Understanding Diabetes Datasets: A Comparative Analysis

Daniele Lotito (PhD AI)

Presentation for the courses SLLD 1 and SLLD 2

Motivation and goals

Through a data-driven approach, we explore three datasets about diabetes and aim at

1. Identifying differences in the datasets, only looking at the data and at the information made available by the data collectors
2. Applying different statistical methods to all the datasets in order to highlight differences between them
3. Warning about potential misuses of the datasets

Datasets used

Three diabetes datasets:

- Stanford dataset - from the “Least angle regression” paper by Hastie and co-authors
- Pima Indian dataset - collected by (Indian) National Institute of Diabetes and Digestive and Kidney Diseases
- Iraqi society dataset - acquired from the laboratory of Medical City Hospital of Baghdad

Sources:
Stanford. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression, Dataset source on Stanford Edu
Pima Indian. (2016) Dataset source on Kaggle
Iraqi society. (2020)Dataset source on Mendeley

Stanford Dataset

In the dataset, the variables include:

- Age
- Sex
- Body mass index
- Average blood pressure
- Six blood serum measurements
- y (a measure of disease progression one year after baseline)

These variables pertain to 442 diabetic patients. All variables except y are mean-centered and scaled by the standard deviation times the square root of the dataset size.

Source: Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression.

Iraqi Society Dataset

The data were collected from the Iraqi society. Specifically, it includes:

- Age
- Gender
- Body Mass Index (BMI)
- Blood Sugar Level
- Fasting Lipid Profile (including LDL, VLDL, Triglycerides(TG), and HDL Cholesterol)
- Diabetes Disease Class (Diabetic, Non-Diabetic, or Predict-Diabetic)

The dataset has data from 103 non-diabetic, 53 predicted-diabetic, and 844 diabetic patients.

Source: Iraqi society, acquired from the laboratory of Medical City Hospital of Baghdad and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital (2020). Dataset on Mendeley

Pima Indians Dataset

The Pima dataset is related to women of Pima ethnic group. Its variables includes:

- Age: Age (years)
- BMI: Body mass index (weight in kg/m^2)
- Glucose: Plasma glucose concentration 2 hours after an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- Insulin: 2-Hours serum insulin (mu U/ml)
- Outcome: Class variable (0 is non-diabetic and 1 diabetic)

Source: Pima Indians Dataset (2016). Dataset on Kaggle

Preliminary observations and background

We soon discover that	There are only few common features
	The meaning of some features is not properly explained
	The response variables are defined differently
	The task associated with each dataset is different

In this presentation

We discuss two different positive messages that have emerged from the analysis.

1. Diabetes presence positively correlates with Age, but the disease progression is mainly determined by other factors that we can address with a healthy lifestyle
2. Predict-diabetic patients are more similar to non-diabetic people. They are perfectly in time to stop the disease progression

Data preprocessing

- We rename and order the features columns in order to deal with datasets of similar structure.
- We check for duplicates in the datasets.
- For each dataset we select a subset of features, favouring the ones shared by different datasets and those easily available.
- We apply the same normalisation used in the Stanford dataset.
- We identify and remove the outliers.

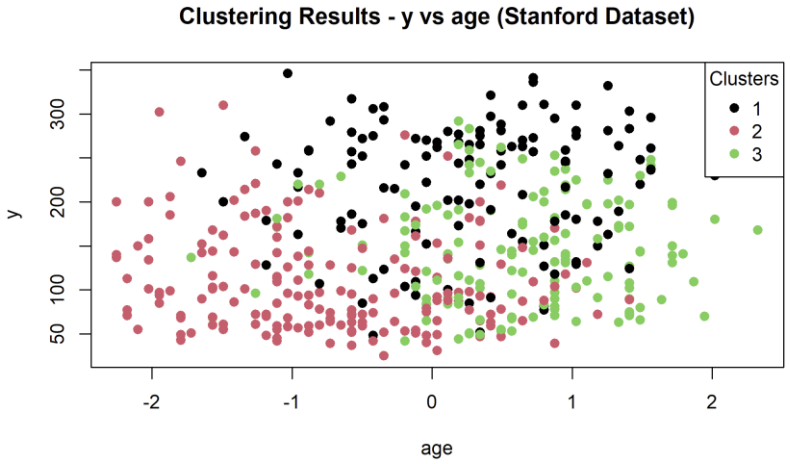
Common features between datasets

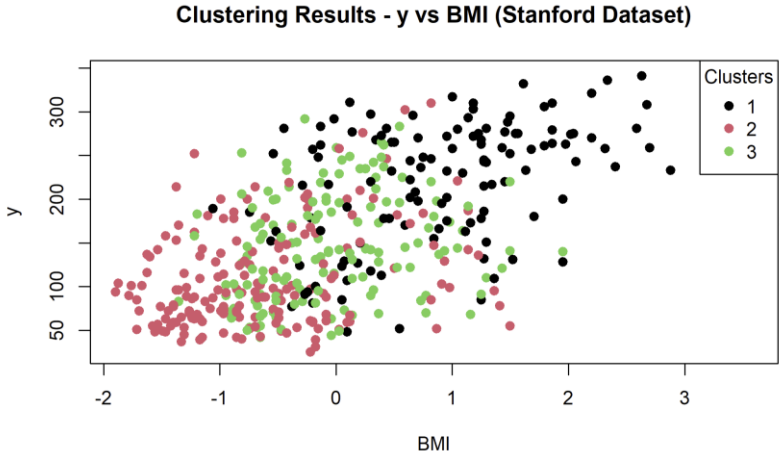
Feature	Stanford Dataset	Iraqi Dataset	Pima Dataset
LDL	Yes	Yes	No
HDL	Yes	Yes	No
Glucose *	Yes	No	Yes
Blood Pressure *	Yes	No	Yes
Age	Yes	Yes	Yes
BMI	Yes	Yes	Yes

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.

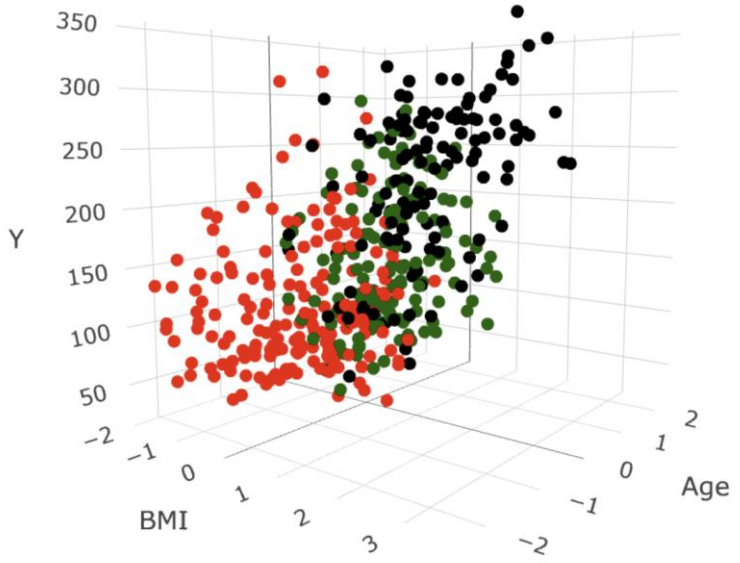
Assessing the relevance of age and bmi features

Stanford dataset



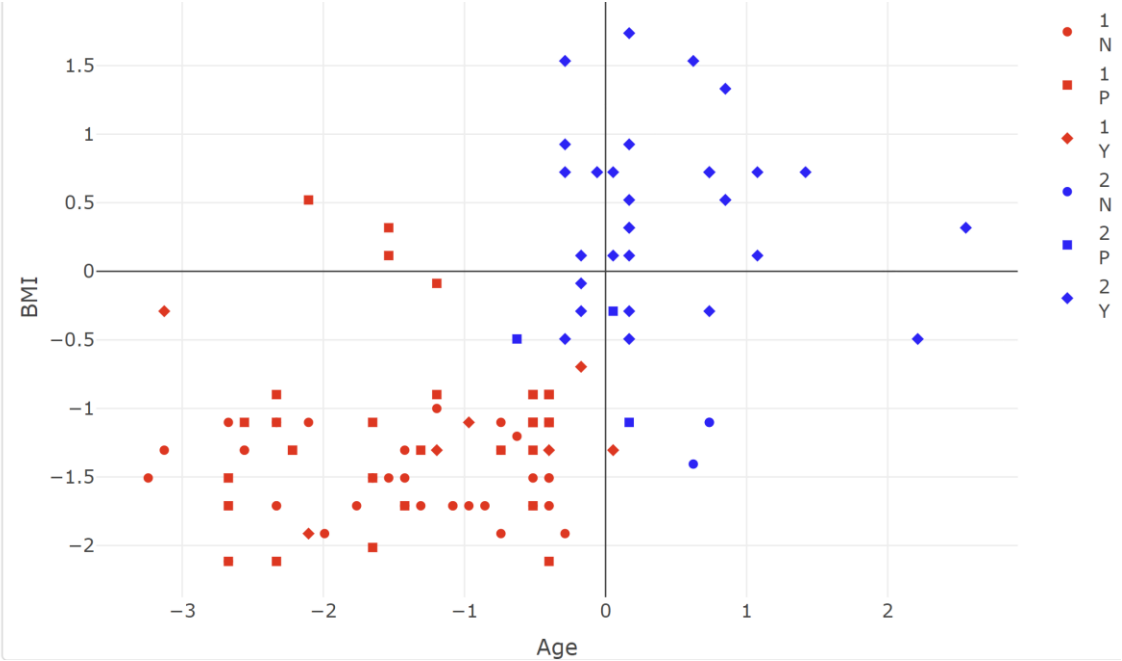


We run a 3-means clustering on "age", "bmi", “map", "ldl", “hdl", “tch", “ltg" and “glu“ features of the Stanford dataset.



Cluster analysis to assess the relevance of age and bmi features

Iraqi dataset



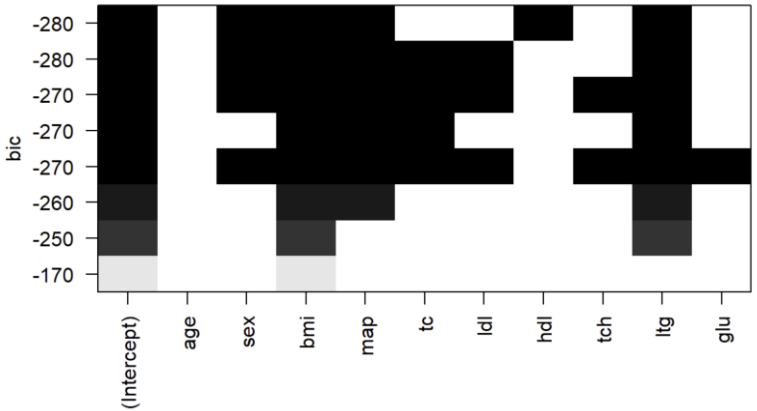
- Clusters on "age", "bmi", "chol", "ldl", "hdl".
- Clusters divide ‘N’ and ‘P’ from ‘Y’.
- Age and bmi equally important.
- Quite robust to seed change and different clustering techniques.

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

Is age more important in classification tasks?

Model selection in **Stanford**

```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         age sex bmi map tc  ldl hdl tch ltg glu
1  ( 1 ) " " " " "*" " " " " " " " " " " " " " "
2  ( 1 ) " " " " "*" " " " " " " " " " " "*" " "
3  ( 1 ) " " " " "*" "*" " " " " " " " " " "*" " "
4  ( 1 ) " " " " "*" "*" "*" " " " " " " " "*" " "
5  ( 1 ) " " "*" "*" "*" " " " " " " "*" " " "*" " "
6  ( 1 ) " " "*" "*" "*" "*" "*" " " " " " " "*" " "
7  ( 1 ) " " "*" "*" "*" "*" "*" " " " " "*" "*" " "
8  ( 1 ) " " "*" "*" "*" "*" "*" " " " " "*" "*" "*"
```



Model selection in **Pima**

```
bestglm(Xy = downsampled_pima, family = binomial, IC = "AIC")
```

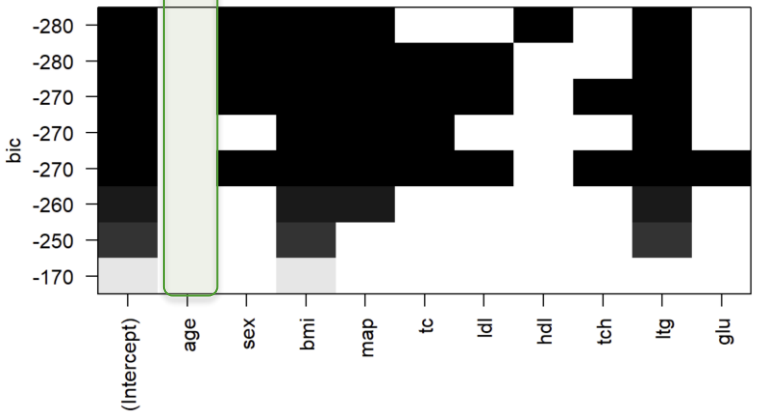
```
bglm.AIC_downsampled_pima$BestModels
```

	age	bmi	BloodP	glu1	insulin	Criterion
1	TRUE	TRUE	TRUE	TRUE	TRUE	527.4840
2	TRUE	TRUE	FALSE	TRUE	TRUE	528.3653
3	TRUE	TRUE	TRUE	TRUE	FALSE	528.9775
4	TRUE	TRUE	FALSE	TRUE	FALSE	529.1363
5	FALSE	TRUE	FALSE	TRUE	TRUE	538.2582
6	FALSE	TRUE	TRUE	TRUE	TRUE	539.7952
7	FALSE	TRUE	FALSE	TRUE	FALSE	540.7404

Is age more important in classification tasks?

Model selection in **Stanford**

```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
          age sex bmi map tc  ldl hdl tch ltg glu
1  ( 1 ) " " " " "*" " " " " " " " " " " " " " "
2  ( 1 ) " " " " "*" " " " " " " " " " " "*" " "
3  ( 1 ) " " " " "*" "*" " " " " " " " " "*" " "
4  ( 1 ) " " " " "*" "*" "*" " " " " " " "*" " "
5  ( 1 ) " " "*" "*" "*" " " " " "*" " " "*" " "
6  ( 1 ) " " "*" "*" "*" "*" "*" " " " " "*" " "
7  ( 1 ) " " "*" "*" "*" "*" "*" " " "*" "*" " "
8  ( 1 ) " " "*" "*" "*" "*" "*" " " "*" "*" "*"
```



Model selection in **Pima**

```
bestglm(Xy = downsampled_pima, family = binomial, IC = "AIC")
```

```
bglm.AIC_downsampled_pima$BestModels
```

	age	bmi	BloodP	glu1	insulin	Criterion
1	TRUE	TRUE	TRUE	TRUE	TRUE	527.4840
2	TRUE	TRUE	FALSE	TRUE	TRUE	528.3653
3	TRUE	TRUE	TRUE	TRUE	FALSE	528.9775
4	TRUE	TRUE	FALSE	TRUE	FALSE	529.1363
5	FALSE	TRUE	FALSE	TRUE	TRUE	538.2582
6	FALSE	TRUE	TRUE	TRUE	TRUE	539.7952
7	FALSE	TRUE	FALSE	TRUE	FALSE	540.7404

Is age more important in classification tasks?

Model selection in **Pima (2)**

```
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

Model selection in **Iraqi**

```
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

What I have not covered in this presentation

In the report and in the notebook are included:

- A detailed preprocessing of the data of all three datasets.
- A comprehensive PCA and cluster analysis (k-means and hierarchical) of the three dataset with dynamic data visualizations as well.
- The assessment of the relevance of common features in the related tasks, with particular attention to the ordered logistic regression and the discussion of its assumptions.
- Feature selection and model selection sections where I applied some of the techniques seen during the course.

The notebook consists of a fully reproducible analysis that can be exported to an html file, in the following days I will gather some feedback and publish it with an open source licence.

Future plans

I envision that what I did during the course could be the basis for a collaboration.

With the help of someone with medical domain knowledge on the topic this project can be developed and presented in a scientific paper.

We could consider also panel datasets and focus on a specific task (e.g. classification or disease progression) to assess the most relevant features for the chosen task.

References

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Harrell Jr, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- Prentice, R. L., & Pyke, R. (1979). "Logistic Disease Incidence Models and Case–Control Studies". *Biometrika*, *66*(3).
- Rajput, M. R., & Khedgikar, S. S. (2022). "Diabetes prediction and analysis using medical attributes: A Machine learning approach". *Journal of Xi'an University of Architecture & Technology*, *14*(1), 98-103.
- Rashid, A. (2020). "Diabetes dataset". *Mendeley Data*.
- Wickham, H., & Grolemund, G. (2017). *R for Data Science*. O'Reilly Media.

Data used

- Stanford. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression, Dataset source on Stanford Edu
- Pima Indian. Dataset source on Kaggle
- Iraqi society. Dataset source on Mendeley

Understanding Diabetes Datasets: A Comparative Analysis

AUTHOR
Daniele Lotito

Abstract

This report investigates three diabetes datasets: the Stanford dataset from the "Least angle regression" paper by Hastie et al., the Pima Indians dataset collected by the National Institute of Diabetes and Digestive and Kidney Diseases, and the Iraqi Society dataset. Despite drawing attention from the public, a simple exploratory analysis reveals issues such as sample heterogeneity, poorly explained variable meanings, and potential variations in the definition of response variables across the datasets. The motivation behind this study is to identify and understand these differences by solely examining the data and information provided by the data collectors. Various statistical methods are applied to the datasets to highlight disparities among potentially comparable data points. Moreover, the report aims to caution against potential misuses of these datasets. By undertaking this analysis, we hope to contribute to a clearer understanding of the nuances within the diabetes datasets and provide insights into their appropriate utilization.

Background

We explore three datasets about diabetes, all the datasets have drawn attention from the public, but even a simple exploratory analysis reveals that

Thank you

daniele.lotito@phd.unipi.it