Chua Feng Qing: EDA, ML
Derek Chang: EDA, ML
Justin Kwek: Dataprep, ML
Koong Kwang Hwee: Backgrd, EDA

1 Introduction

1.1 Background

Diabetes is a disease due to the body's inability to regulate blood sugar (glucose). The WHO estimates that 1.5 million deaths were directly caused by diabetes in 2019 [1]. In the U.S, diabetes is most common among Native Americans [2]. Thus, we analyse medical records of Pima Indian (native) women at least 21 years old. The data's source is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [3]. The data "PimaIndiansDiabetes2" is extracted from the mlbench package in CRAN. The data contains medical examination variables of the women and a class variable for diabetes test. The class variable indicates a positive test for diabetes between 1 to 5 years from the examination, or a negative test for diabetes after 5 or more years from the examination.

1.2 Objective & Questions

Early detection can help prevent diabetes, as pre-diabetes is reversible. We want to detect diabetes outcome using early physical indicators as predictors.
From our data, we aim to investigate:

1. *How many natural clusters do the predictor variables belong to?*

Hypothesis: Two natural clusters if data is highly predictive for diabetes (two level outcome).
We then explore the distinction between clusters and find out:

2. *Can we classify which cases will be diabetic based on their physical indicators?*

2 Data Preparation

Data "PimaIndiansDiabetes2" was extracted instead of "PimaIndiansDiabetes" as the latter contains physical impossibilities which are replaced as NA in "PimaIndiansDiabetes2". Upon inspection, the data contains 9 variables. 8 are numeric predictors and 1 is the two-level factor outcome, "diabetes". The meaning of each variable is compiled in Appendix.
We obtain summary statistics for each variable (in Appendix) and there are 768 total cases, their distribution by outcome shown in Table 1. We also note the number of NA cases for each variable in Table 2.

diabetes	count	% of cases
neg	500	65.1
pos	268	34.9

Table 1: Original distribution of diabetes outcomes

insulin	triceps	pressure	mass	glucose
374	227	35	11	5

Table 2: No. of NA cases for each variable

First, there is a class imbalance in positive to negative cases of 7:13. This indicates some difficulty in predicting positive class as the rarer event has smaller proportion in training data. The imbalance is not severe, so we do not under-sample the negative cases. Also, since a classification model that blindly classifies every case as negative will have 65% accuracy, we establish a baseline performance of 65% accuracy.
We drop all NA cases to enable unbiased investigation of relationships between predictor variables. A total of 376 cases were dropped and the sample size left is 392. The distribution of diabetes outcome remained almost the same (67% neg) corresponding to class imbalance of 1:2.

Finally, diabetes is closely related to insulin resistance [4] and can be indicated by Glucose/Insulin ratio. We then add a 9th feature G/I ratio to the data, transformed by dividing glucose by insulin.

3 Exploratory Data Analysis

3.1 Variable Distribution & Relationship with Diabetes Outcome

To investigate the distribution of variables, we group them by diabetes outcome and plot a boxplot + violin for each predictor variable. First, we observe that a few variables contain extreme outliers, which may be data entry/measurement errors or otherwise unreasonably skew the distribution. An example is a case with insulin>800uU/mL, likely to be an error. We remove several of such outliers (see Appendix), while keeping those not too extreme as they may be natural variations and removing may lower generalisability of our model. We also want to maintain sufficient sample size. After dropping 10 extreme outliers, the final sample size is 382.

From these plots we can also visualise the general relation of predictors variables with diabetes outcome. We see a difference in the distributions for negative vs positive outcomes.

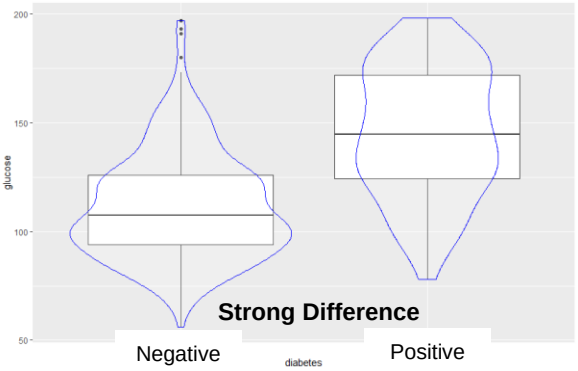


*Figure 1: Boxplot of **glucose** variable.*

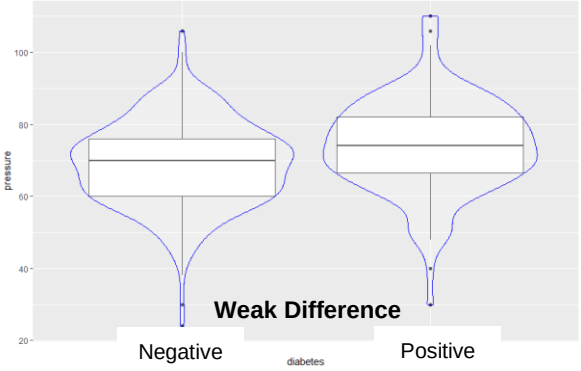


*Figure 2: Boxplot of **pressure** variable.*

For example, from Fig 1, the median glucose for negative is clearly lower than even the lower quartile for positive, suggesting a strong likelihood for women with higher glucose level to develop diabetes. This makes glucose a significant predictor when classifying for diabetes outcome. In contrast, Fig 2 shows a similar distribution of pressure for both groups, so pressure is not as strong a predictor for diabetes class.

We further our analysis by charting the mean outcome which is the probability of positive diabetes for each variable case. An example is shown in Fig 3 below.

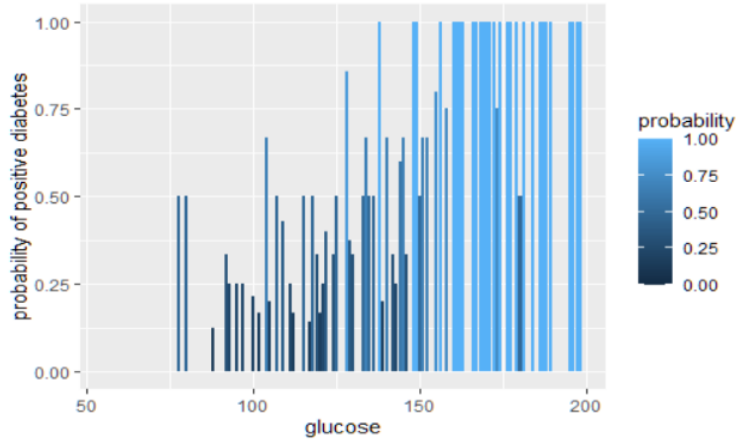


*Figure 3: Probability distribution for **glucose.***

The overall trend shows that probability increases as glucose increases.

It supports the conclusion drawn from Fig 3, that glucose is a strong predictor for diabetes class.

From our EDA, we summarise the association of variables with diabetes class in Table 3 below. We also note that other than G/I ratio which tends to be lower for positive cases, all other variables tend to be higher for positive cases.

Degree of association with diabetes class	Variables
Low	Pressure
Medium	Pregnant, Triceps, Insulin, Mass, Pedigree
High	Glucose, Age, G/I ratio

Table 3: Summary of predictor variables' degree of association with diabetes class.

Next, we check the pairwise correlation of variables with scatterplot and correlation plot.(Appendix) From the scatterplot, we can observe clusters of pink (positive) that are distinct from regions of green (negative), which tells us which variables can separate the class clearly. From the correlation coefficients, we find the highly correlated pairs to be glucose & insulin, triceps & mass, pregnant & age.

3.2 Hierarchical Clustering

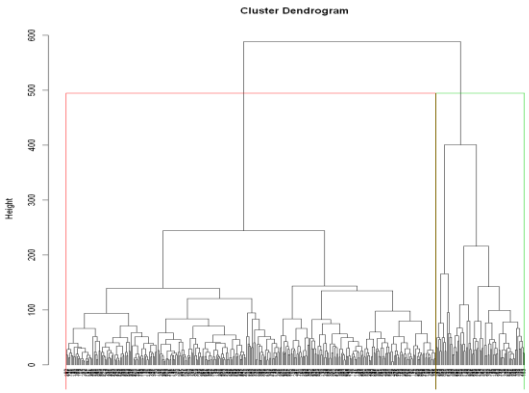


Figure 4: Cluster Dendrogram, 2 natural clusters

Finally, we test our hypothesis of two natural clusters corresponding to diabetes outcome, using hierarchical clustering. We observe 2 distinct clusters (red vs green), at the highest level, across all cases. Cluster 1 has 308 cases and Cluster 2 has 74 cases. Positive cases make up 28% of Cluster 1 and 50% of Cluster 2. Comparing this to the overall dataset which has 35% positive cases, Cluster 1 has higher proportion of negative cases and Cluster 2 has much higher proportion of positive cases. Though not conclusive, it shows that the clusters have separation by diabetes class, supporting hypothesis of two natural clusters, allowing us to proceed with classification using our predictor variables.

4 ML Classification

4.1 Train:Test Split

As the sample size is only 382, the Train:Test split ratio must be optimized as the variance in test data will be too large for accurate testing. We achieved this by running different split ratios with 500 seed values to obtain lowest uncertainty use logistic regression as a testing model.

Train : Test	80:20	70:30	60:40	50:50
Accuracy Uncertainty Range	23.6%	19.3%	15.9%	15.3%

Table 4: Different accuracy range for different split ratio

The model is able to classify with same accuracy even with lower train size and higher test size. This is most likely due to the saturation of model fitting; hence using a larger test data

with less variance is more suitable. The 60:40 split was decided as further uncertainty improvement is not significant. We also kept test data consistent to compare across models.

4.2 Model Tests & Comparison

We attempt Logistic Regression, kNN classification and SVM as they are useful for classification of binary diabetes outcome. All predictor variables were used except pressure, which has low association with diabetes outcome. For kNN, the data is first scaled with min-max normalisation and the best k=20. SVM Polynomial kernel was also tested as we suspected possible non-linear decision surface.

	Logistic Regression	kNN Classification	SVM Linear	SVM Polynomial
Accuracy	75%	77%	76%	76%
Confusion Matrix	86 22 16 29	94 27 8 24	87 22 15 29	90 25 12 26

Table 5: Summary of models' performance

All models beat the baseline 65% accuracy, indicating that all can classify diabetes outcome with some success. There is no obvious winner but kNN has the best accuracy of 77%, beating baseline by additional 12%. It also has the lowest False Positive (FP) rate of 7.8%. Logistic regression and SVM linear both have the lowest False Negative (FN) rate of 43%. The models all have much higher FN rate than FP rate, indicating a higher difficulty in classifying positive cases as they tend to predict negative, possibly due to the class imbalance.

4.3 Final Model

Since the models are less likely to predict positive cases, the low FP rate on kNN could be due to overfitting on the class imbalanced data. Thus, we select SVM linear as our final model due to its lowest FN rate. False negative classifications will lead to undetected cases which defeats the purpose of our model to prevent future diabetes, so we want the lowest FN rate. SVM is also suitable as the predictor variables are all quantitative, so there is no categorical predictor affecting margin calculation.

5 Conclusion

First, using Hierarchical Clustering, we found 2 natural clusters in the dataset, which seem to have distinction in proportion of positive diabetes cases. Second, using SVM linear we are able to classify if a case will develop diabetes within the next 5 years with 76% success rate on the test data.

A limitation of our model is that the FN rate is still considered high even in SVM linear. The class imbalance causing high FN rate may be improved by oversampling the positive cases. Another limitation is the small sample size causing high variance in accuracy which makes it difficult to compare models. An improvement could be to create an artificially larger sample size by imputing the NA cases using regression methods to keep our NA rows. More causal predictors like cholesterol level can be obtained from other datasets to increase predictive power.

Finally, our findings may possibly enable some extrapolation to the general population outside of Pima Indians.

References

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes

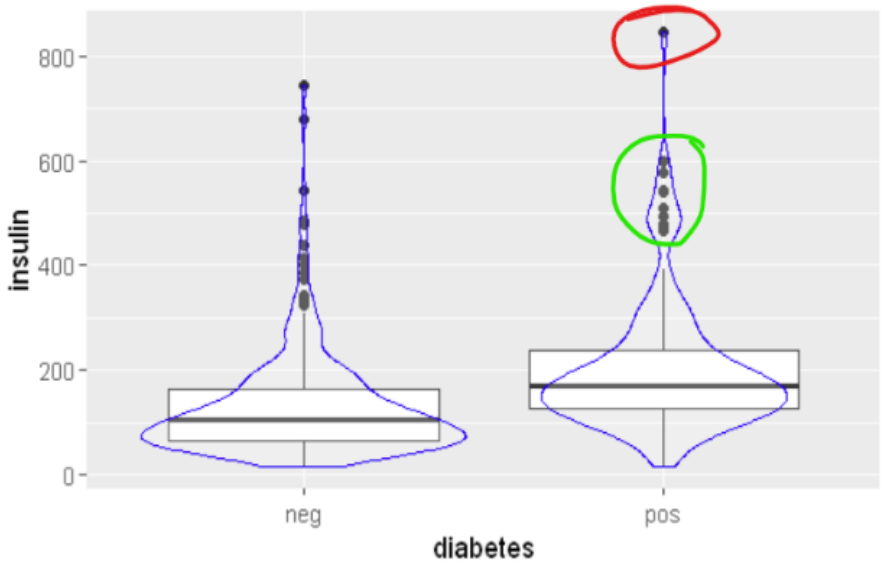
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3830901/

[3] Grace Whaba, Chong Gu, Yuedong Wang, and Richard Chappell (1995), Soft Classification a.k.a. Risk Estimation via Penalized Log Likelihood and Smoothing Spline Analysis of Variance, in D. H. Wolpert (1995), The Mathematics of Generalization, 331-359, Addison-Wesley, Reading, MA.

[4] https://medicine.musc.edu/departments/family-medicine/research/rcmar/insulin-resistance

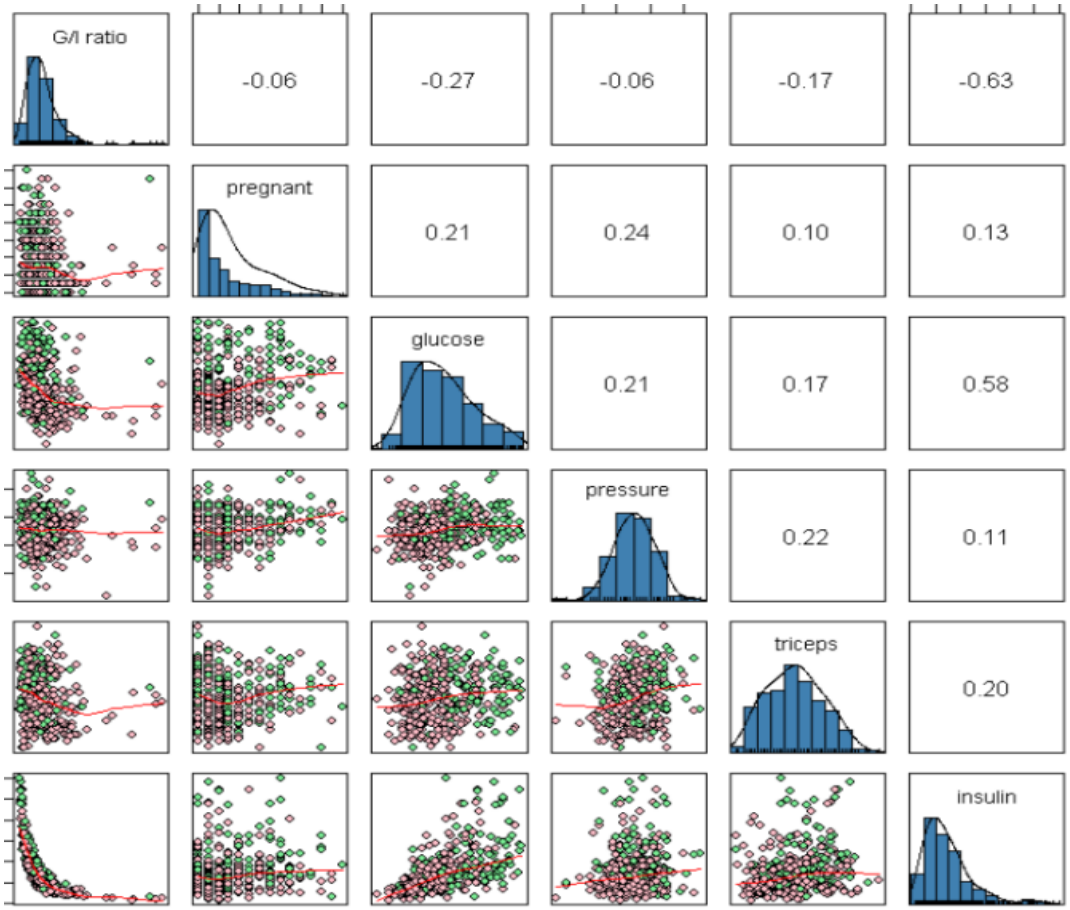
Appendix

pregnant	Number of times pregnant
glucose	Plasma glucose concentration (glucose tolerance test)
pressure	Diastolic blood pressure (mm Hg)
triceps	Triceps skin fold thickness (mm)
insulin	2-Hour serum insulin (mu U/ml)
mass	Body mass index (weight in kg/(height in m)\^2)
pedigree	Diabetes pedigree function
age	Age (years)
diabetes	Class variable (test for diabetes)



RED: Outliers removed

Green: Outliers kept



FUNCTIONS, RCODE FURTHER BELOW

#function for finding best split %

```
x = rep(0,500)
for(i in 1:500){
  set.seed(i)
  training.idx = sample(1:nrow(dia_factored),nrow(dia_factored)*0.4)
  train.data = dia_factored[training.idx,]
  test.data = dia_factored[-training.idx,]
  mlogist <- glm(y~., data=train.data, family="binomial")
  predictions <- predict(mlogist,test.data, type='response')
  y_pred <- factor(ifelse(predictions>0.5,1,0),levels=c(0,1))
  x[i] = mean(y_pred == test.data$y) }
max(x) ; min(x)
max(x)-min(x)
```

#function for geombar

```
for(i in 1:8){
  diageomplot <- dia %>% mutate(y=ifelse(diabetes == 'pos',1,0)) %>% select(-
diabetes) %>% group_by(dia[i]) %>% summarise_at(vars(y), list(probability =mean))
```

```
print(ggplot(diageomplot,aes(x=diageomplot[[1]],y=probability,fill=probability))+geom_bar(
stat='identity')+xlab(colnames(diageomplot[1])) +ylab('% of positive diabetes'))
}
```