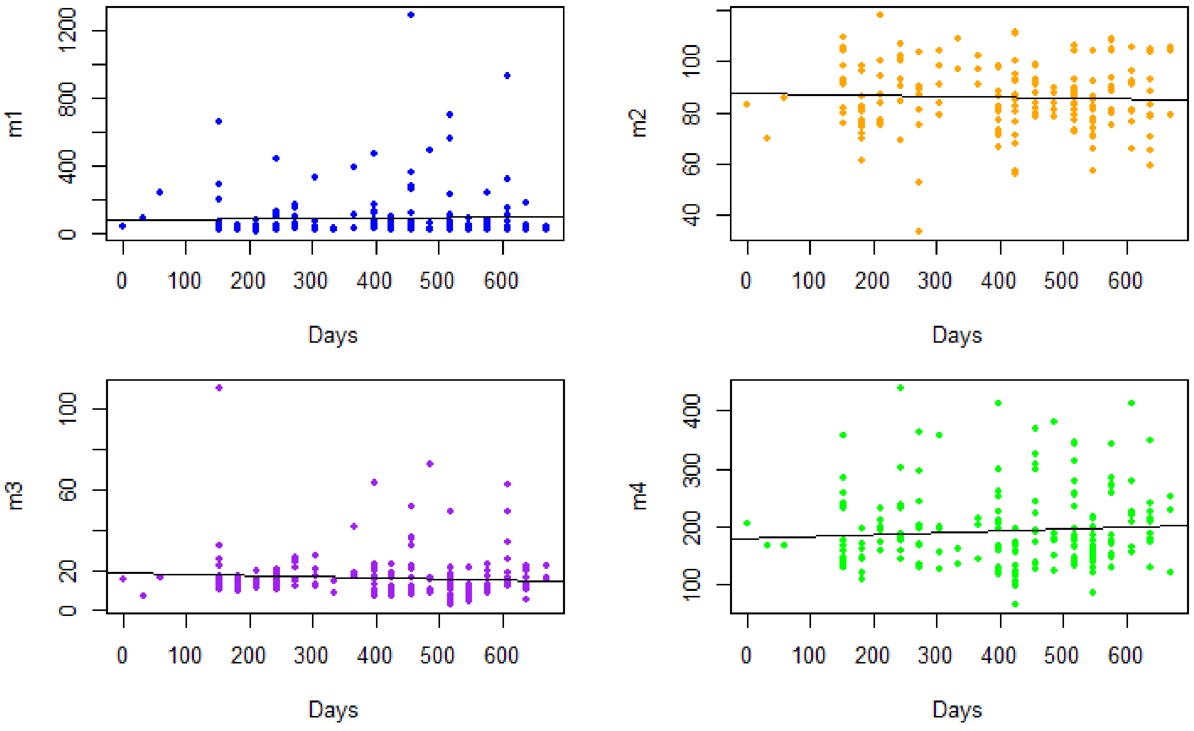

Fig 6: Relationship between number of days after 31/12/1977 blood sample was taken and each measurement

We then tested our three classifiers on randomly selected training data and test data which comprised 70% and 30% of total data, respectively. The results obtained for LDA, QDA and Naïve Bayes can be seen below in Tables 1-6.

		Real	
		Positive	Negative
Predicted	Positive	87	12
	Negative	2	35

Table 1: Confusion matrix obtained from LDA classification on the training set

		Real	
		Positive	Negative
Predicted	Positive	37	5
	Negative	1	15

Table 2: Confusion matrix obtained from LDA classification on the test set

Tables 1 and 2 show us how the LDA classifer performed on both our training and test set. We see that the accuracy of the LDA classifer was 90.0% on the training set and 90.0% on the test set.

		Real	
		Positive	Negative
Predicted	Positive	87	8
	Negative	2	39

Table 3: Confusion matrix obtained from Naïve Bayes classification on the training set

		Real	
		Positive	Negative
Predicted	Positive	37	4
	Negative	1	16

Table 4: Confusion matrix obtained from Naïve Bayes classification on the test set

Tables 3 and 4 show us how the Naïve Bayes classifer performed on both our training and test set. We see that the accuracy of the Naïve Bayes classifer was 92.6% on the training set and 91.4% on the test set.

		Real	
		Positive	Negative
Predicted	Positive	88	11
	Negative	1	36

Table 5: Confusion matrix obtained from QDA classification on the training set

		Real	
		Positive	Negative
Predicted	Positive	37	4
	Negative	1	16

Table 6: Confusion matrix obtained from QDA classification on the test set

Tables 5 and 6 show us how the QDA classifer performed on both our training and test set. We see that the accuracy of the QDA classifer was 91.2% on the training set and 91.4% on the test set.

Ultimately all classifiers performed well in being able to predict the patient's diagnosis, however Naïve Bayes had the highest accuracy across both the training and test sets and therefore appears to be the optimal screening procedure. We investigate this further by comparing ROC curves for each classifier against the independent training and test data.

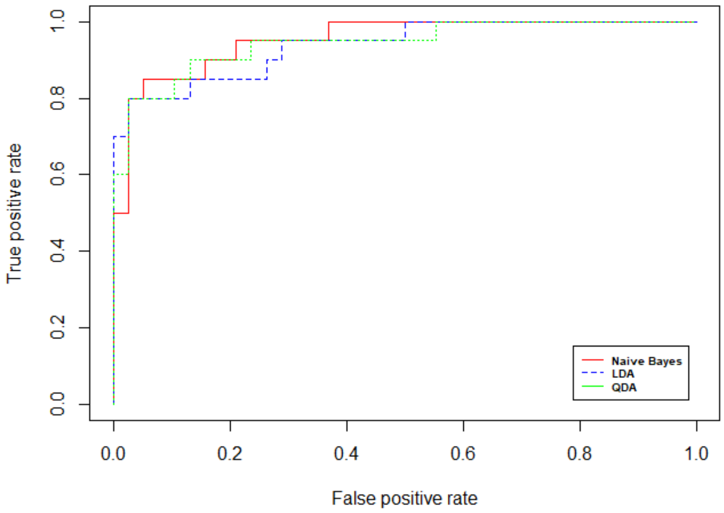

Fig 7: ROC curve comparison for Naïve Bayes, LDA and QDA classification on test data

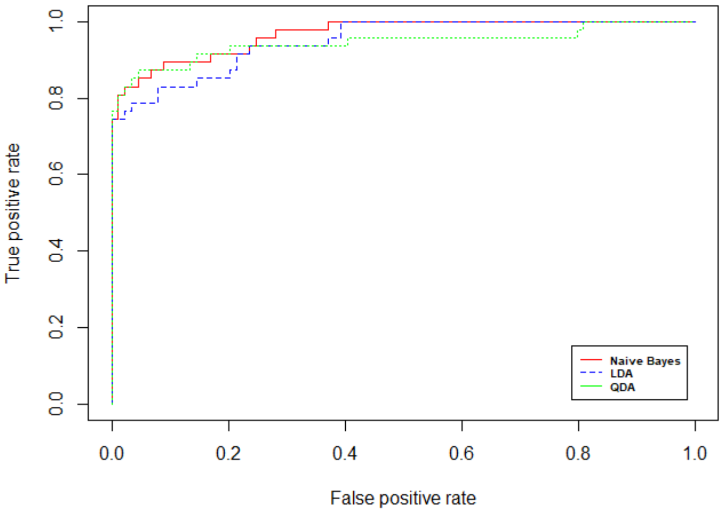

Fig 7: ROC curve comparison for Naïve Bayes, LDA and QDA classification on training data

The area under the ROC curve for Naïve Bayes, LDA and QDA on test data was 0.89, 0.86 and 0.87, respectively. The area under the ROC curve for Naïve Bayes, LDA and QDA on training data was 0.90, 0.86 and 0.88, respectively. We conclude that Naïve Bayes is the optimal screening method against LDA and QDA in that it should minimise the error i.e., incorrect diagnosis.

The screening methods we have used so far however use all the data we have had available to us. We repeated the above screening methods using different combinations of measurements to determine if there were any measurements better than others and whether the measurements should be combined.

Combination of measurements	**Training set accuracy (%)**	**Test set accuracy (%)**
m1, m2, m3, m4	92.6	91.4
m1	89.7	86.2
m2	80.9	70.7
m3	88.2	84.5
m4	89.0	84.4
m1, m2	91.2	86.2
m1, m3	90.4	86.2
m1, m4	91.2	88.0
m2, m3	89.7	82.8
m2, m4	87.5	86.2
m3, m4	90.4	87.9

m1, m2, m3	92.6	86.2
m2, m3, m4	91.2	89.7

We found that using all measurements yields the greatest accuracy in determining the diagnosis for a patient on both the training and test set. Combination m2, m3 and m4 followed with an accuracy rate of 89.7% on the test set. Interestingly, when measurements were used in isolation, m1 performed the best in correctly diagnosing patients in the test set.

Conclusions

The analysis shows that the best screening method for diagnosing patients appears to be Naïve Bayes classification using all data at hand. The current industry standard of using m1 (only) to identify carriers of the disease still relatively performed well and produced more accurate results on our test set than any other standalone measurement. The difference between using m1 alone and all four measurements was 5.2% - depending on the seriousness of this rare genetic disorder, this increase in accuracy may provide a very strong case for combining measurements. Assuming that not all measurements are equally obtained (in terms of cost and accuracy), we may argue that the industry standard m1 measurement should continue being used alone to diagnose patients given its better accuracy rate relative to all other standalone measurements. However, this is dependent on price and ease of obtaining, meaning that if m1 is relatively expensive and difficult to obtain, measurements m3 and m4 are suitable alternatives that provide similar accuracy rates.

We did not find evidence to support the claim that young people tend to have higher measurements (in fact we saw the opposite trend), nor did we find evidence of a systematic drift over time in the measurement process. However, its worth noting that our sample size is limited, and we have less data points for older patients within this sample.