Understanding Diabetes Datasets: A Comparative Analysis

Daniele Lotito, PhD student of the National PhD in AI

SLLD course, modules 1 and 2

Contents

Abstract

This report focuses on the analysis of three diabetes datasets: Stanford, Pima Indians, and Iraqi society. Through an exploratory analysis, we aim to identify differences and potential disparities within the data sets, emphasizing variations in sample characteristics and variable meanings. Using statistical learning methods, our analysis seeks to highlight discrepancies between data sets and the need for rigorous examination of data quality, transparency, and proper documentation. Thus, the goal of this report is not to discover new findings on diabetes. By providing insights into the complexities of these datasets, we also intend to contribute to the understanding of the appropriate use of data and promote data-driven approaches in research.

1 Introduction

Diabetes mellitus is a prevalent chronic metabolic disorder affecting millions of individuals worldwide. It is characterized by elevated blood sugar levels resulting from defects in insulin secretion, insulin action, or both.

This report focuses on the analysis of three diabetes datasets: the Stanford dataset, the Pima Indian dataset, and the Iraqi society dataset, described in the following sections. The Stanford dataset, obtained from the "Least angle regression" paper by Efron, Hastie, Johnstone, and Tibshirani, consists of data of 442 diabetes patients. The Pima Indian dataset, collected by the National Institute of Diabetes and Digestive and Kidney Diseases, includes medical and general information of 768 Pima Indian women, and 268 of them have diabetes. The Iraqi society dataset is constructed from medical information and laboratory analyses extracted from patient files collected by the Medical City Hospital and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital in Iraq, this dataset has data from 103 non-diabetic, 53 predicted-diabetic, and 844 diabetic patients.
Our project primarily aims to utilize statistical learning methods to analyze

and compare various diabetes datasets, rather than discovering new findings about diabetes itself. The focus lies in highlighting differences, potential limitations, and considerations in data interpretation, emphasizing the significance of data quality, transparency, and thorough documentation.

However, throughout our analysis, two encouraging findings have emerged. First, we found that while the presence of diabetes positively correlates with age, the progression of the disease is primarily determined by other factors that can be mitigated through a healthy lifestyle. This finding underlines the importance of lifestyle modifications in diabetes management.
Second, the Iraqi society dataset response variable is the patient's diabetes disease class, that can be diabetic, non-diabetic, or predict-diabetic. Even if the data collectors have not been specific about the meaning of the predict-diabetic label, our analysis suggests that predict-diabetic patients share more similarities with non-diabetic individuals, indicating that they have an opportunity to halt the progression of the disease if proactive measures are taken. In addition to this, we highlight the necessity of providing adequate information about the dataset along with the actual data.

By these analyses, we aim to contribute to the broader understanding of diabetes datasets and provide insights into best practices for their use in research and decision-making, and also for the building of new datasets.

2 A closer look at the datasets

2.1 The Stanford dataset

The Stanford dataset [1], obtained from the "Least angle regression" paper by Efron, Hastie, Johnstone, and Tibshirani, is a widely recognized resource for diabetes research. It consists of baseline variables and serum measurements of 442 diabetes patients. The dataset contains ten feature variables and a response variable representing a quantitative measure of disease progression one year after the baseline assessment.

The ten feature variables in the Stanford dataset are as follows:

- **Age**: Represents the age of the diabetes patients at the time of the baseline assessment.
- **Sex**: Indicates the sex of the diabetes patients. It is a categorical variable, with the specific coding or reference category not mentioned

in the available information.

- **BMI** (Body Mass Index): Measures the body mass index of the patients, providing insights into potential weight-related health risks.
- **MAP** (Mean Arterial Pressure): Represents the average blood pressure of the patients, taking into account both systolic and diastolic measurements.
- **TC or TCH** (Total Cholesterol): This variable is labeled inconsistently in the provided information, as both "TC" and "TCH" could refer to the total cholesterol level. May be the case that TC stands for total cholesterol and "TCH" for Thyroid-Stimulating Hormone, but this measurement is used to be denominated with the acronym "Thyroid-Stimulating Hormone". Further clarification is required to determine the precise meaning of these variables, yet they should be considered with extreme caution.
- **LDL** (Low-Density Lipoprotein): Represents the concentration of low-density lipoprotein cholesterol in the patients' blood. High levels are associated with an increased risk of cardiovascular diseases.
- **HDL** (High-Density Lipoprotein): Indicates the concentration of high-density lipoprotein cholesterol in the patients' blood. Higher levels are associated with a reduced risk of cardiovascular diseases.
- **LTG** (Triglycerides): Triglycerides play a significant role in diabetes by affecting insulin sensitivity and increasing the risk of cardiovascular complications. However, "LTG" acronym is often associated with lamotrigine, a medication used to treat epilepsy and bipolar disorder, in the patients' blood. If this is the case it may serve as a confounding factor, in section 4.4 we will discuss its meaning with data at hand.
- **GLU** (Glucose): Indicates the glucose level in the patients' blood, an essential parameter for diabetes diagnosis and management.

It is important to note that each of the ten feature variables in the Stanford dataset has been mean-centered and scaled by the standard deviation multiplied by the square root of n, where n represents the number of patients (442 in this case). This normalization ensures that the sum of squares of each

column totals 1, facilitating meaningful comparisons and analyses. However, the normalization of the sex variable is of dubious utility, and it is unclear from the provided information if the base group is "male" or "female".

Moreover, another notable limitation of the Stanford dataset is the lack of a comprehensive description of the features. The provided information does not offer explicit explanations or context for the variables, including the blood measurements. It remains unclear which variable specifically stands for total cholesterol, as it is labeled inconsistently as "TC" and "TCH". Moreover, also the "LTG" feature raises some doubts, if the analyses will point that this feature is relevant for the regression task, we could be allowed to conclude that "LTG" stands for triglycerides.

Researchers analyzing this dataset should exercise caution and consider seeking additional information or consulting data collectors and domain experts to ensure proper interpretation and utilization of the data. Further investigation and clarification of variable meanings are necessary to derive accurate conclusions and insights from the Stanford and Iraqi datasets.

In the subsequent sections, we will employ various statistical methods to explore and analyze the dataset, but first we will introduce the other two datasets used in this study.

2.2 The Pima Indians Dataset

The Pima Indians dataset [2], originating from the National Institute of Diabetes and Digestive and Kidney Diseases, is a valuable resource for studying diabetes. The primary objective of this dataset is to predict the likelihood of a patient having diabetes based on specific diagnostic measurements. The dataset exclusively focuses on female patients who are at least 21 years old and belong to the Pima Indian heritage.

Here are the variables contained in the Pima Indians dataset:

- **Pregnancies**: This variable represents the number of times a patient has been pregnant.
- **Glucose**: The concentration of plasma glucose measured 2 hours after an oral glucose tolerance test.
- **BloodPressure**: The patient's diastolic blood pressure.
- **SkinThickness**: The thickness of the triceps skinfold.

- **Insulin**: The patient's serum insulin level measured 2 hours after ingestion.
- **BMI**: The Body Mass Index of the patient.
- **DiabetesPedigreeFunction**: A function that scores likelihood of diabetes based on family history.
- **Age**: The patient's age at the time of data collection.
- **Outcome**: This is the class variable. 0 indicates a non-diabetic patient, while 1 indicates a diabetic patient.

The Pima Indians dataset serves as a rich resource for exploring the factors related to diabetes, particularly in the context of the Pima ethnic group. It has data of 268 diabetic patients and 500 non-diabetic patients. The detailed information it provides about each variable allows for a comprehensive analysis of the different factors that may influence the onset of diabetes.

2.3 The Iraqi Society Dataset

The Iraqi Society dataset [3] is a detailed collection of diabetes-related medical information and laboratory analyses from a diverse range of patients within the Iraqi population. The dataset includes the following variables:

- **Age**: The patient's age at the time of data collection.
- **Sex**: The patient's gender.
- **BMI**: The patient's Body Mass Index.
- **Chol**: The patient's total blood cholesterol level.
- **LDL**: The patient's blood concentration of Low-Density Lipoprotein cholesterol.
- **HDL**: The patient's blood concentration of High-Density Lipoprotein cholesterol.
- **VLDL**: The patient's blood concentration of Very-Low-Density Lipoprotein cholesterol.

- **CLASS**: The patient's diabetes disease class (Diabetic, Non-Diabetic, or Predict-Diabetic).

The dataset includes 103 non-diabetic patients, 53 predicted-diabetic patients, and 844 diabetic patients. As we can see already, there are variables common to other dataset, and no specific meaning of the "Predict-Diabetic" class is provided. The only thing we can say at this stage is that the task associated to this dataset is a multiclass classification, that can be addressed with a multiclass logistic regression. To be more specific, the response variable is qualitative and ordered, and we could also approach the task with ordered logistic regression, provided that the hypothesis of the ordered logistic regression are satisfied. We will discuss this point later in this report.

3 Data Preprocessing

In our project, preprocessing plays a critical role in preparing the datasets for further analysis. This step begins with renaming and ordering the feature columns, ensuring a consistent structure across all datasets. Subsequently, we perform a check for duplicate entries and remove them to maintain data integrity. To facilitate cross-dataset comparison, we select a subset of features that are common across different datasets or are easily accessible. This uniform selection of features enables us to apply the same normalization method used in the Stanford dataset to the other datasets. Finally, we identify and remove outliers to mitigate their potential impact on the accuracy of our analyses.

3.1 Identification and Removal of Outliers

Each dataset underwent an initial visual inspection for potential outliers. The Iraqi dataset needed more care due to potential data entry errors [**?**], but the Pima and Stanford datasets were also investigated for outliers within their respective variables. The first common step was to scale the data in the same way.

3.1.1 Outlier Removal

A step was dedicated to the Iraqi dataset to check for potential anomalies in the feature distributions. Box plots and histograms were employed for this

purpose. In particular, we inspected the distribution of features per each response class, to be sure that the number of outlier was roughly proportional to the number of data points in each class. In other words, we wanted to be sure that outliers were not valid data points characterized by high values of some features caused by the belonging to a certain class. After the identification of outliers in the Iraqi dataset, a removal process was initiated, and the data was re-plotted using histograms and scatterplots to verify the successful exclusion of outliers. Subsequently, the dataset was scaled again to allow for a more accurate comparison with the Stanford dataset.

3.1.2 Outlier Management in Pima and Stanford Datasets

Stanford and Pima datasets underwent a similar process, where outliers were identified based on a 3.5 z-score threshold and were subsequently removed from the dataset, and the data was re-plotted to confirm the effectiveness of the outlier management. Only few points ($\simeq 5$) per dataset were removed at this stage. As the data has been rescaled again after the first removal, some outliers (according to the same criteria) still remained in all the datasets, but for the same reason -i.e. the second scaling- they did not necessitate any handling.

4 Data Exploration and Analysis

The data exploration process is fundamental to better understand the data at our disposal. In the following we will focus on more specific questions that have at their core the comparison between the data. To this end, we first sum up the common features of the datasets in table 1 and then start by considering the two features that are common to all the datasets. These features are age and body mass index, two general features that are accessible immediately for everybody.

4.1 Basic exploratory analysis of Age and BMI

We now discuss one of our main findings: *diabetes presence positively correlates with Age, but the disease progression is mainly determined by other factors that we can address with a healthy lifestyle.* To this purpose we will assess the relevance of these feature in the task related to each dataset, and

Feature	Stanford Dataset	Iraqi Dataset	Pima Dataset
LDL	Yes	Yes	No
HDL	Yes	Yes	No
Glucose *	Yes	No	Yes
Blood Pressure *	Yes	No	Yes
Age	Yes	Yes	Yes
BMI	Yes	Yes	Yes

Table 1: Common Features in Stanford, Iraqi, and Pima Datasets. Stanford and Pima Indians datasets share glucose and blood pressure measurements, but these measurements are taken in different clinical settings, thus these features are not comparable, at least without using domain knowledge.

see if the age feature is less relevant in determining the disease progression rather than in the diagnostic process.

4.1.1 Correlation Analysis

Scatter plots were used to visualize the distribution of "age" and "bmi" in each dataset. These scatter plots provided a preliminary understanding of the relationship between age and BMI in each dataset. Focusing only on linear relations, correlation coefficients between "age" and "bmi" were computed for each dataset to assess the strength of their linear relationship. The correlation coefficient ranges from -1 to 1, where -1 indicates a strong negative linear relationship, 1 indicates a strong positive linear relationship, and 0 indicates no linear relationship. The results are shown in table 2, while there is positive correlation, we observe that it is not strong, as all the correlations are below 0.5. The fact that these variables are not strongly correlated encourages the searching for the "separate effect" of each of them in the various tasks.

Table 2: Correlation between Age and BMI in different datasets

Dataset	Correlation (age and BMI)
Stanford	0.185
Pima	0.031
Iraqi	0.403

4.2 Assessment of "age" and "bmi" in the various tasks

The relevance of "age" and "bmi" in regression and classification tasks for each dataset were then assessed.

4.2.1 Regression analysis in the Stanford analysis

For the Stanford dataset, linear regression was used with "age" and "bmi" as predictors. The results are shown in figure 1. The use of only two regressors worsens the performance a lot, but we can already observe that here "bmi" seems to be a more relevant feature.

```
Residuals:
     Min       1Q   Median       3Q      Max
-156.147  -45.139   -7.835   46.276  152.735

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.133      2.962  51.360   <2e-16 ***
bmi          924.816     63.369  14.594   <2e-16 ***
age          133.014     63.369   2.099   0.0364 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.27 on 439 degrees of freedom
Multiple R-squared:  0.3504,    Adjusted R-squared:  0.3475
F-statistic: 118.4 on 2 and 439 DF,  p-value: < 2.2e-16
```

Figure 1: Relevance of age and bmi in the multiple linear regression for the Stanford dataset.

4.2.2 Classification Analysis in the Pima dataset

For the Pima dataset, logistic regression was used for binary classification with "age" and "bmi" as predictors. The results are in figure 2. We can observe that "age" and "bmi" are equally important predictors for this task.

4.2.3 Multiclass Classification in the Iraqi Dataset

The Iraqi dataset includes the "CLASS" feature, which represents the patient's diabetes disease class. This categorical feature can take the values "Y" (Diabetic), "N" (Non-Diabetic), or "P" (Predict-Diabetic). To perform

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.76007    0.08755  -8.682  < 2e-16 ***
age          0.57367    0.08461   6.780  1.2e-11 ***
bmi          0.69882    0.09126   7.657  1.9e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 928.96  on 721  degrees of freedom
Residual deviance: 816.53  on 719  degrees of freedom
AIC: 822.53

Number of Fisher Scoring iterations: 4
```

Figure 2: Relevance of age and bmi in the binary classification for the Pima dataset.

multiclass classification, ordered logistic regression was used with “age” and “bmi” as predictors. The results are shown in figure 3. Here, “bmi” seems to be a more relevant feature, a more comprehensive analysis will be carried out in section 4.4.

4.2.4 Interpretation of Regression Coefficients

The coefficients for “age” and “bmi” reflect the change in the log-odds of the outcome for a one-unit increase in the predictor, holding all other predictors constant. A positive coefficient suggests that an increase in the predictor is associated with an increase in the log-odds of a higher category of the outcome variable [4, 5].

4.2.5 Interpretation of Regression Intercepts

The intercepts obtained from the ordered logistic regression were also interpreted. These intercepts represent the log-odds of the outcome when all predictors are zero. The intercepts for “N—P” and “P—Y” compare the likelihood of being in one category versus the reference category.

```
Call:
polr(formula = response_iraqi ~ age + bmi, data = iraqi_std)

Coefficients:
     Value Std. Error t value
age 0.3623     0.1303    2.78
bmi 2.7900     0.2694   10.36

Intercepts:
    Value    Std. Error t value
N|P  -4.3833   0.3144   -13.9422
P|Y  -3.5970   0.2868   -12.5426

Residual Deviance: 486.7355
AIC: 494.7355
```

Figure 3: Relevance of age and bmi in the ordered logistic regression for the Iraqi dataset.

4.2.6 Assessment of Model Assumptions

The proportional odds assumption or parallel regression assumption of ordinal logistic regression was discussed. This assumption states that the relationship between each successive pair of outcome groups is consistent. Checking the validity of this assumption is important to ensure the appropriateness of the model.

4.2.7 Summary

To sum up, the relevance of the "age" and "bmi" features in each dataset was assessed. It was found that both features are relevant in all three datasets for their respective tasks, with exception to "age" in the Stanford and Iraqi dataset. This suggests to further explore the importance of age in the various tasks, we will address this issue in section 4.4 when we will carry out a feature selection analysis.

4.3 A little cluster tour

Here we briefly show that the cluster analysis reveals that

- In the Stanford dataset, clustering on the regressors, we group people with similar values of the response variable. This shows that the clini-

cal measurements alone characterize a medical profile more inclined to favor disease progression. This can be viewed in figure 4.

- In the Iraqi society dataset, the response variable represents the diabetes disease class of the patient, categorized as diabetic, non-diabetic, or predict-diabetic. The latter category, predict-diabetic, is not explicitly defined by the data collectors. However, our analysis suggests that patients in the predict-diabetic category exhibit a significant similarity to non-diabetic individuals. This assumption is further supported by a cluster analysis, which appear to group non-diabetic and predict-diabetic patients together, reinforcing the similarities between these two categories, as can be seen in Figure 5.

The clustering technique used is the k-mean clustering, the feature considered for the clustering were "age", "bmi", "map", "ldl", "hdl", "tch", "ltg" "glu" for the Stanford dataset and "age", "bmi", "chol", "ldl", "hdl" for the Iraqi dataset. We also performed jointly a PCA analysis and hierarchical clustering, they are discussed in the notebook in the supplementary documentation.

4.4 Feature selection and the role of age

For each of the three models we have run an exhaustive search for the best predictors and leveraged the best models with an information criteria. The results are shown in figures [Fig. 6], [Fig. 7] and [Fig. 8]. We now clearly see that the "age" feature is more relevant in the classification tasks. We observe that although in section 4.2.3 "age" was less relevant than "bmi" in the multiclass classification tasks related to the Iraqi dataset with only two regressors, when compared with other predictors it is a really important explanatory variable. We can observe from [Fig. 8] that "LTG" is a very important feature, as a consequence the more probable interpretation "LTG" is *triglycerides*, and we can surely exclude that "LTG" stands for lamotrigine.

4.4.1 Summary

One of the central discoveries of our analysis concerns the correlation between age and diabetes. While the occurrence of diabetes does indeed rise with age, it's crucial to note that the progression of the disease is primarily shaped by other factors, which are also often correlated with age but can be

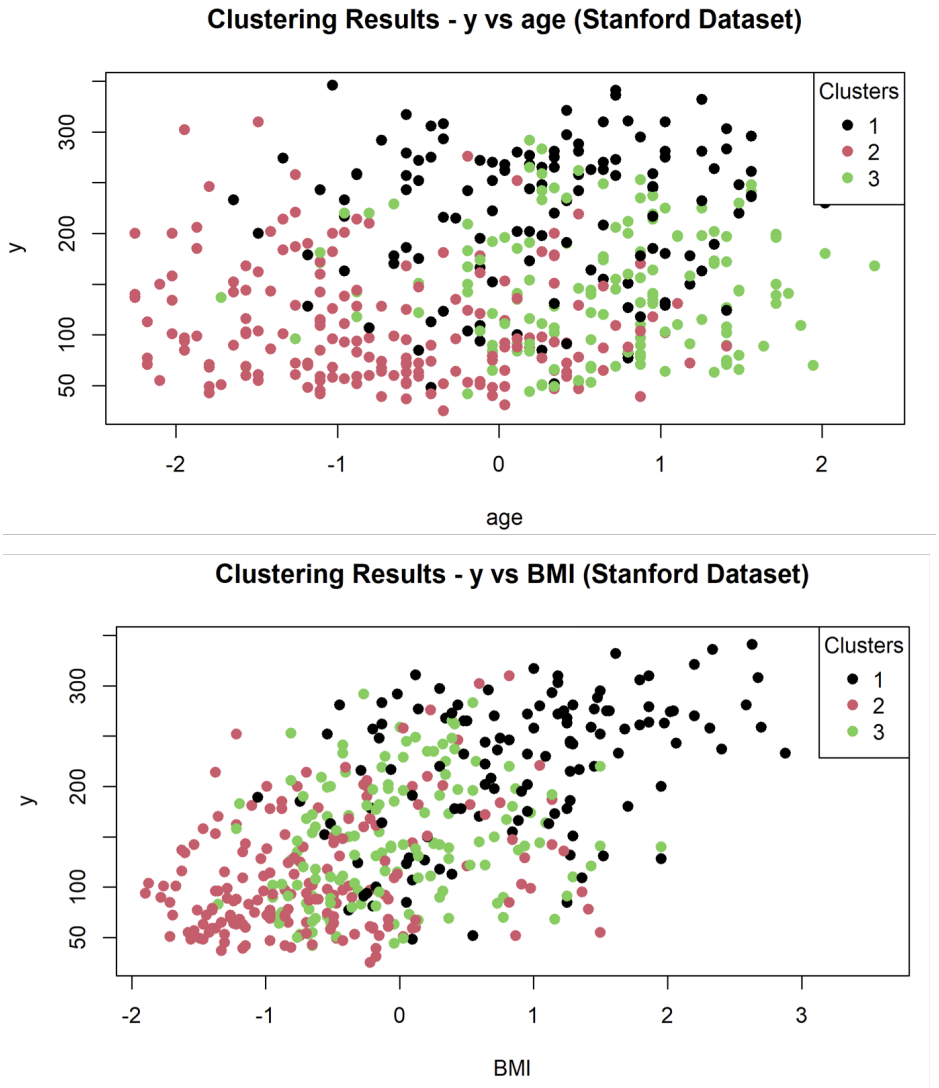
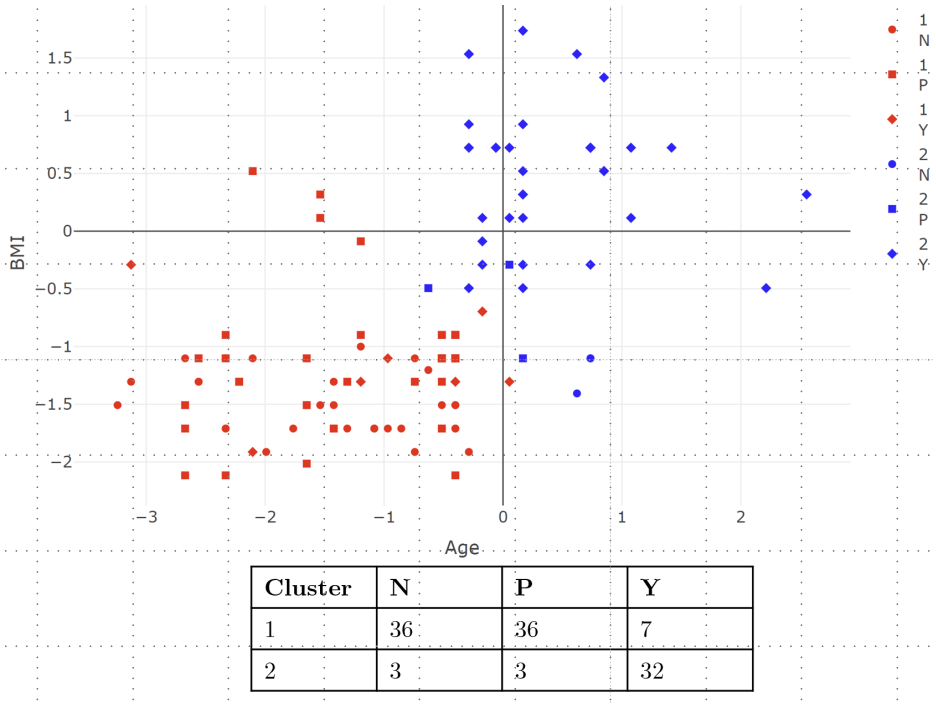


Figure 4: Clustering analysis of the Stanford dataset. The figure showcases the grouping of individuals based on similar values of the response variable when clustering is applied on the regressors.



Cluster	N	P	Y
1	36	36	7
2	3	3	32

Figure 5: Cluster analysis and membership table of the Iraqi dataset showing the grouping of non-diabetic and predict-diabetic patients.

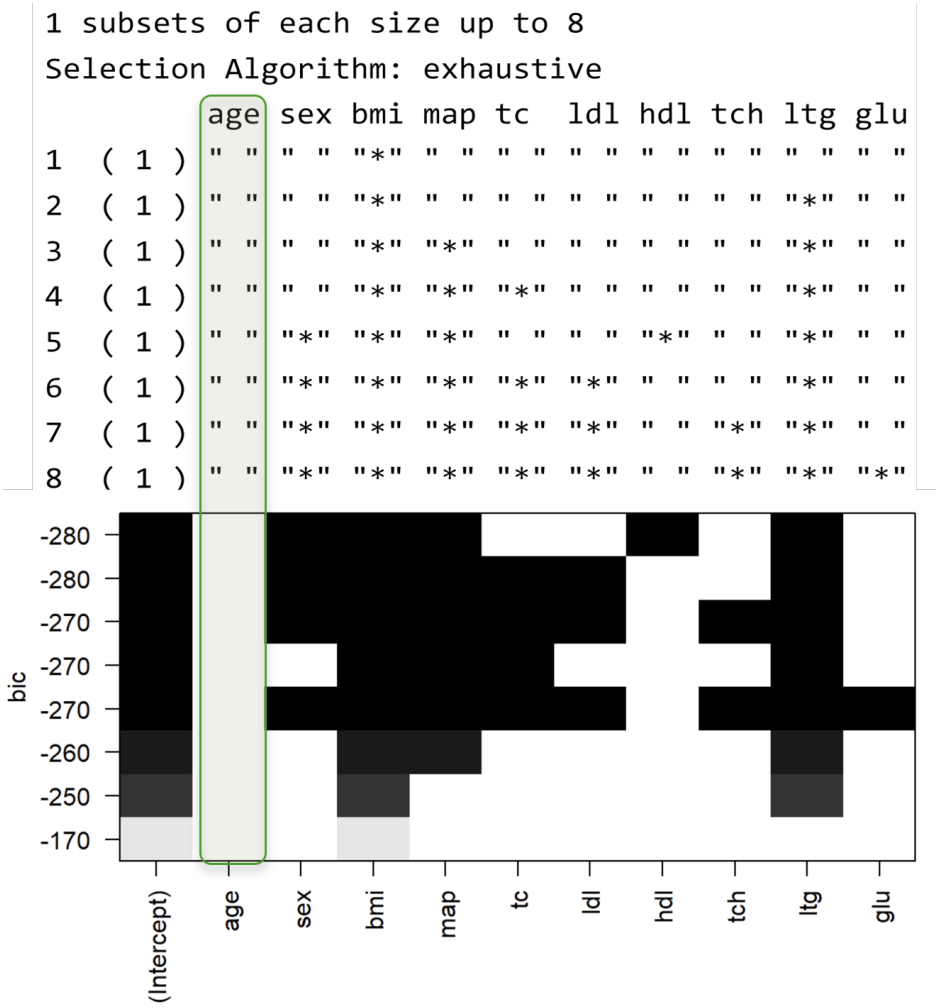


Figure 6: Best model selection through exhaustive search and visualization of their performance with BIC information criteria for the Stanford dataset.

```
bglm.AIC_pima = bestglm(Xy = pima_std, family = binomial, IC = "AIC",
    TopModels = 7)
```

age <lgl>	bmi <lgl>	BloodP <lgl>	glu1 <lgl>	insulin <lgl>	Criterion <dbl>
TRUE	TRUE	FALSE	TRUE	FALSE	699.9520
TRUE	TRUE	TRUE	TRUE	FALSE	700.8072
TRUE	TRUE	FALSE	TRUE	TRUE	701.1338
TRUE	TRUE	TRUE	TRUE	TRUE	701.7049
FALSE	TRUE	FALSE	TRUE	TRUE	714.8213
FALSE	TRUE	FALSE	TRUE	FALSE	715.3604
FALSE	TRUE	TRUE	TRUE	TRUE	716.8213

Figure 7: Best models selection through exhaustive search and their assessment with AIC information criteria for the Pima dataset.

```
bglm.AIC_iraqi = bestglm(Xy = iraqi_bin, family = binomial, IC = "AIC",
    TopModels = 5)

bglm.AIC_iraqi$BestModels
```

age <lgl>	sex <lgl>	bmi <lgl>	chol <lgl>	ldl <lgl>	hdl <lgl>	Criterion <dbl>
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	272.1818
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	272.6340
TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	273.9787
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	274.0679
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	274.5267

Figure 8: Best models selection through exhaustive search and their assessment with AIC information criteria for the Iraqi dataset.

positively influenced through healthy habits. This means that although older age groups may be more susceptible to diabetes due to various physiological changes and potentially ingrained unhealthy lifestyle habits, age is not an absolute predictor of how the disease will progress. Instead, our data analysis may suggest that lifestyle factors such as diet, physical activity, stress management, and regular medical check-ups play a significant role in managing the disease's progression. In conclusion, the age-diabetes relationship isn't as straightforward as it might seem at first glance. The connection between lifestyle factors and disease progression underlines the essential role of proactive health management, regardless of age, in diabetes control. We hope our findings will prompt further exploration into the role of lifestyle interventions in managing diabetes and inform future research and policy directions.

5 Conclusions

Our analysis of the Stanford, Pima Indian, and Iraqi society diabetes datasets demonstrated the importance of data quality, transparency, and comprehensive documentation for research and decision-making purposes. Our key findings indicate that while diabetes incidence correlates with age, disease progression is largely influenced by lifestyle factors, obviously correlated with age too, but that can be addressed with healthy habits. Moreover, the predict-diabetic class in the Iraqi dataset suggested that proactive intervention might help these individuals avoid or delay diabetes onset. However, the vague nature of the predict-diabetic category emphasizes the need for clear data documentation. From a personal point of view, the work conducted throughout this course could be the basis for potential future collaborations. Envisioning the expansion of this project, it would be beneficial to team up with an expert possessing medical domain knowledge. Furthermore, it could be interesting to explore panel datasets and focus on a specific task.

Supplementary material

The supplementary material related to this report is a quarto R notebook. The notebook presents the computational details of this analysis and makes it fully reproducible and exportable to a `.html` file.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[3] M. Kuhn, and K. Johnson. *Applied Predictive Modeling*. Springer, 2013.

[4] R. L. Prentice, and R. Pyke. *Logistic Disease Incidence Models and Case–Control Studies*. Biometrika, 1979.

[5] M. R. Rajput, and S. S. Khedgikar. *Diabetes prediction and analysis using medical attributes: A Machine learning approach*. Journal of Xi, 2022.

[6] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus*. In Proceedings of the Symposium on Computer Applications and Medical Care, pp. 261–265. IEEE Computer Society Press, 1988.

[7] H. Wickham, and G. Grolemund. *R for Data Science*. O'Reilly Media, 2017.

Online resources

[1] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). *Least angle regression Stanford diabetes dataset*. Available at: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

[2] Pima Indian. (2016). *Diabetes Dataset*. Available at: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

[3] Iraqi society. (2020). *Diabetes Dataset*. Available at: https://data.mendeley.com/datasets/wj9rwkp9c2/1

[4] Ordinal logistic regression. Available at: https://www.karlin.mff.cuni.cz/~pesta/NMFM404/ordinal.html

[5] Harrell Jr, F. E. (2001). *Regression Modeling Strategies*. Available at: https://hbiostat.org/doc/rms.pdf

[6] UCLA: IDRE (Institute for Digital Research and Education). *Data Analysis Examples*. Available at: http://www.ats.ucla.edu/stat/dae/

Appendix - Preprocessing in Iraqi Dataset

This appendix makes explicit the first step in the pre-processing of the Iraqi dataset. More information can be found in the notebook cited in the supplementary material, which discusses the entire analysis reproducibly and in the same way as it is illustrated below, with the addition of figures and graphs. We take the Stanford dataset as a base and begin the preprocessing of the Iraqi dataset. The Pima dataset required fewer preprocessing steps, so in the next subsections we will discuss only the Iraqi dataset.

Renaming Variables

We transform the column names to lowercase for uniformity and ease of use.

```
names(iraqi)[1:13] <- tolower(names(iraqi)[1:13])
names(iraqi)[3] <- "sex"
```

Listing 1: Renaming Variables

Data Normalization

The Iraqi dataset is ordered and normalized to match the Stanford dataset:

```
iraqi_o <- iraqi[c("age", "bmi", "chol", "ldl","hdl"
   ,"vldl")]
iraqi_o <- data.frame(scale(iraqi_o))
iraqi_o <- cbind(iraqi_o["age"], c(iraqi["sex"]),
                 c(iraqi_o[c("bmi", "chol", "ldl","
                    hdl","vldl")]), iraqi["CLASS"])
```

Listing 2: Data Normalization

Here, “age”, “bmi”, “chol”, “ldl”, “hdl”, and “vldl” are normalized using the `scale` function. The normalized data is then combined with the original “age”, “sex”, and “CLASS” columns to form the preprocessed dataset.

Checking and Removing Duplicates

The dataset contains duplicated entries, which can distort the analysis. These duplicates are identified and removed:

```
iraqi_d <- iraqi[, !(names(iraqi) %in% c("id", "no_pation"))]
iraqi_o <- iraqi_o[!duplicated(iraqi_o), ]
```

Listing 3: Removing Duplicates

It is important to note that the Stanford and Pima datasets were also checked for duplicates, and none were found.

Handling Categorical Variables

In this step, we handle the categorical variables in the dataset. Specifically, we use male individuals as the base group for the “sex” variable and make adjustments to the “CLASS” variable. These adjustments come from the fact that there are different spacing and caps choices for the same variable.

```
    library(plyr)
    iraqi_o$sex <- revalue(iraqi_o$sex, c("F"=1))
    iraqi_o$sex <- revalue(iraqi_o$sex, c("f"=1))
    iraqi_o$sex <- revalue(iraqi_o$sex, c("M"=0))

    iraqi_o$CLASS <- revalue(iraqi_o$CLASS, c("Y "="Y"))
    iraqi_o$CLASS <- revalue(iraqi_o$CLASS, c("N "="N"))
```

Listing 4: Handling Categorical Variables

Exploring Categorical Variables

Finally, we explore the categorical variables in the preprocessed dataset using basic plotting techniques. The code snippet below demonstrates the explo-

ration of the “sex” and “CLASS” variables:

```
library(ggplot2)
library(plyr)

ggplot(iraqi_o, aes(x=reorder(sex , sex,
   function(x)-length(x)))) +
geom_bar(fill='red') + labs(x='sex')

ggplot(iraqi_o, aes(x=reorder(CLASS , CLASS,
   function(x)-length(x)))) +
geom_bar(fill='blue') + labs(x='class')
```

Listing 5: Exploring Categorical Variables

For all the plots we refer to the notebook file cited in the supplementary material.