Danmarks Tekniske Universitet

Introduction to Machine Learning and Data Mining

Project 1

02450, Spring 2024

Group 32

Nikolaj Jakobsen - s234818
David Lindahl - s234817
Mikkel Nielsen Broch-Lips - s234860

	Section 1	**Section 2**	**Section 3**	**Section 4**
David	60%	20%	20%	40%
Mikkel	20%	60%	20%	40%
Nikolaj	20%	20%	60%	20%

Table 1: Completion Status of Sections by Authors

1 Description of the data set

This paper will be working with the dataset 'Physical Characteristics of Urine With and Without Crystals'. The contributor of the data is 'National Cancer Institute'. The data was recorded in 1985 and has since been made open source. The dataset is available via kaggle.com [1]

The overall problem of interest of our dataset is to explore whether we can predict the presence of kidney stones in a urine sample. Additionally, we seek to predict calcium concentration based on urea concentration using regression analysis.

The data contains 79 urine specimens, given in Table 2. These were analyzed in an effort to determine if certain physical characteristics of the urine might be related to the formation of calcium oxalate crystals (kidney stones). Every sample has been gathered in the laboratory of James S. Elliot M.D, Stanford University School of Medicine.

Patient number	Specific gravity	pH	mOsm	mMho	Urea	Calcium	Target
1	1.021	4.91	725	14.0	443	2.45	0
2	1.017	5.74	577	20.0	296	4.49	0
3	1.008	7.20	321	14.9	101	2.36	0
...	...	...	...	...	...	...	...
79	1.015	6.03	416	12.8	178	9.39	1

Table 2: Selected patient urine analysis data

Our dataset has not been cited in other scientific literature. However, in 2017 computer engineers from University of Guilan, Iran made a study predicting the presence of kidney stones in urine samples via. machine learning. They achieved a classification accuracy of [95.9% ; 99.6%]. They achieved this with a dataset with $N = 936$ samples and $M = 42$ features. [3]

Although there are no scientific papers citing our data, many users on kaggle.com have tried to make an accurate classifier, that can predict if a sample has kidney stones or not. One such user is 'Ganesh Jainarain'.. Using a XGBClassifier, he achieved a classification accuracy of 78.31%. The XGBClassifier works by training a series of decision trees and optimizing a specified loss function. [2]

The goal of our project is:

- Make an accurate classifier, that can predict if a sample has kidney stones based on the 6 attributes.
- Make a regression model, that can predict the calcium concentration in a urine sample, given the urea concentration.

This last section of the introduction will be discussing, how we decided, which attributes we want to predict via regression.

When choosing which attribute we want to predict, there are 2 conditions that needs to be met:

1. Correlation: For regression to be successful, and to make a model with low error, it helps to have correlated data. If the data has a high correlation, a linear plot will generally have less error, and the regression will be more successful.
2. Relevance: Being able to predict a persons hair color based in the urine's specific gravity would likely have little to no relevance. This illustrates the importance of selecting attributes with meaningful relevance to the problem domain, as predictive models are most valuable when they provide insights or solutions relevant to the target problem.

With these two conditions in mind, we decided have our regression model predict the concentration of calcium based off the concentration of urea. The concentration of calcium is a useful attribute, since low levels of calcium intake has been linked to diseases such as Osteoporosis. The correlation between urea and calcium can be seen in figure 2 to be 0.5. This should make for a somewhat accurate model. In figure 1, we can also see, that the data seems somewhat correlated.

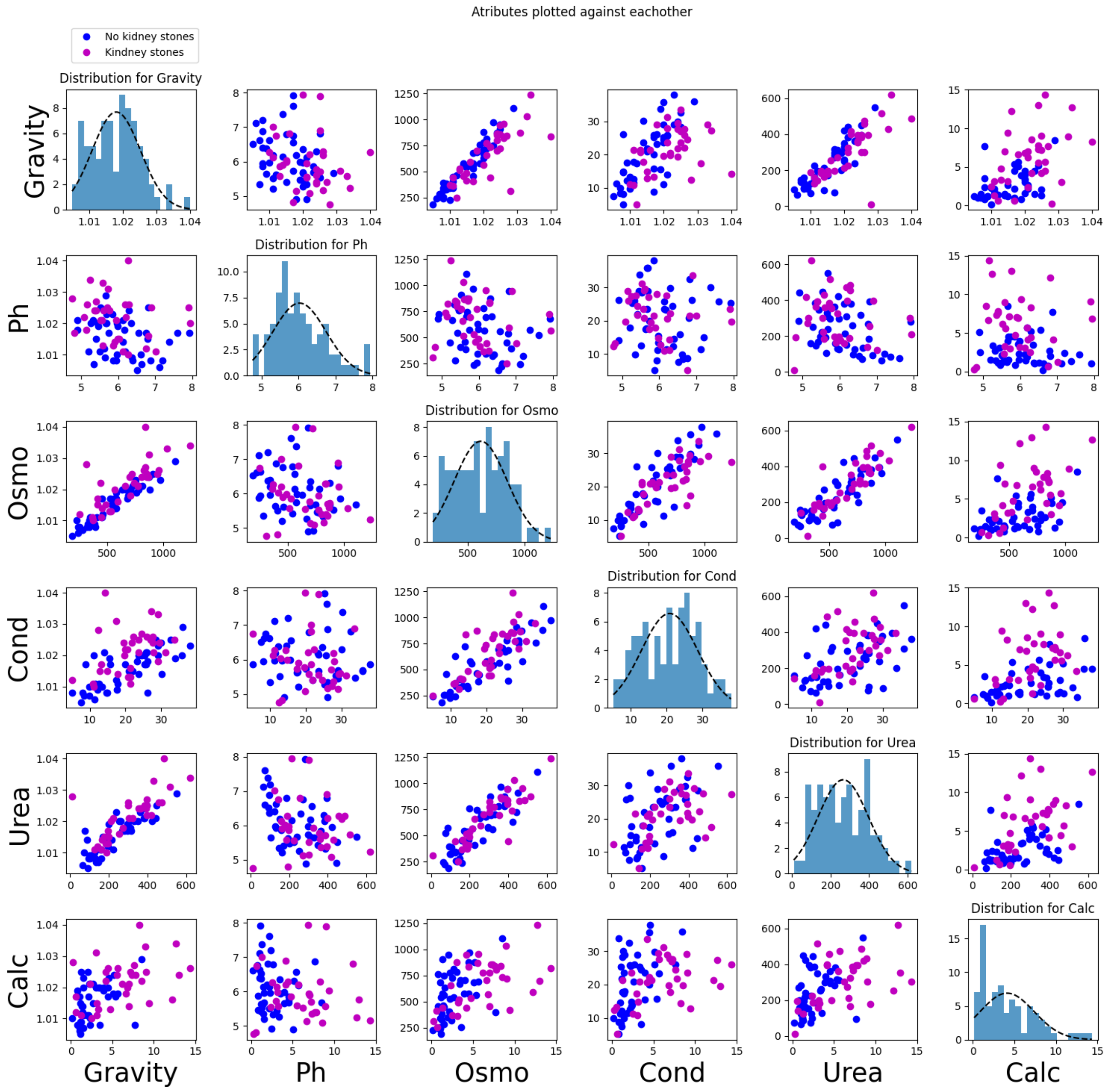

Figure 1: '$m \times m$' pairplot of each attribute.

Given the small size of our dataset, we plan to explore the performance of machine learning models using both the original and normalized data. This approach will allow us to determine which data preparation technique yields better accuracy in classification tasks. Normalizing the data could potentially improve model performance by adjusting the scale of the features, making the model less sensitive to the scale of attributes. This comparison is crucial for identifying the most effective data preprocessing method for our specific dataset and achieving higher classification accuracy.

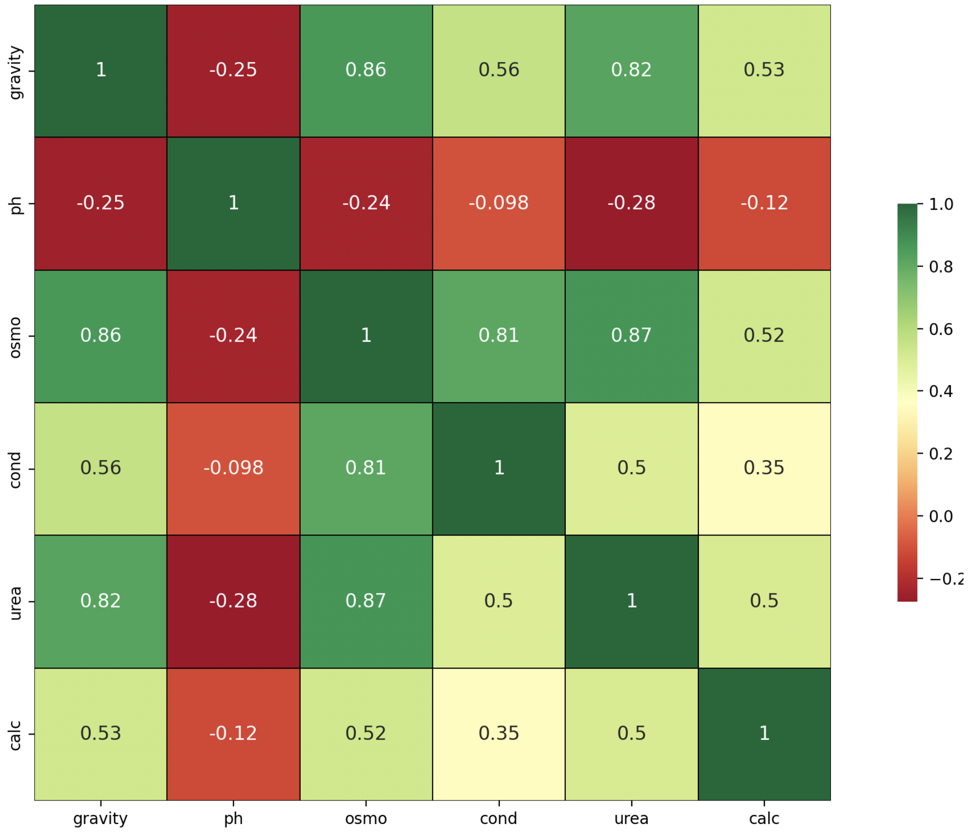

Figure 2: $m \times m$ Each attribute's correlation

2 Attributes

There are 79 observations for each 7 attributes. The data is paired and is extracted from urine samples and tests of 79 patients. There are no missing values or obvious issues with the data.

1 Specific gravity of urine
The specific gravity of urine is a measure that compares the density of urine to the density of water. This attribute is a continuous ratio, because it has a meaningful zero point. The specific gravity of 1 corresponds to the density of pure water.

2 pH of urine
The pH of urine is a measure in the interval 1 to 14 where 7 is neutral, ¿7 is base, ¡7 is acid. This is a continuous and interval attribute. It's not a ratio as the attribute doesn't have a true zero point.

3 Osmolarity of urine
Osmolarity is measured in milliosmoles per liter (mOsm/L) and is a measure of the concentration of solutes in urine. It's a continuous and ratio attribute.

4 Conductivity of urine
The conductivity of urine is a measure of the ability of urine to conduct electricity. It's measured in Siemens per meter (S/m) and is a continuous and ratio attribute.

5 Concentration of urea in urine
The concentration of urea in urine is a measure of the amount of urea nitrogen present in a urine sample. It's a continuous and ratio attribute.

6 Concentration of calcium in urine
The concentration of calcium in urine is a measure of the amount of the mineral calcium present in a urine sample. It's a continuous and ratio attribute.

7 Target
This attribute is either 0 for absence of kidney stone or 1 for presence of kidney stone. It's therefore discrete

and nominal. This attribute is central as it holds the answer to the main question of this study: Is there a correlation between features from the urine samples and the presence of kidney stones. This attribute will later be used as y, when training a model for classification.
As of this and the fact that it's a binary attribute, we will treat this attribute as a class index.

The main statistics for each attribute are can be examined in Table 3.

	Gravity	pH	Osmo	Cond	Urea	Calc
Mean	1.018	6.028	612.848	20.814	266.405	4.139
Std	0.007	0.724	237.515	7.939	131.255	3.260
Min	1.005	4.760	187.000	5.100	10.000	0.170
25%	1.012	5.530	413.000	14.150	160.000	1.460
50%	1.018	5.940	594.000	21.400	260.000	3.160
75%	1.023	6.385	792.000	26.550	372.000	5.930
Max	1.040	7.940	1236.000	38.000	620.000	14.340

Table 3: Summary statistics of urine and kidney stones data.

To visualize the data we made box plots for each attribute. Be aware that the box plots are standardized for better comparison.

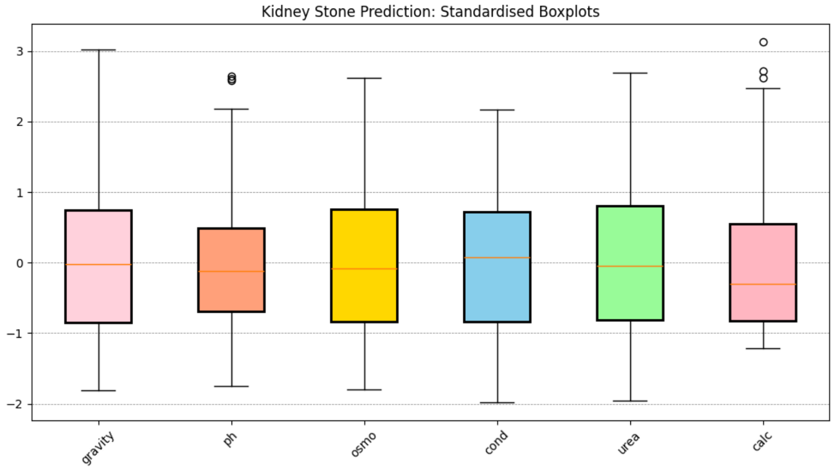

Figure 3: Standardized Boxplot

There is a small trend for in the box plots (fig. 3) . They all seem to have a longer gab between the the "Max" and the upper quantile than between "Min" and the lower quantile (except "cond"). This indicates that the distributions of the attributes is skewed with a longer tail towards the maximums. This trend is most clear for "calc".

Looking at the histograms it becomes clear that the trend indicated by the box plots is not as strong as first interpreted. For "calc" it's still clear and it actually seems like it could have something like a χ^2 distribution. The other attributes seem to weakly follow normal distributions.

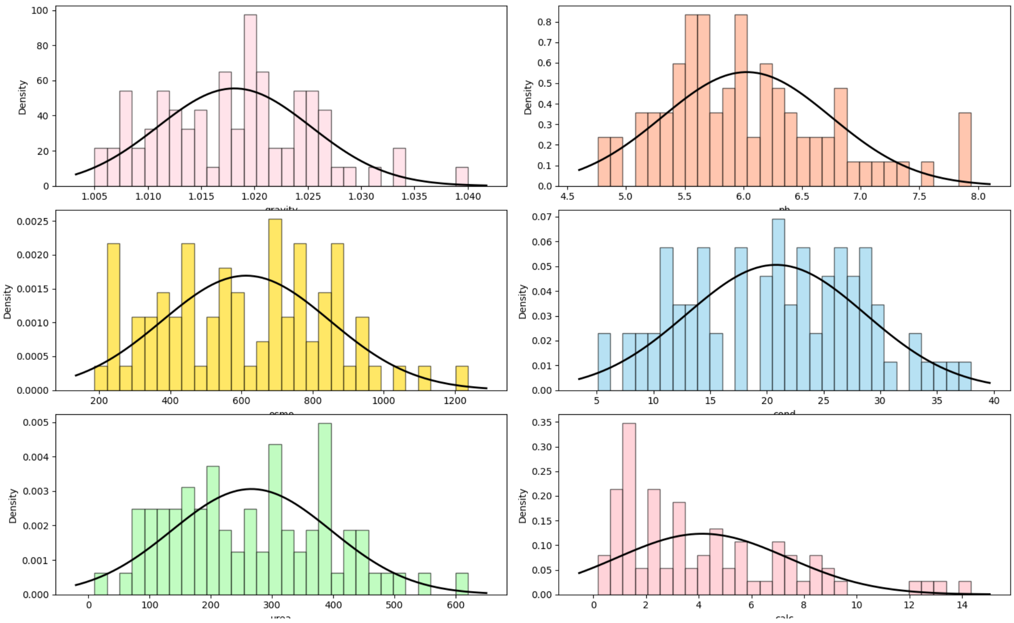

Figure 4: Histograms

3 Data Visualization

3.1 Correlation, Distribution, Outliers, and Modeling

As seen in Figure 2, there is a notable correlation between some of the attributes, for instance, osmolarity and urea have a correlation coefficient of 0.87. This observation is further supported by Figure 1, where attributes plotted against each other demonstrate linear relationships. While examining the distribution of attributes, some exhibit characteristics closer to a normal distribution than others, with urea deviating significantly.
The analysis also reveals a few outliers, particularly in pH and calcium levels, as depicted in Figure 3. However, these outliers do not significantly deviate from the overall data distribution.
Based on these visualizations, we believe that primary machine learning modeling is feasible. The separation observed among data points, especially when multiple attributes are considered, suggests potential for predictive modeling in identifying kidney stone presence.

3.2 PCA Analysis

Prior to PCA, attributes were standardized to have a mean of 0 and a standard deviation of 1, ensuring that variations across different scales are comparable.

Figure 5 illustrates that 80% of the data variance is captured within the first three PCA components. Hence, our analysis will focus on these components:

$$\text{PCA1} = \begin{bmatrix} 0.47 \\ -0.17 \\ 0.51 \\ 0.39 \\ 0.47 \\ 0.34 \end{bmatrix}, \quad \text{PCA2} = \begin{bmatrix} -0.01 \\ 0.95 \\ 0.09 \\ 0.27 \\ -0.06 \\ 0.12 \end{bmatrix}, \quad \text{PCA3} = \begin{bmatrix} 0.04 \\ 0.07 \\ -0.20 \\ -0.53 \\ 0.06 \\ 0.82 \end{bmatrix}$$

$$\text{Attributed to: } \begin{bmatrix} \text{Gravity} \\ \text{pH} \\ \text{Osmolarity} \\ \text{Conductivity} \\ \text{Urea} \\ \text{Calcium} \end{bmatrix}$$

The first component, accounting for approximately 60% of the variance, significantly includes nearly all attributes, indicating their collective importance in predicting kidney stones. Component 2 emphasizes pH, with a minor contribution from conductivity, while Component 3 focuses on calcium and conductivity, highlighting specific attributes' influence.

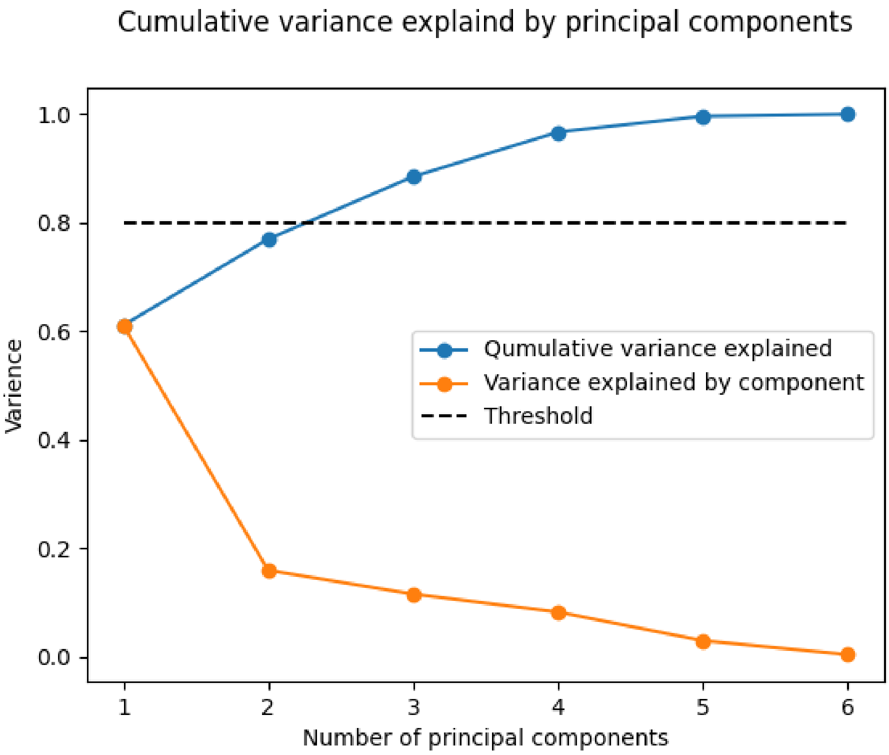

Figure 5: Cumulative variance explained by each PCA component.

The projection onto PCA components (Figure 6) does not immediately reveal clear separation between categories, suggesting that a combination of these components is necessary for distinction. This is consistent with the need for 3 components to explain 80% of the variance.

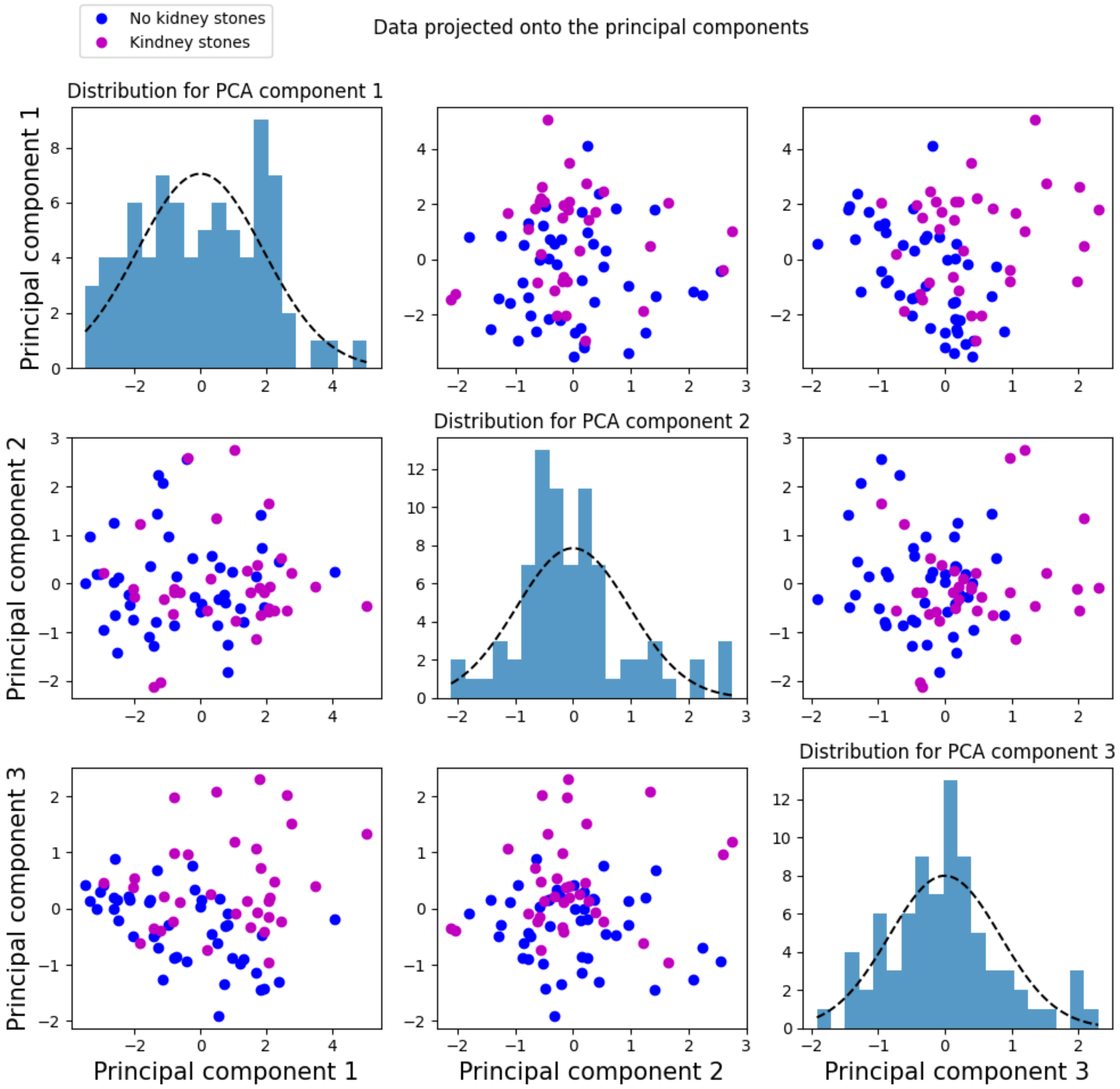

Figure 6: Data projected onto the first three PCA components.

4 Discussion

4.1 Summary

To finish this introductory paper, we want to summarize what we have learned from our data in the context of using it for building a classification model.

The data does not have any issues but is rather small with only 79 patients included. Every patient represents one observation for each attribute, as seen in Table 1.
Six attributes are continues and are measures of urine. This is a good type of data to train a classification model, as it gives more nuances to each data point than discrete would have. The seventh attribute ”Target” is different. Target is a binary measure of kidney stones and is therefore a perfect feature to predict with a classification model. ”Target” will be used as the value we would expect our model to predict. In other words, it will be the y-value, when training and testing our model. In this way we build a classification model that predicts if a person has kidney stones from information about that persons urine.

4.2 Evaluation: How will the classification/regression perform?

Compared to larger datasets, such as the University of Guilan's kidney stone dataset with 936 × 42 dimensions, our dataset's smaller size means we cannot expect classification accuracy in the high 90s. A way to counter this, could be via simulating more data via bootstrapping. While this would give more data points, a downside is that this would not introduce new information or variability to the dataset.

Furthermore, we can see in figure 1, that there is no 100% clear distinction between our two classes, 'kidney stone' and 'no kidney stone'. it's therefore unrealistic for our dataset get near 100% accuracy. Other computer engineers have used our dataset and gotten a classification accuracy of 78.31%.

Based on this information, classification seems feasible, but we are not expecting to get a classification accuracy above 90.

Since there is 34 datapoints with kidney stones and 45 datapoints without kidney stones, a classification model that classifies everything as "no kidney stone" would be correct about 57% of the time. We expect a future model to perform better than this. To perform regression successfully, it helps if the data is correlated. The correlation between urea concentration and calcium concentration is 0.5, which is not good nor bad. Regression in this case seems feasible, but the plot will most likely have high error function $E(\mathbf{w})$

References

[1] D. F. Andrews and A. M. Herzberg. *Physical Characteristics of Urines With and Without Crystals*, pages 249–252. Springer New York, New York, NY, 1985.

[2] GANESH JAINARAIN. Kidney stone prediction eda binary classification. `https://www.kaggle.com/code/richeyjay/kidney-stone-prediction-eda-binary-classification/notebook`, 2023.

[3] Yassaman Kazemi and Seyed Abolghasem Mirroshandel. A novel method for predicting kidney stone type using ensemble learning. *Artificial Intelligence in Medicine*, 84:117–126, 2018.

5 Exam questions

Question 1

Answer: D.
First we look at the attribute $x1$: Time of day, which is described as '30 minute interval (coded)'. This attribute is interval, since the difference between say 1-3 is the same as 9-11. it's not ratio, since it has no particular meaning saying 4 is 'twice as large' as 2.
The only answer which has $x1$ as 'interval' is D. By process of elimination, the correct answer must be D.

Question 2

Answer: A.
You calculate this using the formula for finding p-distance:

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

In the case of A, we have $p = \infty$:

$$\max\{|x_1-y_1|, |x_2-y_2|, \ldots, |x_M-y_M|\} = max\{|26-19|, |0-0|, |2-0|, |0-0|, |0-0|, |0-0|, |0-0|\} = max\{7,2\} = 7$$

The answer is 7, there is statement A correct.

Question 4

Answer: D.
The answer D tells me to look at the 2nd component, which is the 2nd column of the matrix v. There is also some information about the input: "An observation with a low value of Time of day (x1), a high value of Broken Truck(x2), a high value of Accident victim(x3), and a high value of Defects(x4) will typically have a positive value". So x1 is -0.5 in the second column of the v matrix, this will be multiplied by a low input (< 0) and we can therefore expect a positive output from x1. I do this with all the given inputs and concludes that x1,x2,x3,x4 are all going to have positive outputs going through component 2. But there is an uncertainty as there isn't any information about the second strongest parameter x5, that's why answer D says "typically have a positive value".

Question 5

the equation for jaccard similarity is :

$$J(i,j) = \frac{a}{a+b+c}$$

where:

- a is the number of words both in s_1 and s_2
- b is the number of words in s_1 and not in s_2
- c is the number of word in s_2 and not in s_1

after some quick counting we find that $a = 2$, $b = 6$ and $c = 5$. This gives us

$$\frac{2}{2+5+6} = 0.153846$$

The answer is therefore A