Introduction to Machine Learning and Data Mining
Course no. 02450

Feature extraction and data visualization

Students:
Siddhanta Gupta - s210217
Hiten Chadha - s210208
Martin Blanck - s202281

October 28, 2022

Contents

Introduction

This document reports on the first assignment for the project on data visualization and analysis of the course 02450 Introduction to Machine Learning and Data Mining.
It aims to describe how a dataset can be handled and visualized using the content of the first part of the course. The dataset chosen is a medical dataset whose aim is to predict if a patient is likely to have diabetes. It is presented in details in the first and second sections of the report. This document further presents how the chosen dataset can be visualized and analyzed using vizulation techniques such as a principal component analysis (PCA).
The main purpose of the report is to get familiar with the dataset and to start thinking about how machine learning methods such as regression and classification can be performed in the second report later this semester. The major part of the code used within the range of this work is inspired from the scripts provided with the course 02450 Introduction to Machine Learning and Data Mining and downloaded from DTU Learn.

1 Description of the dataset

In the medical field, it is not seldom that a disease is diagnosed after the onset of symptoms that can potentially jeopardize the patient's health. Being able to predict if some patient is diseased before the onset of symptoms could potentially be a revolution in terms of treatments. On one hand, it could for example allow the beginning of a treatment early on to prevent the onset of symptoms. In other cases, it could be useful to stop or slow down the progression of some disease. In most diseases today, that cannot be performed. But machine learning tools could give new leads toward this objective.

The Pima Indians Diabetes Database is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The population from this study was the PIMA Indian population near Phoenix, Arizona. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. This dataset is a well validated data resource which can help to create a working training model and predict the date of onset of diabetes on future test data. The objective of this model is to forecast that whether an individual would develop diabetes mellitus within five years given the value of the eight diagnostic input variables.

The original dataset consists of 768 objects and 9 attributes (8 diagnostic input

independent variables and 1 outcome dependent target variable). The objects or instances contain valid data as well as corrupted data, outliers etc.

The description and numerical range of attributes are as given below:

Pregnancies - Number of times pregnant
Glucose - Plasma glucose concentration in mg/dl at 2 Hours in an Oral Glucose Tolerance Test (GTT)
Blood Pressure - Diastolic blood pressure in mm Hg
Skin Thickness - Triceps skin fold thickness in mm
Insulin - 2 Hour serum insulin in µU/mL
BMI - Body Mass Index in kg/m^2
Diabetes Pedigree Function (DPF) - DPF provides a measure of the expected genetic influence of the affected and unaffected relatives on the subject's eventual diabetes risk. DPF increases as the number of relatives who developed DM increases, as the age at which those relatives developed DM decreases, and as the percentage of genes that they share with the subject increases. Also notice that the value of the DPF decreases as the number of relatives who never developed DM increases, as their ages at their last examination increase, and as the percent of genes that they share with the subject increases.
Age – Age of the subject in years
Outcome – Class '0' or '1' based on whether the subject has developed diabetes mellitus in the next 5 years from the index examination.

The dataset was obtained from Kaggle database with a standard usability index of 8.8 (see [1]). The reference of the study paper for analysing the raw dataset is obtained from Kaggle (see [2]). The authors of the paper used 75% percent of the data to train the model (learning mode) and remaining 25% data for the test phase. An ADAP Learning Algorithm was designed from the data which gives a real number in the output. A cutoff value was chosen as the value to discriminate whether the person under test will develop diabetes or not in the next 5 years. With a cut-off value of 0.448 from ADAP, it was noted that the sensitivity and the specificity of the model is 76%.

In the range of this report, it is analyzed how classification and regression could be applied to this medical data set about diabetes prediction. It is not the purpose of this first report to perform the actual computing of classification and regression. But it is rather to analyze deeply the attributes and class, in order to give food for thought on what is available and can be done for the next report.
Naturally, the classification task will aim to classify whether a patient is diabetic or not based on the attributes described in the beginning of this section. In the range of this report, it can be done by applying a classification algorithm and then comparing the output to the true laboratory output.

On the other hand, the outcome of a regression applied to this data set is not obviously appearing relevant. It is hoped that it would be possible to predict crucial attributes such as glucose level, blood pressure or body mass index (BMI) based on the other attributes. In a medical point of view, it would be relevant to be able to predict such parameters as their measures can sometimes be uncomfortable for the patient (intravenous measure of blood glucose level for example). It would represent a gain of time and well-being for the patient. The BMI being a continuous variable, it would be possible to apply regression to predict its value for some patient depending on independent parameters such as insulin level, skin thickness, blood pressure, blood glucose level and if the person is diabetic or not.

2 Detailed explanation of the attributes

The attributes in the dataset are described in the table from figure 1.

Attribute	Range	Mean	Std. Deviation	Type
Pregnancies	0-17	3.85	3.37	Discrete/Ratio
Glucose	0-199	120.89	31.95	Continuous/Ratio
BloodPressure	0-122	69.11	19.34	Continuous/Ratio
SkinThickness	0-99	20.54	15.94	Continuous/Ratio
Insulin	0-846	79.80	115.17	Continuous/Ratio
BMI	0-67.1	31.99	7.88	Continuous/Ratio
Diabetes Pedigree Function	0.078-2.42	0.47	0.33	Continuous/Interval
Age	21-81	33.24	11.75	Discrete/Ratio

Figure 1: A brief explanation of the attributes

In order to investigate data issues, box-plots of each attributes are plotted on figure 2. It can be seen that some data attributes such as skin thickness, glucose concentration, blood pressure and BMI contain missing or "zero" value, which means the data must be corrupted as it is biologically not possible.

Abnormally low diastolic blood pressure values can also here be observed. According to [3], a blood pressure below 60 mm Hg is considered as low blood pressure. Therefore, it has been decided to exclude the observations with a blood pressure below 50 mm Hg to get rid of biologically unfeasible values or potentially corrupted data.

All such instances are filtered out of the original data set for creating a well-informed learning model. There are 392 instances remaining after filtering,

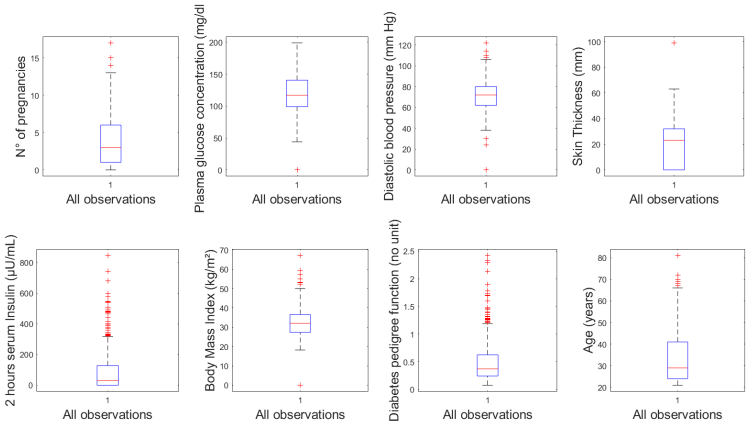

Figure 2: Boxplots of each of the attributes used to predict diabetes before pre-processing of the outliers and corrupted values

which still gives us a decent amount of observations for further study. In the following sections, all figures and explanations are related to the above described cleaned dataset. A summary of the final dataset used within this report is available in the table from figure 3.

Attribute	Range	Mean	Standard Deviation
Pregnancies	0-17	3.36	3.24
Glucose	56-199	123.03	31.06
BloodPressure	50-110	71.69	11.29
SkinThickness	7-99	29.16	10.49
Insulin	14-846	157.09	119.76
BMI	18.2-67.1	33.04	6.91
Diabetes Pedigree Function	0.078-2.42	0.52	0.34
Age	21-81	31.02	10.29

Figure 3: A brief explanation of the attributes (After removal of missing data and outliers)

3 Data visualization

3.1 General visualization of the data

A first visualization of the dataset was provided by the boxplots on figure 2. It allowed the identification and the further removal of outliers and corrupted data.

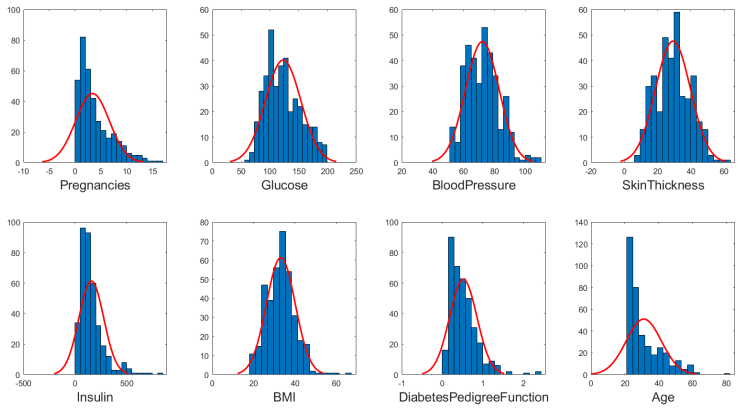

Figure 4: Histogram of each of the attributes used to predict diabetes after removal of outliers and corrupted values

To go further, figure 4 shows the histogram of each attribute in order to get a hint on their overall distribution. The red curve represents how a normal distribution would look like and is plotted to assess whether or not some attributes follow a normal distribution.

It looks like some of the attributes have a distribution resembling a normal distribution. For example for BMI, blood pressure, skin thickness or blood glucose level, the histograms almost fit the shape of the associated normal distribution. However, in a medical context such as the one of this dataset, that would not have had any significant interpretation and would not be relevant.

Further information are provided thanks to the correlation matrix on figure 5. It appears that some attributes are significantly correlated to others. This can be easily explained within the range of this dataset. For example, it is known that blood pressure is generally increasing with age. Blood pressure is higher as well for overweight people having higher BMI and skin thickness.
There is also a significant correlation between the glucose and insulin levels. It makes sense because patients with diabetes are likely to have abnormally high insulin levels [4]. Therefore, people with high insulin level are likely to be diabetes patients and thus to have higher glucose level.

Based on these visualizations, there should not be significant issues regarding the machine learning modeling. After removal of corrupted values and outliers, there is still a decent number of observations. Besides, the attributes are mainly continuous and ratio while the diabetes output is a binary attribute, leading to diverse pretty straight-forward possibilities for the next report in terms of

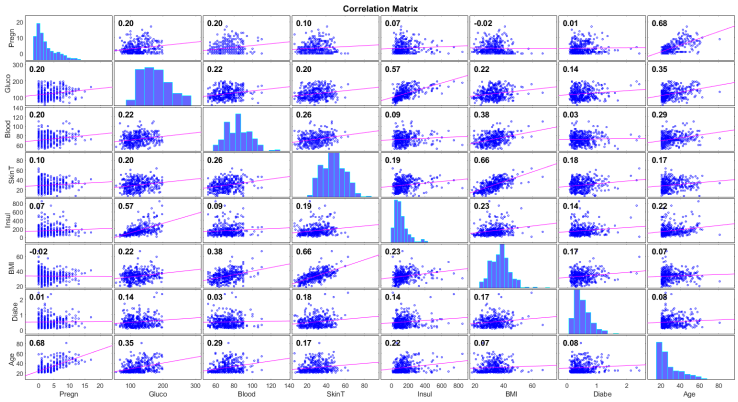

Figure 5: Matrix displaying the correlation between each attribute. The histograms are available on the diagonal.

classification and regression as explained in section 1.

3.2 Principal Component Analysis

Given the above-mentioned correlations between some of the variables and the high dimensionality of the data (8 dimensions), it is relevant to try to perform a principal component analysis (PCA) on this dataset. It has been computed in MATLAB following the code provided by the course content using the singular value decomposition. Before going through the PCA, the data has been standardized by subtracting the mean and dividing the standard deviation from each attribute. Indeed, it is necessary because the attributes have different scales as it can be seen on the table from figure 3.

The explained variance is plotted on figure 6. From that, it can be seen that 6 components are needed to account for 90% of the attributes' variations. Using 5 principal components accounts for 87% of the total variations.

For an easier representation, first five (5) principal components are chosen. They have been plotted one against another on figure 7.

No relevant conclusions can be drawn from these plots. It seems like a tendency is present with more non-diabetic patients for low value of the $PC1$ axis and more diabetic patients for high value on the $PC1$ axis (first line). But no distinct clusters of points are observed. It could be due to the fact that the attributes are not linked or that the relation between them is not linear.

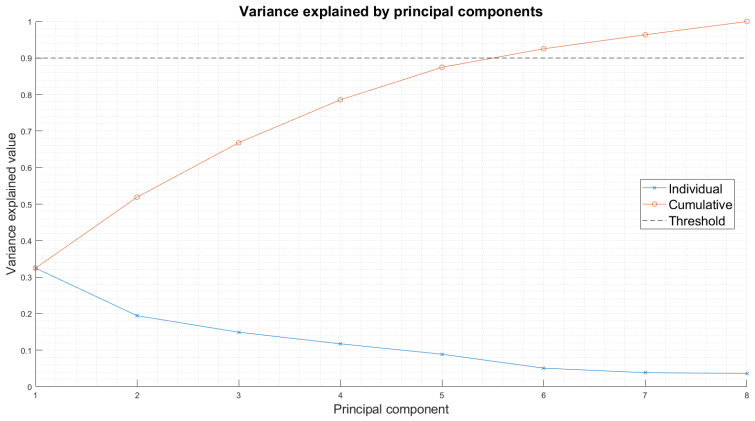

Figure 6: Variance explained of the principal components with a threshold set to 90%.

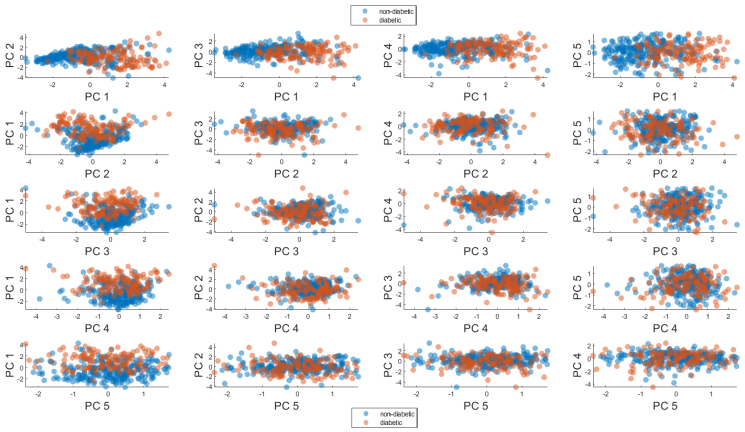

Figure 7: Projection of the data on the 5 principal components obtained via PCA

Furthermore, the directions of the considered PCA components are represented on the feature matrix on figure 8. It has been chosen to represent the Eigenvectors as a heat map to highlight the main coordinates of the vectors. As seen on the figure, the main component (PC1) has values that are relatively close to each other for all coordinates. It is noted that the line 7 corresponding to the Diabetes Pedigree Function (DPF) has a slightly smaller value. Looking at this map, the hypothesis is that the dataset is well spread almost equally across all its directions but the DPF.

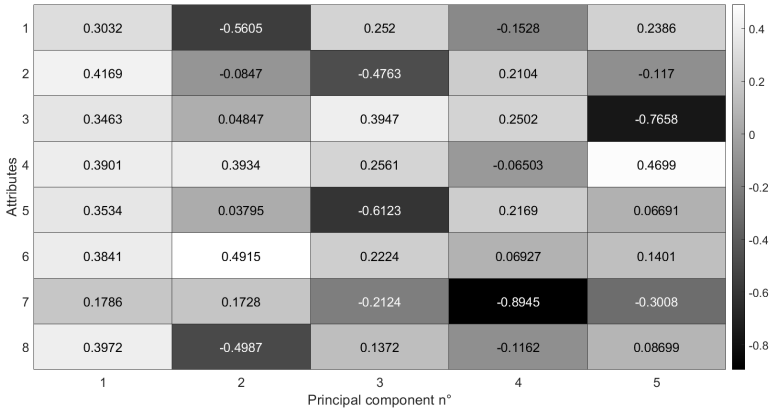

Figure 8: Value of the coordinate of the eigenvectors obtained through PCA

4 Discussion

This assignment's purpose was to choose a dataset, import it in MATLAB and pre-process it in order to prepare for further classification and regression tasks.

The dataset was first analyzed with the help of basic visualization tools such as boxplots and histograms in order to identify and subsequently removing plausible outliers and corrupted values. The data being measured is realized in a medical context, the identification of the outliers was made easier with the help of the standard biological ranges of the attributes.

Furthermore, the rectified dataset without these corrupted values, appears to be suitable for further machine learning processing.The dependent 'outcome' variable in this dataset is binary in nature which can be predicted through classification. Also, the attribute 'BMI' is continuous/ratio in nature which makes it easier to perform regression analysis. So it can be safely concluded that we can perform classification and regression algorithms easily on the rectified dataset.

Moreover, an attempt to reduce the dataset's dimensionality through PCA was performed. No conclusion could be drawn from this process, except that the variation of the data is not linear in nature, rather well spread across almost all the new dimensions. So, there is a risk of loss of information through PCA.

References

[1] Kaggle datasets, *Pima Indians Diabetes Database*, URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv, Accessed: 08-03-2021

[2] Source research paper by Smith, J.W., Everhart, J.E.Dickson, W.C.Knowler, W.C. Johannes, R.S. (1988). *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus*, Accessed: 08-03-2021

[3] MayoClinic website, *Low blood pressure (hypotension)*, URL: https://www.mayoclinic.org/diseases-conditions/low-blood-pressure/symptoms-causes/, 08-03-2021

[4] MayoClinic website, *Hyperinsulinemia: Is it diabetes?*, URL:https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/expert-answers/hyperinsulinemia/faq-20058488, Accessed: 06-03-2021

A Appendix

A.1 Diabetes Pedigree Function (quoted from [2])

$$DPF = \frac{\sum_i K_i(88 - ADM_i) + 20}{\sum_j K_j(ALC_j - 14) + 50}$$

Figure 9: DPF expression taken from [2]

The DPF was developed by the authors to provide a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject. It uses information from parents, grand parents, full and half siblings, full and half aunts and uncles, and first cousins.

The function is given by the above expression, where

i: ranges for all $relatives_i$ who had developed diabetes by the subject's examination date;
j: ranges over all $relatives_j$, who had NOT developed diabetes by the subject's examination date;
K_x: is the percent of genes shared by the $relative_x$ and
- equals 0.500 when the $relative_x$ is a parent or full sibling,
- equals 0.250 when the $relative_x$ is a half sibling, grandparent, aunt or uncle, and
- equals 0.125 when the $relative_x$ is a half aunt, half uncle or first cousin;
ADM_i: is the age in years of $relative_i$, when diabetes was diagnosed;
ACL_j: is the age in years of $relative_j$ at the last non-diabetic examination (prior to the subject's examination date);

Constants:
The constants 88 and 14 represent, with rare exception, the maximum and minimum ages at which relatives of the subjects in this study developed DM.
The constants 20 and 50 were chosen such that:-
1. A subject with no relatives would have a DPF value slightly lower than average
2. The DPF value would decrease relatively slowly as young relatives free of DM joined the database
3. The DPF value would increase relatively quickly as known relatives developed DM.

B Table of contribution

	Hiten	Siddhanta	Martin
Section 1	20%	60%	20%
Section 2	60%	20%	20%
Section 3	20%	20%	60%
Discussion	33%	34%	33%