Technical University of Denmark

ML Project 1 - PCA on Diabetes Dataset

02450 Introduction to Machine Learning and Data Mining

Lecturer: Bjørn Sand Jensen

Authors:

Nicholas Borch, Robin Braagaard, Peter V. Larsen

s234841, s234856, s234839

29^{th} of February 2024

1 Responsibility Distribution

Section	2	3	4	4.1	5
Borch	30%	30%	30%	40%	30%
Braagaard	40%	30%	30%	30%	40%
Larsen	30%	40%	40%	30%	30%

2 Data description

The primary aim of this dataset is to predict, with the use of various diagnostic measurements, whether or not diabetes will develop in a patient within the following five years. The dataset only contains female subjects above the age of 21 from a diabetes high-risk population, Pima Indians. None of the subjects were diagnosed with diabetes at the time of the study. If a patient received a diabetes diagnosis within the following year, the patient was removed from the data. A study of this specific group originates from the 1980s when the data from our dataset was collected.
The study was conducted by the Logistics Management Institute, the National Institute of Diabetes Digestive and Kidney Diseases, and The Johns Hopkins University School of Medicine. All institutes are situated in North America. The study collected eight variables which previously had been deemed significant risk factors for diabetes. The purpose of the study was to train a neural network learning algorithm called ADAP on 576 cases, and then aimed to predict diabetes onset in 192 test cases over five years. After determining a specific threshold from the output of the prediction, the algorithm proved to achieve a sensitivity and specificity of 76%. [2]
The dataset includes several medical predictor variables and a single dependent binary variable called "Outcome", which determines whether or not diabetes occurred in the patient within five years from the first examination.[3]

We seek to be able to gain insight into the relation between various medical attributes and the future onset of diabetes in the patients. We would especially like to gain knowledge of which combination of medical attributes that are associated strongest with the future onset of diabetes. With the use of Principal Component Analysis (PCA), we aim to reduce the dimensions of the dataset so that we more effectively can visualize the data by identifying the most relevant attributes.
In terms of using classification on our dataset, it makes sense to try to predict the binary outcome of the future onset of diabetes based on the various medical attributes.
In the regression analysis of our next report, the goal would be to predict the value of one of the continuous variables through the information we have from the remaining variables.

We will transform our entire dataset by standardizing the variables so they are all scaled similarly. This is done by subtracting the mean from every observation and dividing that by the standard deviation. We have executed this via Python but the formula used looks as follows: $z = \frac{x_i - \mu}{\sigma}$.

3 Explanation of Attributes

In the following table, a description of the attributes from the dataset is given:

Table 1: Attributes of the Dataset

Attribute	Description	Details	Unit of Measurement
Outcome	Binary indicator, '1' = Diabetes	Discrete, nominal	No Unit
Pregnancies	No. pregnancies	Discrete, ratio	No Unit
Age	Age of subject	Discrete, ratio	Years
Glucose	Glucose in OGTT	Continuous, ratio	mg/dL
Blood Pressure	Measured Blood Pressure of Subject	Continuous, ratio	mm Hg
Skin Thickness	Measured in triceps-area	Continuous, ratio	mm
Insulin	2-Hour Serum Insulin	Continuous, ratio	μU/ml
BMI	Body Mass Index of Subject	Continuous, ratio	kg/m^2

To conduct a PCA, rows with missing values (NaN) will be removed. Our dataset contains zeroes instead of missing values, so we have replaced the zeroes, where they are physically impossible to be present (eg. Blood Pressure), with the value NaN. However, this means that we aren't able to determine whether a zero in the "Pregnancies" attribute is a missing value or a legitimate zero. In our report we have chosen to keep all the values in the "Pregnancies" attribute as we deem that the missing values in the other categories are due to measurement faults whereas the number of times a subject has been pregnant more likely is a question that all the patients have been asked and can answer without undergoing any examination. Furthermore, in our PCA, we excluded the outcome variable to make sure that the analysis was purely based on the predictor variables, and only used the outcome data to illustrate the results of our analysis.

The following table shows a statistical summary of the attributes of our dataset:

Table 2: Summary of Data

Attribute	Min	Max	Mean	Median	Variance	Q1	Q3
Pregnancies	0.00	15.00	3.29	2.00	9.82	1.00	5.00
Glucose	56.00	198.00	121.68	119.00	920.29	99.00	142.00
Blood Pressure	24.00	110.00	70.56	70.00	151.12	62.00	78.00
Skin Thickness	7.00	60.00	28.89	29.00	107.29	21.00	36.00
Insulin	15.00	600.00	151.31	125.00	11200.92	76.50	190.00
BMI	18.20	57.30	32.84	33.10	43.88	28.35	36.90
Diabetes Pedigree Function	0.09	1.70	0.51	0.45	0.09	0.27	0.68
Age	21.00	63.00	30.68	27.00	96.47	23.00	36.00

Regarding glucose, any measurement below 140 mg/dL is considered normal. Measurements equal to or above 200 mg/dL indicate diabetes.[1][2] In our data, no measurements above 200 mg/dL are recorded, which supports the claim that none of the registered participants were indicated diabetics. The Diabetes Pedigree Function is calculated based on a series of genetic factors relating to diabetes (see Smith et al., 1988[2]).

4 Data Visualization and PCA

Before visualizing and analysing our dataset, we have chosen to perform outlier control in order for our analysis not to be affected by abnormal data points, and to reduce potential biases in the dataset. We have done this in an as objective way as possible, by removing the points that are more than 4 standard deviations away from the mean of each attribute to ensure effective removal of data points beyond the typical range of the data. This way extreme outliers are removed from the data without ruining its integrity.

In the following segment, a histogram, box plot and QQ-plot are visualized for each attribute in the dataset. By analysing these plots, we will be able to determine the extent to which each attribute follows a normal distribution.

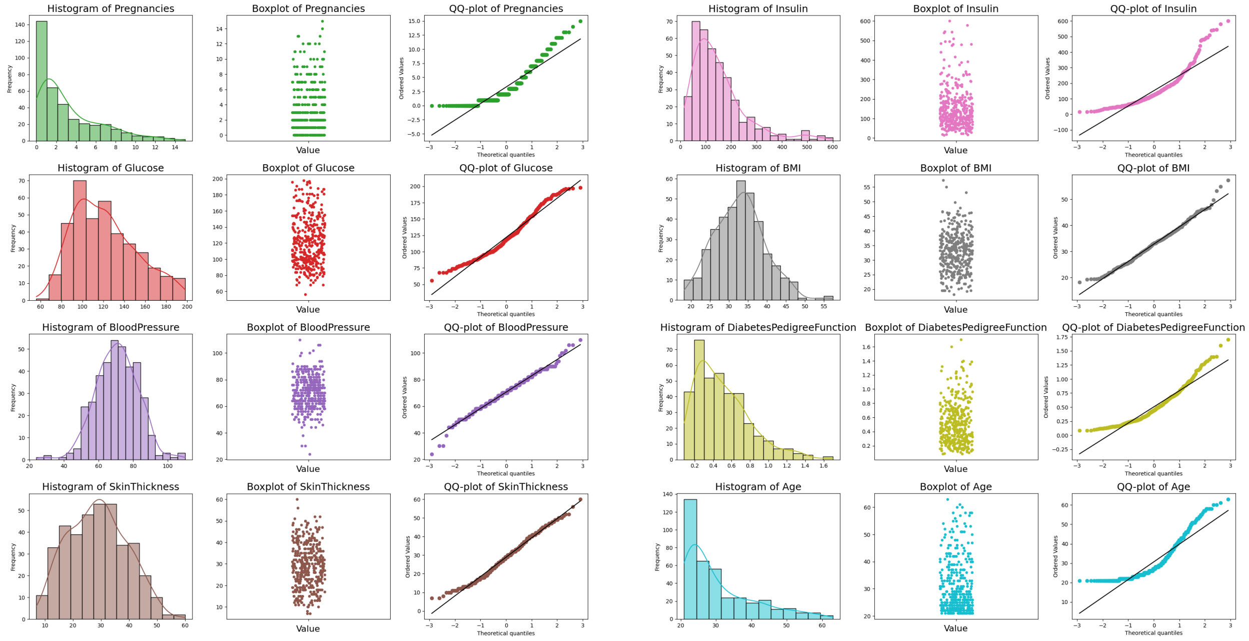


Figure 1: Histogram, box-plot, and QQ-plot of selected attributes.

By analysing the histograms and QQ plots, a few of the attributes look to follow a normal distribution relatively well. The "Blood Pressure" attribute especially seems to follow a normal distribution with only the tails of the distribution skewing the data a little. The "Skin Thickness" and "BMI" attributes also follow normal distributions relatively well, even though it is noticeable that both of the attributes are skewed a bit to the right. The rest of the attributes from our dataset seem not to be following a clear normal distribution.

In this next segment, paired scatter plots and a matrix plot have been executed to provide a visual overview of the attributes' correlation with one another. In the scatter plot, the orange dots represent the future onset of diabetes within the patient.

Figure 2: Paired Attributes Scatter Plots

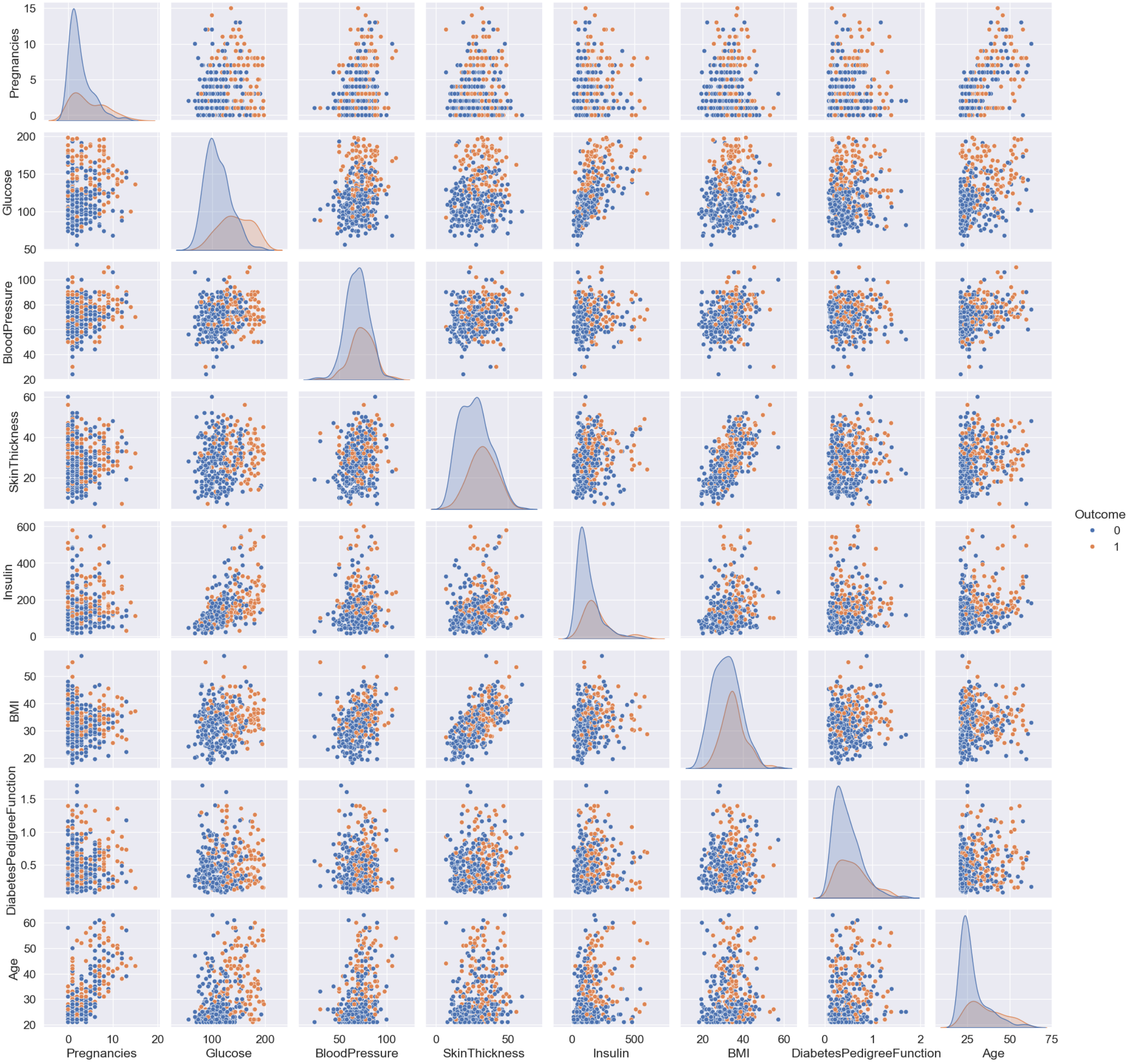
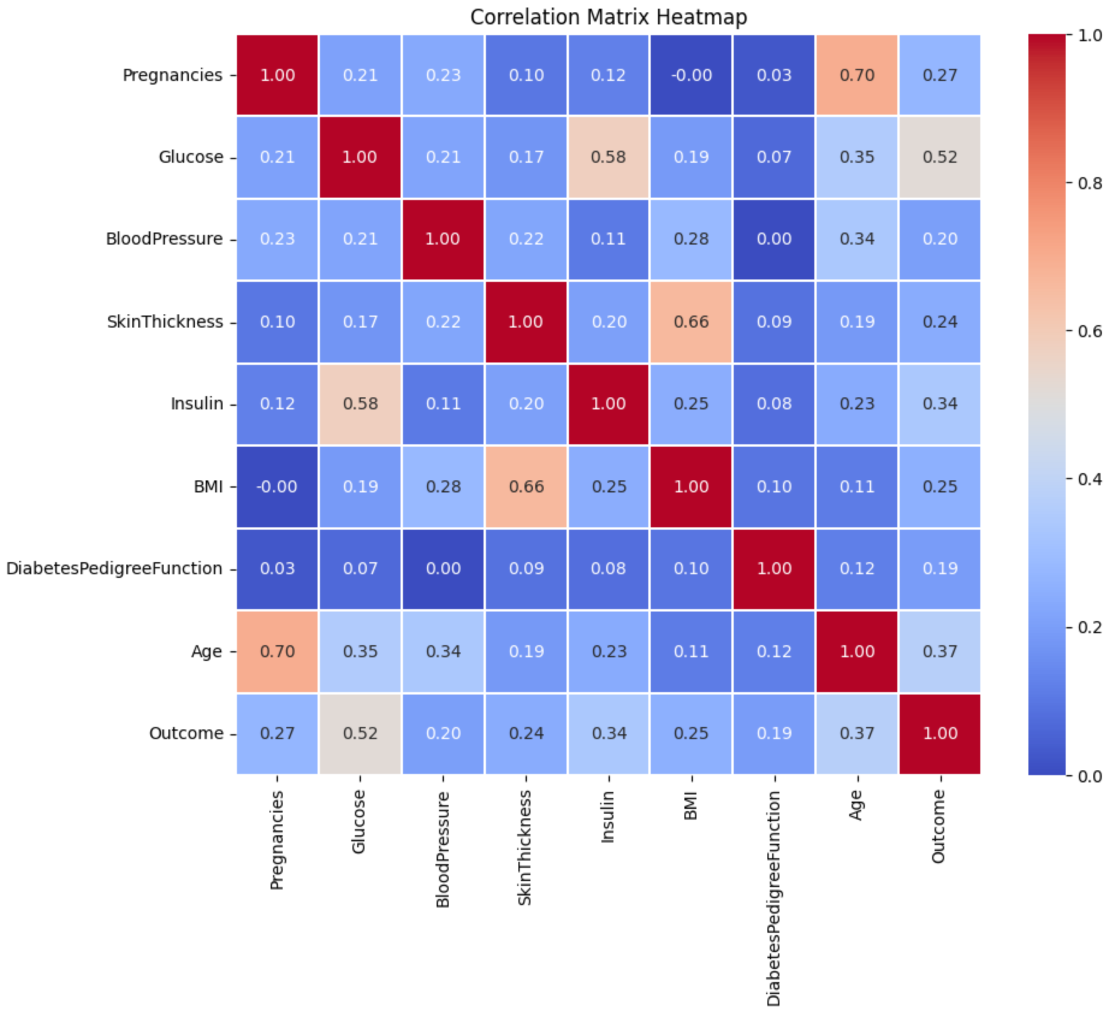


Figure 3:



When analysing the plots above, a correlation between a few of the attributes seems to be present. The most evident correlation is the one between a patient's age and their number of pregnancies which per the "Correlation Matrix Heatmap" has a positive correlational value of 0.70. The Glucose and Insulin levels of the patients also seem to correlate significantly positively; the same can be said about the "Skin Thickness" and "BMI" attributes. However, there doesn't seem to be any obvious correlation between the remaining attributes. When analysing the density plot and other scatter plots containing the "Glucose" attribute, it seems to split the data points more clearly than any of the other attributes. This knowledge instils a belief that the glucose levels of a patient might be a significant indicator of the future onset of diabetes. As seen in Figure 3, there is a noticeable positive correlation of 0.52 between outcome and glucose levels.

After extensive analysis, examination and visualization of our dataset, our primary machine learning modelling aim of predicting the future onset of diabetes seems feasible. Pairing the presence of a clear correlation between a selection of attributes with the fact that it is possible to observe a relatively normal distribution of some of the variables, instils confidence in the potential of being able to predict the onset of diabetes.

4.1 PCA

Following the introductory sections of our data, PCA was conducted. We found the following principal components, each represented as a linear combination of vectors forming the space wherein the observations lie (see Table 3). These components are orthonormal, meaning the information captured by each is uncorrelated with each other.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0.3468	0.5189	0.2645	-0.0412	0.3258	0.1287	0.0651	0.6417
0.4083	0.0608	-0.5319	0.1319	-0.1230	-0.6954	0.0886	0.1508
0.3361	0.0178	0.3428	0.2220	-0.8220	0.1236	-0.1479	0.0845
0.3632	-0.4775	0.2657	-0.0135	0.3373	-0.1750	-0.6518	0.0095
0.3691	-0.0718	-0.6158	0.0931	0.0431	0.6674	-0.1518	-0.0220
0.3515	-0.5514	0.2016	0.0202	0.1086	0.0860	0.7151	0.0291
0.1160	-0.0623	-0.0816	-0.9562	-0.2338	-0.0002	-0.0229	0.0810
0.4415	0.4309	0.1804	-0.0901	0.1450	-0.0346	0.0792	-0.7418

Table 3: Principal Components

Succeeding this, we calculated the principal components' explained variance ratio in Python and illustrated the distribution in Figure 4 (Left).

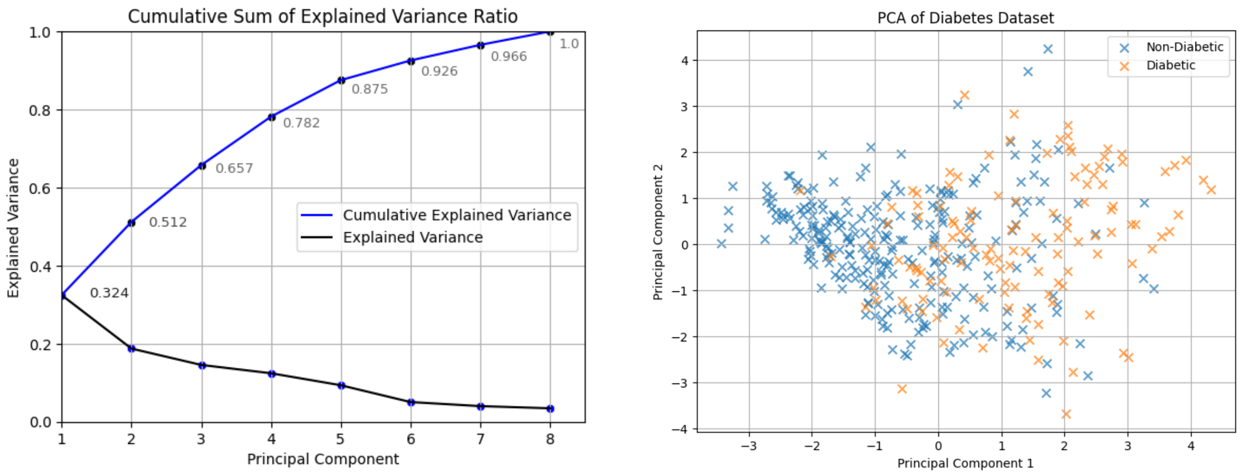


Figure 4: Left: Principal components' explained variance ratio; Right: Plot of the data projected onto PC1 and PC2.

As seen above, principal components 1 and 2 constitute a total of 51.2% of the explained variance in the dataset. PC1 with an explained variance of 32.4% and PC2 with an explained variance of 18.7%. It is also noticeable, that to obtain an explained variance of above 80% we would need to incorporate five principal components, meaning much of the dataset would be explainable in five dimensions instead of the eight that it includes.

When selecting the number of principal components to retain, we wanted to see what information the first two principal components contained, as a two-dimensional plot is a simple way to visualize data. The "PCA of Diabetes Dataset"-plot (Figure 4 (Right)) shows a partial separation of future onset of diabetes detected contra not detected. Even though it only explains 51.2% of the variance in the data, the plot can visualise some of the information from the eight-dimensional dataset in a two-dimensional space to a limited extent. The figure generally suggests that a positive PC1-value of an observation could be correlated to a future diabetes diagnosis.

Following this, we computed the first three principal components' coefficients as seen in Figure 5 on the next page. Our aim with including the third principal component was to uncover any additional insights related to the outcome that might not be clear in the initial plots. We decided not to create a three dimensional plot as these can pose interpretive challenges and for that reason chose to stick with the two dimensional plot representations.

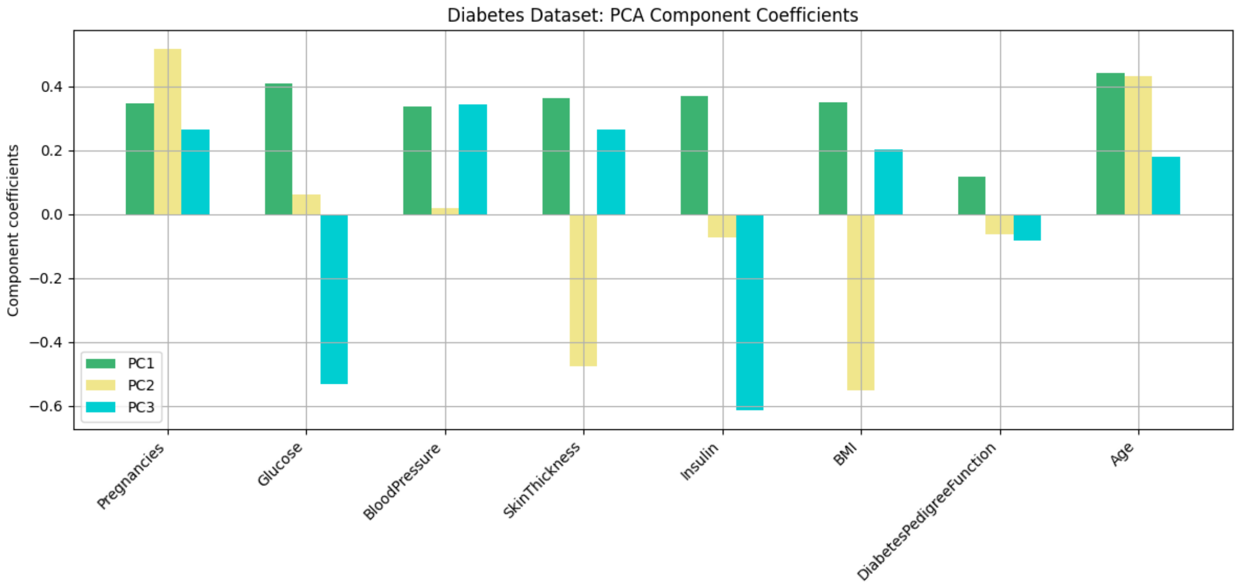


Figure 5: Plot of principal components 1, 2, and 3's coefficients

We also projected the observations onto each of the first three principal components to analyze their relation to diabetes onset. See Figure 6

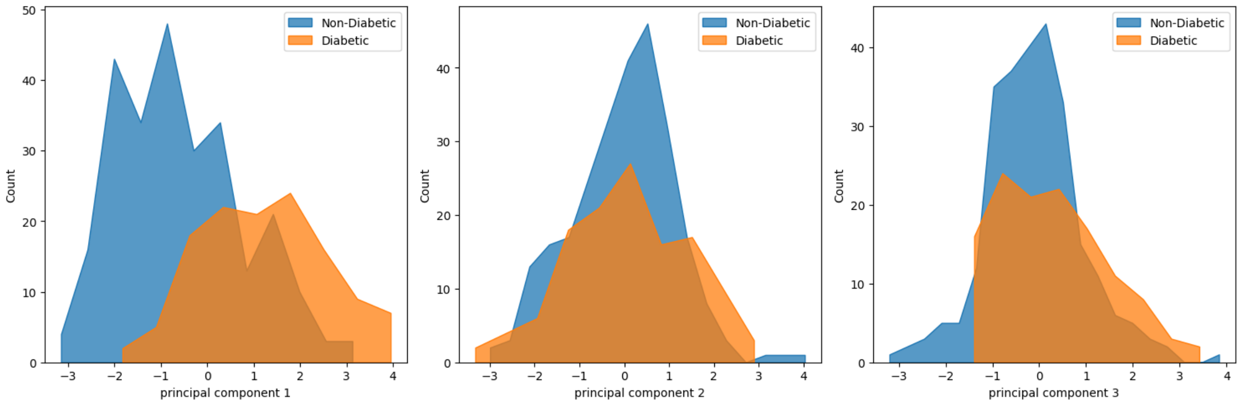


Figure 6: Density plots of projections onto PC1, PC2, PC3.

Figure 5 above shows which attributes contribute in which ways to the first three principal components. An analysis of the coefficients of PC1 (Figure 5) highlights that an observation of relatively high values for all attributes typically results in a positive value of the projection onto PC1. Which also can be seen in Figure 4: Right. As discussed earlier, an observation with a high PC1-value appears to be somewhat correlated with receiving a diabetes diagnosis within 5 years, which is supported by the separation of the density plots that can be seen in Figure 6. This result doesn't come as a surprise, as all the variables are known in the medical sphere, as being significant predictors of a diabetes diagnosis. Furthermore, the fact that all the values are aligned in the same positive manner also suggests that when one attribute increases for the patient the rest also tend to increase.

When analysing PC2 and PC3, some slightly more complex patterns appear. The pattern in PC2 suggests that a subject's age and pregnancies tend to increase/decrease together whilst having the inverse effect on a subject's BMI and skin thickness. This could indicate that a subject with a high number of pregnancies also tends to be of older age while having lower BMI and skin thickness measurements. However, the lack of any clear separation between the density plots seen in Figure 7 suggests that PC2 isn't a strong predictor of the onset of diabetes on its own.

PC3 appears to discriminate between people with a combination of relatively high values for all attributes, but low values of Glucose, Insulin and Diabetes Pedigree Function. However, similarly to PC2, the lack of any clear separation between the density plots for that principal component, makes it difficult to use PC3 to draw any clear conclusions about the future onset of diabetes.

5 Discussion

Conducting statistical summary and detailed attributes analysis on the dataset, containing several medical predictor variables, we gained an understanding of the data and were able to showcase compelling information of a multi-dimensional dataset. However, the dataset also had its limitations, such as the replacement of missing values with zeroes, leading us to make assumptions on which attributes a zero could be viable and when it would be physically unrealistic.

A PCA analysis was conducted and the first two principal components with a cumulative explained variance of 51.2% were able to illustrate a partial tendency in the data and separate people who in the future would receive a diabetes diagnosis and who would not. Although not a comprehensive distribution, it suggests that a significant amount of information can be compressed into fewer dimensions without crucial loss.

The discovery of this tendency makes our machine learning objective, to be able to predict the future onset of diabetes, seem feasible. We know from the 1980s study that it could be done with sensitivity and specificity of 76% and with today's technological advancement we believe we can achieve an accuracy even better.

In conclusion, we were able to perform Principal Component Analysis on our dataset, revealing several interesting complex patterns. So far, we have not been able to find any methods to unambiguously separate individuals based on attributes, but we suspect that a further analysis utilizing machine learning will allow us to extract more information.

References

[1] Cathi J. Swift. Emily Eyth, Hajira Basit. Glucose tolerance test. `https://www.ncbi.nlm.nih.gov/books/NBK532915/`, 2023,
[Visited 27th of February 2024].

[2] WC Dickson WC Knowler RS Johannes Jack W. Smith, JE Everhart. Using the adap learning algorithm to forecast the onset of diabetes mellitus. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf`, 1988,
[Visited 26th of February 2024].

[3] Akshay Dattatray Khar. Diabetes dataset. `https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data`, 2023,
[Visited 26th of February 2024].

6 Problems

6.1 Question 1 Answer:

Nominal is defined as "If the variable is not ordered and only uniqueness matters.". Therefore, (Time of day) is Nominal. Ratio is defined as "If the value 0 of the variable has a specific, physical meaning. I.e. it makes sense to say one value of the variable is “twice as large” as another.". Therefore (Traffic lights) and (Running over) is ratio. Ordinal is defined as "If the variable is ordered (smaller, larger).". Therefore (Congestion Level) is ordinal.

The answer for Question 1 is A.

6.2 Question 2 Answer:

Given: $x_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad x_{18} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$

The p-norm distance is given by

$$d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p = \begin{cases} \left(\sum_{i=1}^{M} |x_i - y_i|^p \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty \\ \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_M - y_M|\} & \text{if } p = \infty. \end{cases}$$

We check for the first case $d_{p=\infty}(\mathbf{x_{14}}, \mathbf{x_{18}})$:

$$d_{p=\infty}(\mathbf{x_{14}}, \mathbf{x_{18}}) = \max\{|26 - 19|, |0 - 0|, |2 - 0|, |0 - 0|, |0 - 0|, |0 - 0|, |0 - 0|\} = 7$$

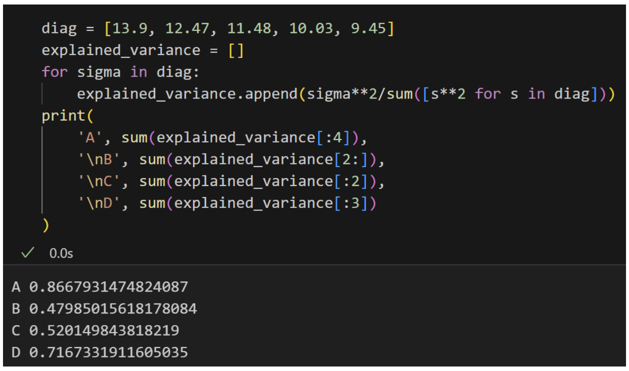
The answer for Question 2 is A.

6.3 Question 3 Answer:

The following code was written to equation 3.18 from the lecture notes:

$$\text{Explained Variance} = \frac{\Sigma_{i=1}^{n} \sigma_i^2}{\Sigma_{i=1}^{M} \sigma_i^2}$$

We used the equation to calculate the sum of explained variance of the first four PCs, the last three PCs, etc. We see that the only correct answer would be A; The variance explained by the first four principal components is greater than 0.8.



```
diag = [13.9, 12.47, 11.48, 10.03, 9.45]
explained_variance = []
for sigma in diag:
    explained_variance.append(sigma**2/sum([s**2 for s in diag]))
print(
    'A', sum(explained_variance[:4]),
    '\nB', sum(explained_variance[2:]),
    '\nC', sum(explained_variance[:2]),
    '\nD', sum(explained_variance[:3])
)
```

```
A 0.8667931474824087
B 0.47985015618178084
C 0.520149843818219
D 0.7167331911605035
```

Figure 7: Python Code used for Calculations

The answer for Question 3 is A.

6.4 Question 4 Answer:

When you standardize data, lower values will be negative and higher values will be positive. When looking at the matrix $\mathbf{V}$, we see that the correct answer for Question 4 is D, since ”An observation with a low value of Time of day, a high value of Broken Truck, a high value of Accident victim, and a high value of Defects will typically have a positive value of the projection onto principal component number 2.” **The correct answer to the Question 4 is therefore D.**

6.5 Question 5 Answer:

The Jaccard similarity is defined as the intersection of two sets of words divided by the union of two sets of words: $J(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|}$.
The length of the intersection is 2 and the length of the union is 13, resulting in a calculation that looks as follows: $J(s_1, s_2) = \frac{2}{13} = 0.153846$.
The correct answer to Question 5 is therefore A.