Danmarks Tekniske Universitet

DTU

Prostate Cancer Prediction

Machine Learning & Data Mining
02450

Authors

Buster Bøgild Nielsen - s211548
Ludvig Lui Reck Madsen - S205011

Section	Buster	Ludvig
Chapter 1	60%	40%
Chapter 2	40%	60%
Chapter 3	60%	40%
Chapter 4	40%	60%
Question 2	x	
Question 3	x	
Question 5	x	
Question 6	x	

October 3, 2024

Contents

1 Data set

The prostate cancer dataset focuses on exploring the relationship between various clinical measures and the level of prostate-specific antigen (PSA) in men who are about to undergo radical prostatectomy. PSA levels are crucial in diagnosing and monitoring prostate cancer, and understanding the factors that influence PSA can provide insights into cancer progression and aid in clinical decision-making. The data consists of 97 observations and includes nine variables, capturing both continuous and categorical predictors related to prostate cancer.

1.1 Source

This dataset originates from a study conducted by Stamey et al. (1989), which investigated the correlation between PSA levels and multiple clinical measurements in prostate cancer patients.

1.2 Previous data analysis

Previous analyses of this dataset typically focus on regression modeling, specifically linear regression, to understand how clinical measurements predict PSA levels. For example, studies have examined the influence of factors such as prostate volume, Gleason score, and the presence of seminal vesicle invasion on PSA levels. These analyses often use standard statistical techniques, including feature scaling and variable selection, to optimize model performance.

1.3 Classification and Regression Tasks

We want to examine the relationship between PSA levels (prostate-specific antigen) and five other attributes (age, cancer volume, prostate weight, benign prostatic hyperplasia, and capsular penetration).

2 Attributes

The prostate cancer dataset includes several clinical measurements related to prostate cancer and PSA levels. Below is a detailed explanation of the attributes, including their units, types, and descriptions.

graphicx

Table 1: Attribute Descriptions and Types

Attribute	**Unit**	**Type**	**Description**
lcavol (Log Cancer Volume)	cm^2	Continuous, Ratio	Logarithm of the tumor volume in the prostate.
lweight (Log Prostate Weight)	grams	Continuous, Ratio	Log-transformed weight of the prostate.
age (Age)	Years	Continuous, Ratio	Age of the patient at the time of measurement.
lbph (Log of Benign Prostatic Hyperplasia)	cm^2	Continuous, Ratio	Logarithm of the benign prostatic hyperplasia amount.
svi (Seminal Vesicle Invasion)	Binary	Discrete, Nominal	Binary indicator for whether cancer cells have invaded the seminal vesicle.
lcp (Log of Capsular Penetration)	cm	Continuous, Ratio	Linear extent of cancer penetration into the prostate capsule.
gleason (Gleason Score)	N/A	Discrete, Ordinal	Gleason Grade representing the aggressiveness of prostate cancer.
pgg45 (Percent of Gleason Scores 4 or 5)	Percentage	Continuous, Ratio	Percentage of Gleason scores 4 or 5, indicating aggressive cancer cells.
lpsa (Log PSA)	ng/mL	Continuous, Ratio	Logarithm of prostate-specific antigen levels.
Train (T/F)	N/A	Binary	Indicator for training set membership.

The dataset is mostly complete, with valid measurements for the majority of observations. One error in the `lweight` variable was corrected from earlier versions of the dataset. The subset of continuous attributes selected for this project excludes the following variables: svi, gleason, pgg45 and train

2.0.1 Basic Summary Statistics

Table 2: Summary Statistics of Key Continuous Attributes

Attribute	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
lcavol	0.58	1.35	-1.39	3.64
lweight	3.63	0.54	2.37	4.78
age	63.90	7.31	41.00	79.00
lbph	-0.53	0.70	-1.39	2.33
lcp	-1.39	0.74	-1.39	2.11
pgg45	15.28	21.92	0.00	100.00
lpsa	2.47	0.76	-0.43	5.58

3 Data visualizations.

Visualizing the data using both a scatter plot and histograms helps identify potential outliers and irregularities in the distribution. Because the attributes intervals are so different, the data has been standardised before making the scatter plot, by subtracting the mean and deviding by the standard deviation.

Figure 1: Scatter Plot

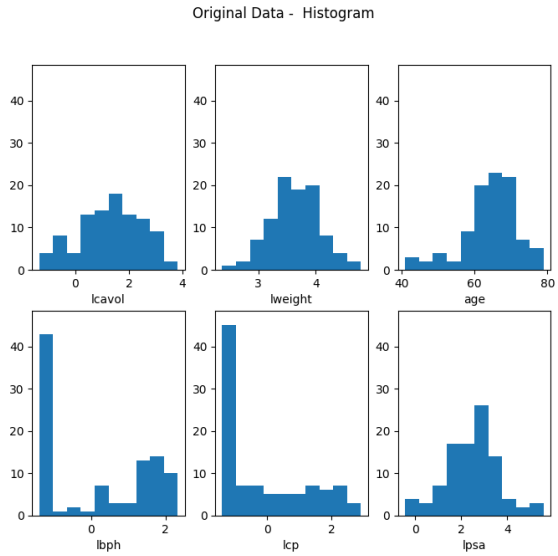

Figure 2: Histograms

From the histograms (Figure 2), it is evident that variables such as **lbph** and **lcp** show a high concentration of values near zero, which suggests the presence of potential outliers. Particularly, **lbph** has a significant clustering near zero, indicating that many observations might need special attention before further analysis.

In the scatter plot (Figure 1), the normalized and standardized data highlights individual points that lie far outside the main data clusters. For example, **lcavol** and **lpsa** show data points that are distant from the bulk of the data, suggesting these might be outliers. This is crucial to address before applying machine learning models, as these outliers could impact the accuracy of the predictions.

To evaluate whether the attributes in the dataset follow a normal distribution, we examined the histograms for each variable. The results are mixed, with some variables approximating a normal distribution, while others deviate significantly:

- **lcavol** (Log Cancer Volume) appears to be normally distributed but is slightly right-skewed. The distribution shows a peak near the center with a gradual tapering off towards higher values.

- **lweight** (Log Prostate Weight) also shows characteristics of a normal distribution, though it is slightly right-skewed like lcavol. However, the skewness is less pronounced, indicating that this variable is closer to a normal distribution.
- **age** is clearly right-skewed, with the majority of observations clustering towards the lower age range. This skewness suggests that age does not follow a normal distribution and that a transformation might be necessary for further analysis.
- **lbph** (Log Benign Prostatic Hyperplasia) does not follow a normal distribution. A significant number of observations cluster near zero, which suggests that most patients in the dataset have low values for benign prostatic hyperplasia. This heavy concentration near zero indicates that further preprocessing, such as a transformation, may be required.
- **lcp** (Log Capsular Penetration) shows a similar pattern to lbph, with a large number of observations clustered near zero. This indicates a deviation from a normal distribution, which could complicate the modeling process if left untreated.
- **lpsa** (Log PSA) appears to be closer to a normal distribution than some of the other variables but is slightly left-skewed. This slight deviation from normality might still be acceptable for certain models, but it is something to consider depending on the chosen approach.

In conclusion, while variables like **lcavol** and **lweight** approximate normal distributions with minor skewness, others, such as **lbph** and **lcp**, deviate significantly from normality. Addressing these non-normal distributions through transformations (e.g., log or square-root transformations) or other techniques will be essential for improving the performance of machine learning models that assume normality.

3.1 Attribute correlation

To explore the relationships between the attributes, we use a scatter plot matrix, which helps visualize the pairwise correlations between the variables. This matrix enables us to quickly identify potential linear relationships.

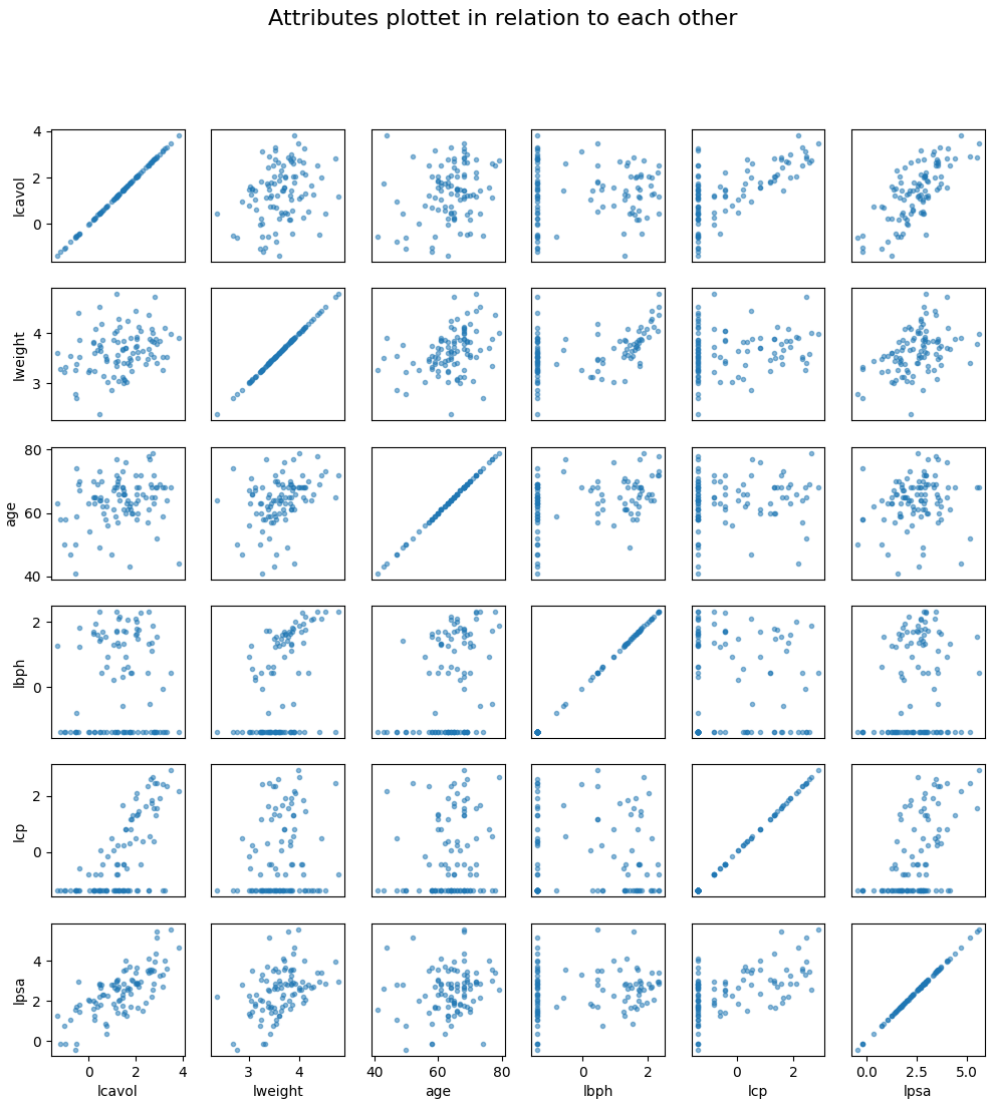

Figure 3: Pair scatterplot

From the scatter plot matrix (Figure 3), we can observe the following correlations between attributes:

- *lcavol* (Log Cancer Volume) and *lpsa* (Log PSA) exhibit a strong positive correlation, as evident by the clear linear trend between the two. This suggests that higher tumor volumes are associated with higher PSA levels.
- *lcavol* also shows a moderate positive correlation with *lweight* (Log Prostate Weight), meaning that larger prostate weights tend to be associated with larger tumor volumes.
- *lpsa* and *lweight* show a noticeable positive correlation as well, indicating that prostate weight may be a predictor for PSA levels.

- *age* does not show any clear linear relationship with the other variables, suggesting that it may not be strongly correlated with other clinical measures in this dataset.
- *lbph* (Log Benign Prostatic Hyperplasia) and *lcp* (Log Capsular Penetration) both exhibit low or no significant correlation with most other attributes. These variables have large clusters of values near zero, which may obscure potential linear relationships.
- Other attribute pairs, such as *lcp* and *lpsa*, or *age* and *lpsa*, show weak or no obvious linear relationships based on the scatter plots.

In conclusion, the most prominent correlations are between *lcavol*, *lpsa*, and *lweight*, indicating that these variables are likely to be important for predictive modeling. Attributes like *age*, *lbph*, and *lcp* do not show strong relationships with the other variables and may require further investigation or transformation before use in modeling.

Noticeable correlations: *lcavol* and *lpsa*, *lcavol* and *lweight*.

3.2 Feasibility of Machine Learning Models

Based on the visualizations, the primary machine learning modeling aim appears feasible. Strong correlations between key variables such as *lcavol*, *lpsa*, and *lweight* suggest that regression models will be effective for predicting PSA levels. These relationships provide a solid foundation for linear or advanced machine learning models.

However, some attributes, like *lbph* and *lcp*, exhibit large clusters of values near zero, indicating potential issues with outliers or non-normal distributions. Preprocessing steps such as transformations may be necessary to improve model performance.

In conclusion, while regression models show promise, careful data preprocessing is required to handle outliers and weakly correlated variables, ensuring the best possible model accuracy.

3.3 PCA

3.3.1 Variability

Figure 4 llustrates the variance explained by principal components analysis (PCA).

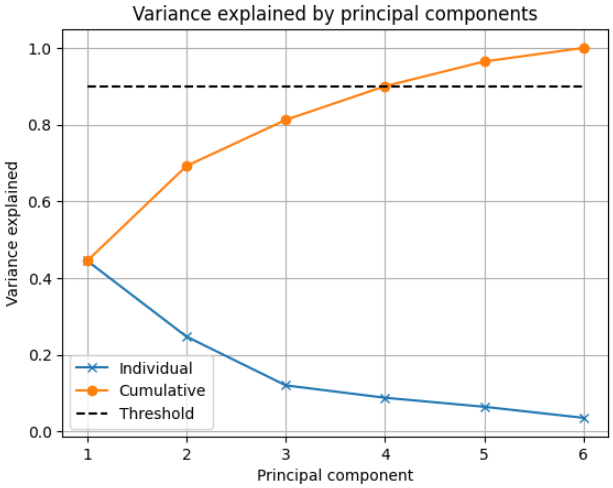

Figure 4: PC variance

X-Axis: This represents the number of principal components included in the analysis.

Y-Axis: This shows the variance explained, ranging from 0 to 1.

Individual Variance: These points represent the amount of variance explained by each individual principal component. The first few components explain a significant portion of the variance, but the contribution decreases with each additional component.

Cumulative Variance: This line shows the cumulative variance explained as more principal components are added. It starts at the variance explained by the first component and increases with each subsequent component.

Threshold: The horizontal dashed line indicates a threshold for the minimum amount (90%) of variance that is preferred to explain the dataset.

3.3.2 PCA Directions

Figure 8 shows a bar chart representing the principal component coefficients.

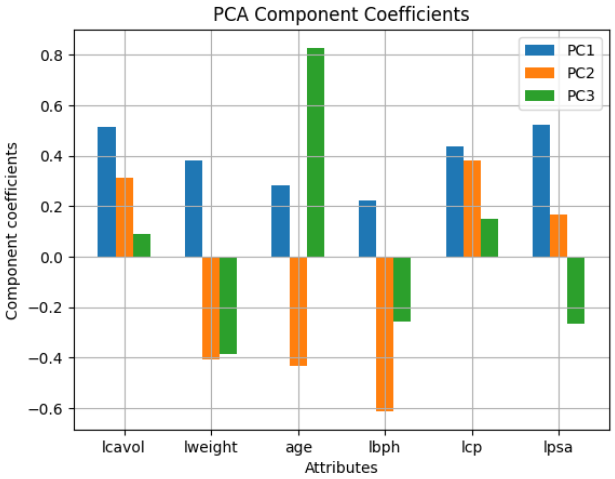

Figure 5: Principal component coefficients

The x-axis represents the attributes, while the y-axis represents the coefficients for each principal component (PC1, PC2, and PC3). The height of each bar indicates the contribution of the corresponding attribute to that principal component.

PC1: The attributes lcavol, lweight, and age have the highest positive coefficients for PC1, suggesting that these attributes are strongly correlated with the first principal component. This means that the variation in these attributes explains a significant portion of the overall variation in the dataset.

PC2: The attribute lbph has the highest negative coefficient for PC2, indicating that it is strongly negatively correlated with this component. This suggests that lbph is capturing a different aspect of the variation in the data compared to PC1.

PC3: The attributes lcp and lpsa have the highest positive and negative coefficients, respectively, for PC3. This suggests that these attributes are capturing a third dimension of variation in the data that is orthogonal to PC1 and PC2.

3.3.3 PCA Projection

Figure 9 represents a scatter plot of centered data projected onto the first two principal components.

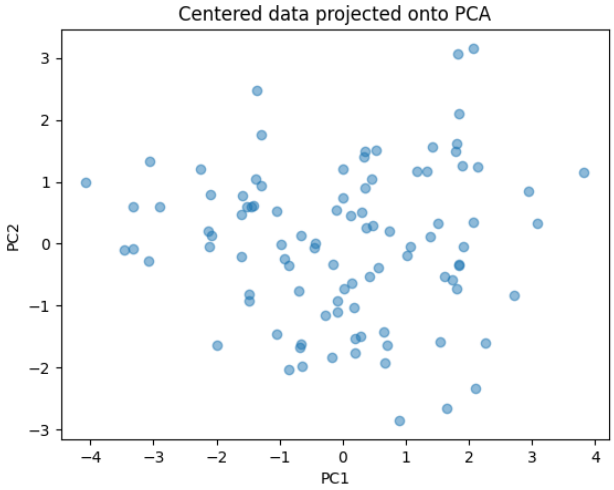

Figure 6: PCA data projection

The data points are spread out along both PC1 and PC2 axes, indicating that both components contribute to explaining the variation in the data.

The scatter plot is slightly elongated along the PC1 axis, suggesting that PC1 captures a larger portion of the variance in the data compared to PC2.

4 Learning reflections.

Project 1 - Predicting prostate cancer - has been a challenging experience, providing valuable insights into data analysis, visualization, and the application of principal component analysis (PCA).

One of the key learnings was the importance of thorough data exploration and visualization. By creating scatter plots and histograms, we were able to understand the distribution of various attributes. This process highlighted the need for careful data preprocessing, especially for variables like lbph and lcp, which showed significant clustering near zero.

Applying PCA to the dataset was interesting. It demonstrated how we can reduce the dimensionality of complex datasets while retaining most of the important information. The variance explained by each principal component and the PCA projection visualizations provided insights into the underlying structure of the data.

Most importantly for us, this project made us aware of the iterative nature of machine learning. From initial data exploration to feature selection and model evaluation, each step informed the next, highlighting the need for a flexible and adaptive approach to data science projects.

5 Question 2

General p-norm formula: $\|v\|_p = (\sum_{i=1}^{n} |v_i|^p)^{\frac{1}{p}}$
Where $v_i =$

$$\mathbf{x}_{14} - \mathbf{x}_{18} = \begin{pmatrix} 26-19 \\ 0-0 \\ 2-0 \\ 0-0 \\ 0-0 \\ 0-0 \end{pmatrix} = \begin{pmatrix} 7 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

1. $p = \infty$ (Maximum norm):

This norm takes the maximum absolute value of the vector components:

$$\|v\|_\infty = \max(|7|, |0|, |2|, |0|, |0|, |0|) = 7$$

2. $p = 3$:

We use the formula for the 3-norm:

$$\|v\|_3 = \left(|7|^3 + |0|^3 + |2|^3 + |0|^3 + |0|^3 + |0|^3\right)^{\frac{1}{3}} = \left(7^3 + 2^3\right)^{\frac{1}{3}} = (343 + 8)^{\frac{1}{3}} = 351^{\frac{1}{3}} \approx 7.054$$

3. $p = 1$:

For the 1-norm, we sum the absolute values of the components:

$$\|v\|_1 = |7| + |0| + |2| + |0| + |0| + |0| = 7 + 2 = 9$$

4. $p = 4$:

For the 4-norm:

$$\|v\|_4 = \left(|7|^4 + |0|^4 + |2|^4 + |0|^4 + |0|^4 + |0|^4\right)^{\frac{1}{4}} = \left(7^4 + 2^4\right)^{\frac{1}{4}} = (2401 + 16)^{\frac{1}{4}} = 2417^{\frac{1}{4}} \approx 7.012$$

Therefore A) must be correct, rest is false

6 Question 3

The singular values in matrix S represent the amount of variance explained by each principal component. These values are squared to get the variance explained by each component. The diagonal elements of S are: 13.9, 12.47, 11.48, 10.03, 9.45.

To compute the variance explained by each principal component, sum the squared singular values:

$$\text{Total variance} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2$$

After calculating this, the variance explained by any combination of principal components can be determined by summing their squared singular values and dividing by the total variance.

Here are the results for the variance explained by different combinations of principal components:

- The first two principal components explain 52.01% of the variance. - The first three principal components explain 71.67% of the variance. - The first four principal components explain 86.68% of the variance. - The last three principal components explain 47.99% of the variance.

Therefore A) is correct, rest is false

7 Question 5

The Jaccard similarity is calculated as follows:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Finding the intersection and union of words.

$A \cap B = \{the, words\}$

$A \cup B =$ {the, bag, of, words, representation, becomes, less, parsimonous, if, we, do, not, stem}

Counting the number of words

$|A \cap B| = 2$

$|A \cup B| = 13$

$$Jaccard(A, B) = \frac{2}{13} = 0.1538461538$$

The correct answer is **A**

8 Question 6

I need to find

$$p\left(\hat{x_2}=0 \middle| y=2\right)$$

This can be done by summing alle the propabilities where $\hat{x_2}=0$

$$p\left(\hat{x_2}=0 \middle| y=2\right)=p\left(\hat{x_2}=0,\hat{x_7}=0 \middle| y=2\right)+p\left(\hat{x_2}=0,\hat{x_7}=1 \middle| y=2\right)=0.81+0.03=0.84$$

So the correct answer is **B**