University of Science and Technology of Hanoi

Machine Learning & Data Mining II

Labwork I

Group Members:

Nguyen Tuan Thanh	22BA13289	thanhnt.22ba13289@usth.edu.vn
Nguyen Chi Quang	22BA13262	quangnc.22ba13262@usth.edu.vn

Hanoi, May 2025

Contents

Chapter 1

STUDY THE DATASET

1.1 Dataset 1: Student habits performance

The dataset used in this study represents the academic and personal profiles of 1,000 students. It includes information about each student's demographics, study habits, mental health, and lifestyle, along with their final exam scores. The main purpose is to analyze and predict how personal attributes such as age, gender, study habits, mental health, and lifestyle factors influence academic performance.

Student Id	Age	Gender	Study Hours per Day	Social Media Hours	Netflix Hours	Part-Time Job	Attendance Percentage	Sleep Hours	Diet Quality	Exercise Frequency	Parental Education Level	Internet Quality	Mental Health Rating	Extracurricular Participation	Exam Score
S1000	23	Female	0.0	1.2	1.1	No	85	8	Fair	6	Master	Average	8	Yes	56.2
S1001	20	Female	6.9	2.8	2.3	No	97.3	4.6	Good	6	High School	Average	8	No	100
S1002	21	Male	1.4	3.1	1.3	No	94.8	8	Poor	1	High School	Poor	1	No	34.3
S1003	23	Female	1.0	3.9	1.0	No	71	9.2	Poor	4	Master	Good	1	Yes	26.8
S1004	19	Female	5.0	4.4	0.5	No	90.9	4.9	Fair	3	Master	Good	1	No	66.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
S1995	21	Female	2.6	0.5	1.6	No	77	7.5	Fair	2	High School	Good	6	Yes	76.1
S1996	17	Female	2.9	1.0	2.4	Yes	86	6.8	Poor	1	High School	Average	6	Yes	65.9
S1997	20	Male	3.0	2.6	1.3	No	61.9	6.5	Good	5	Bachelor	Good	9	Yes	64.4
S1998	24	Male	5.4	4.1	1.1	Yes	100	7.6	Fair	0	Bachelor	Average	1	No	69.7
S1999	19	Female	4.3	2.9	1.9	No	89.4	7.1	Good	2	Bachelor	Average	8	No	74.9

Table 1.1: **Student Habits and Performance Dataset**

1.1.1 Dataset classification

Dimensions: There are 16 dimensions (features) in this dataset.
Discrete features: There are 9 discrete features, since these features are counted in whole numbers or categories.

- **Student Id:** (unique identifier)
- **Gender:** (Male, Female)
- **Part Time Job:** (Yes, No)
- **Diet Quality:** (Good, Fair, Poor)
- **Exercise Frequency:** (integer values indicating frequency)
- **Parental Education Level:** (High School, Master, etc.)
- **Internet Quality:** (Good, Average, Poor)
- **Mental Heath Rating:** (integer values between 1 and 10)

- **Extracurricular Participation:** (Yes, No)

Continuous features: There is 7 continuous feature (Age, Study Hours per Day, Social Media Hours, Netflix Hours, Attendance Percentage, Sleep Hours, Exam Score) because they are often represented by real numbers, including decimals
Quantitative and Numerical features: There are 9 features (Age, Study Hours per Day, Social Media Hours, Netflix Hours, Attendance Percentage, Sleep Hours, Exercise Frequency, Mental Health Rating, Exam Score). They represent measurable data, typically expressed numerically.
Qualitative features: There are 7 qualitative features (Student Id, Gender, Part Time Job, Diet Quality, Parental Education Level, Internet Quality, Extracurricular Participation). They represent categories or attributes that describe characteristics or qualities of the data. These features are typically non-numeric or numerical data encoded by numbers) and can be observed but not measured.

1.1.2 Data issues and preprocessing

The dataset contains 1000 rows and 16 columns, including both numerical and categorical attributes. A preliminary inspection shows that no data is missing, all columns have 100% completeness, with null count and null percentage equal to 0.
Additionally, the dataset was checked for duplicate entries, and no duplicate rows were found. All feature names are well-structured and values are consistent with the expected format.
The dataset is clear and complete with Exam Score label in the dataset. This label represents the outcome or target variable that we aim to predict based on other features (Study Hours per Day, Attendance Percentage, etc).

1.1.3 Statistical Measures

Assume that the number of observations: n.

- **Mean ($\bar{x}$):** Calculate the mean of each feature.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

	Age	Study Hours per Day	Social Media Hours	Netflix Hours	Attendance Percentage	Sleep Hours	Exercise Frequency	Mental Health Rating	Exam Score
Mean	20.4980	3.5501	2.5055	1.8197	84.1317	6.4701	3.0420	5.4380	69.6015

Table 1.2: Mean of features

- **Variance (Var(X)):** Calculate the variance of each feature

$$\text{var}(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2$$

where σ is the standard deviation of x (a vector).

	Age	Study Hours per Day	Social Media Hours	Netflix Hours	Attendance Percentage	Sleep Hours	Exercise Frequency	Mental Health Rating	Exam Score
Variance	5.327323	2.157638	1.374574	1.155878	88.345831	1.504000	4.102338	8.108264	285.223591

Table 1.3: Variance of features

- **Covariance:** a measure of how much two random variables change together
  - **Population covariance**

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

  - **Sample covariance**

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

	Age	Study Hours per Day	Social Media Hours	Netflix Hours	Attendance Percentage	Sleep Hours	Exercise Frequency	Mental Health Rating	Exam Score
Age	5.3273	0.0135	-0.0248	-0.0029	-0.5652	0.1061	-0.0179	-0.2964	-0.3471
Study Hours per Day	0.0135	2.1576	0.0349	-0.0492	0.3626	-0.0500	-0.0854	-0.0158	20.4765
Social Media Hours	-0.0248	0.0349	1.3746	0.0145	0.4461	0.0262	-0.0886	0.0050	-3.3014
Netflix Hours	-0.0029	-0.0492	0.0145	1.1559	-0.0211	-0.0012	-0.0140	0.0255	-3.1190
Attendance Percentage	-0.5652	0.3626	0.4461	-0.0211	88.3458	0.1586	-0.1496	-0.5017	14.2605
Sleep Hours	0.1061	-0.0500	0.0262	-0.0012	0.1586	2.5203	-0.1496	0.0491	2.5203
Exercise Frequency	-0.0179	-0.0854	-0.0886	-0.0140	-0.1496	-0.1496	4.1023	0.0491	5.4767
Mental Health Rating	-0.2964	-0.0158	0.0050	0.0255	-0.5017	0.0491	0.0246	8.1086	15.4621
Exam Score	-0.3471	20.4765	-3.3014	-3.1190	14.2605	2.5203	5.4767	15.4621	285.2239

Table 1.4: Covariance Matrix

- **Correlation (ρ):** Correlation normalizes covariance by dividing it by the standard deviations of the two variables

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

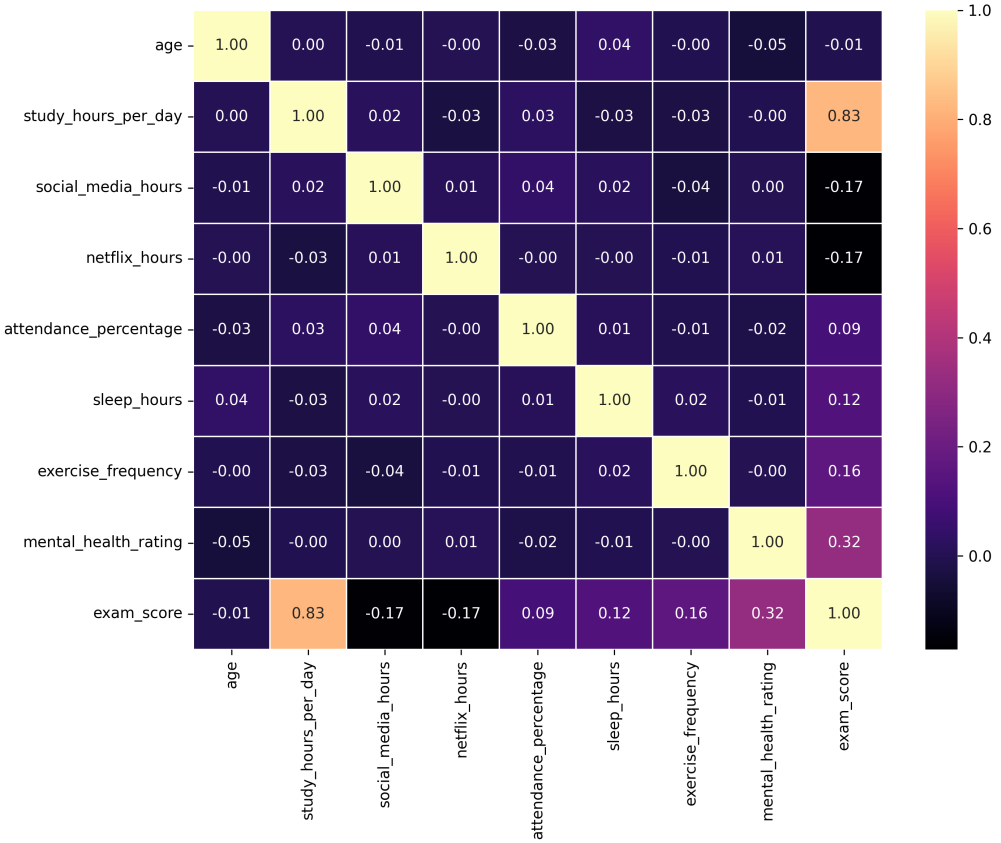

Figure 1.1: Correlation Matrix of Dataset 1

Based on the correlation matrix, the most correlated pair of features in your dataset is between Study Hours per Day and Exam Score, with a correlation of 0.83, which is nearly 1 (a strong positive relationship). This strong positive correlation suggests that as the number of study hours per day increases, exam scores also tend to increase. This relationship could be useful for future studies focused on improving student performance by increasing study time.

- **Calculate these measures for categorical features.**
  Categorical data cannot be directly measured by mean, variance, covariance, or correlation since these are measures defined for numerical data. Although there are two most popular methods to perform them:
  - Count the frequency of each value.
  - Encode categorical values to numerical values (Label Encoding, Ordinal Encoding).

1.2 Dataset 2: Insurance Premium

The dataset used in this study is a publicly available health insurance dataset that contains demographic and lifestyle information of individuals, along with their corresponding medical insurance expenses. Its primary purpose is to analyze and predict how personal attributes such as age, gender, body mass index (BMI),... influence insurance costs.

Age	Sex	BMI	Children	Smoker	Region	Expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33.0	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
...	...	...	...	...	...	...
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41

Table 1.5: **Insurance Cost Data**

1.2.1 Dataset classification

Dimensions: There are 7 dimensions (features) in this dataset.
Discrete features: There are 4 discrete features, since these features are counted in whole numbers or categories.

- **Sex:** (Male, Female)
- **Children:** (a count and takes integer values)
- **Smoker:** (Yes, No)
- **Region:** (northeast, northwest, southeast, southwest)

Continuous features: There are 3 continuous features (Age, BMI, Expenses) because they are often represented by real numbers, including decimals.
Quantitative and Numerical features: There are 4 features (Age, BMI, Children, Expenses). They represent measurable data, typically expressed numerically.
Qualitative features: There are 3 qualitative features (Sex, Smoker, Region). They represent categories or attributes that describe characteristics or qualities of the data. These features are typically non-numeric or numerical data encoded by numbers) and can be observed but not measured.

1.2.2 Data issues and preprocessing

The dataset contains 1,338 rows and 7 columns, including both numerical and categorical attributes. A preliminary inspection shows that no data is missing, all columns have 100% completeness, with null count and null percentage equal to 0
Additionally, the dataset was checked for duplicate entries, and no duplicate rows were found. All feature names are well-structured and values are consistent with the expected format.
The dataset is clear and complete with Expenses label in the dataset. This label represents the outcome or target variable that we aim to predict based on other features.

1.2.3 Statistical Measures

Assume that the number of observations: n.

- **Mean ($\bar{x}$):** Calculate the mean of each feature.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

	Age	BMI	Children	Expenses
Mean	39.20	30.66	1.09	13270.42

Table 1.6: Mean of features

- **Variance (Var(X)):** Calculate the variance of each feature

$$\text{var}(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2$$

where σ is the standard deviation of x (a vector).

	Age	BMI	Children	Expenses
Variance	197.40	37.19	1.45	146,652,400

Table 1.7: Variance of features

- **Covariance:** a measure of how much two random variables change together
  - **Population covariance**

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

  - **Sample covariance**

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

  The Covariance of Age & Expenses is: 50874.802

- **Correlation (ρ):** Correlation normalizes covariance by dividing it by the standard deviations of the two variables

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  The correlation of Age & Expenses is: 0.299008192285

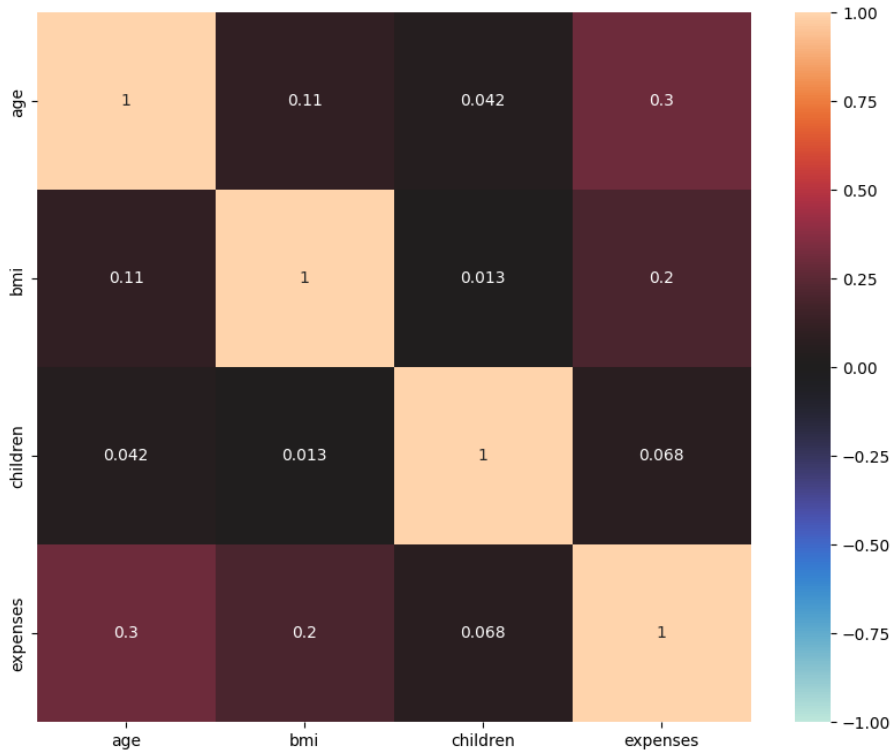

Figure 1.2: Correlation Matrix of Dataset 2

Above correlation and covariance value inform that there exist the strongest relationship between Expenses and Age: 0.3.

- **Calculate these measures for categorical features.**
  Categorical data cannot be directly measured by mean, variance, covariance, or correlation since these are measures defined for numerical data. Although there are two most popular methods to perform them:
  - Count the frequency of each value.
  - Encode categorical values to numerical values (Label Encoding, Ordinal Encoding).

Chapter 2

PRINCIPAL COMPONENTS ANALYSIS (PCA)

2.1 Dataset 1: Student habits performance

2.1.1 Apply Principal Components Analysis

Before applying PCA, we need to encode the categorical features and standardize the data. PCA relies on variance to identify the principal components.
For Student habits performance dataset:

- **Gender:** One-Hot Encoding (Convert "Female" and "Male" and "Other")
- **Part-Time Job:** Label Encoding (No = 0, Yes = 1)
- **Diet Quality:** Ordinal Encoding (Poor < Fair < Good) ⇒ Poor = 0, Fair = 1, Good = 2
- **Parental Education Level:** Ordinal Encoding (None < High School < Bachelor < Master) ⇒ None = 0, High School = 1, Bachelor = 2, Master = 3
- **Internet Quality:** Ordinal Encoding (Poor < Average < Good) ⇒ Poor = 0, Average = 1, Good = 2
- **Extracurricular Participation:** Label Encoding (No = 0, Yes = 1)

Select the principal components

To select the number of principal components, we analyzed the `explained_variance_ratio_` and visualized it through a Scree Plot. We observed that the first few components retain the majority of the variance. Based on the "elbow method" and the cumulative variance curve, we selected k components, which retained approximately XX% of the total variance in the dataset.

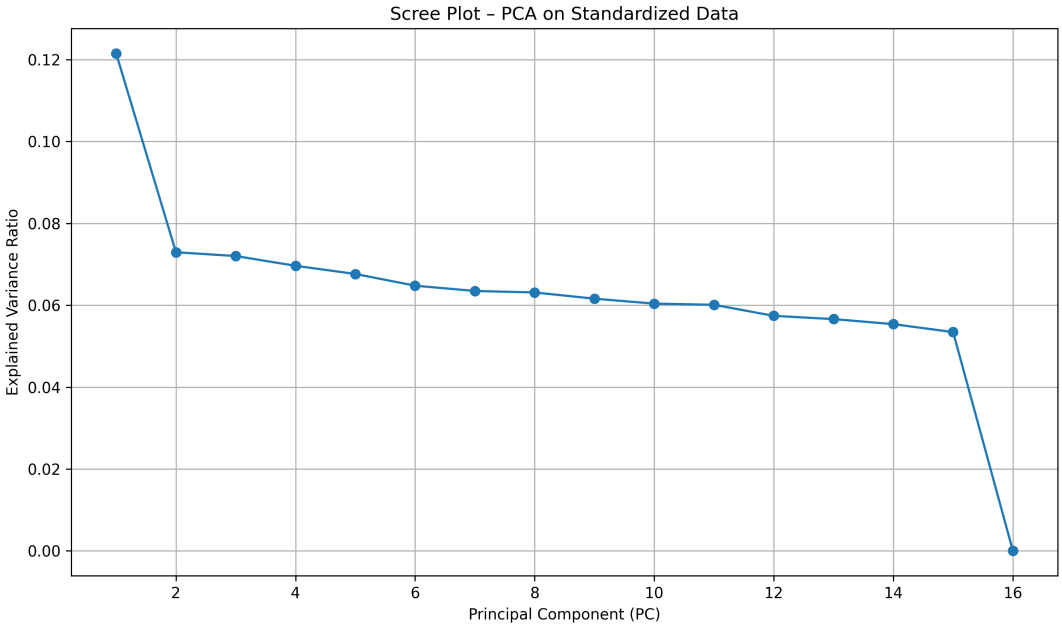

Figure 2.1: The Scree Plot of Dataset 1

Select the first 13 principal components with a total explained variance of 89.12%.

Difference while using principal components with highest and lowest values

- Using the principal components with the **highest explained variance**, they capture the major trends and patterns in the data. In the 2D PCA scatter plot, we can see some general patterns or gradual changes in the data, even though the points do not form clearly separated groups. These components contain meaningful information that helps distinguish between different regions in the dataset.
- The components with the **lowest explained variance** contain noise or redundant information. When plotting data using these components, the result is typically a scatter of points with no visible structure or separability.

Conclusion: The components with the highest explained variance are more useful for dimensionality reduction, visualization, and interpretation. The components with the lowest explained variance do not contribute much meaningful information and can be safely discarded without significantly affecting the data representation

2.1.2 Vary the number of used principal components

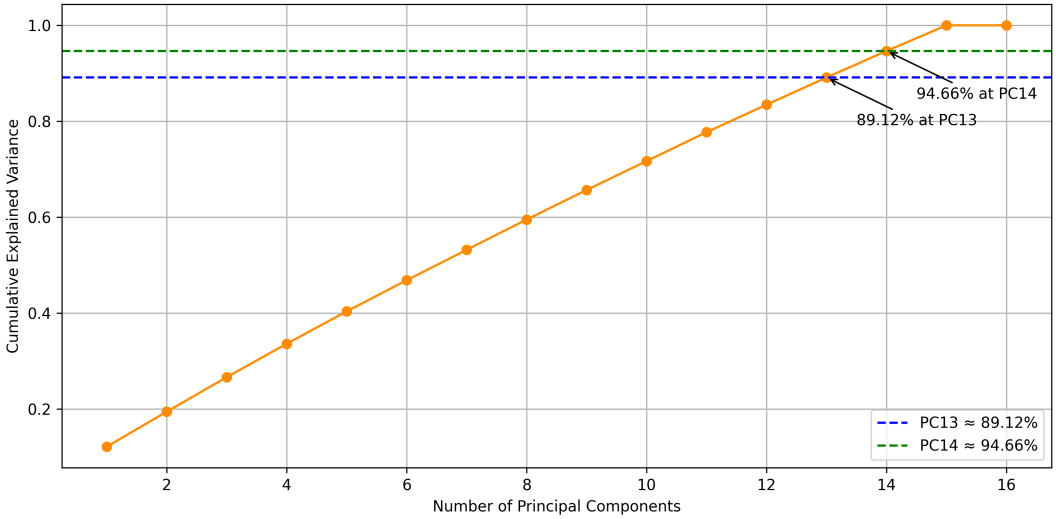

Figure 2.2: Cumulative Explained Variance vs Number of Principal Components of Dataset 1

To evaluate how much information is retained as we vary the number of principal components, we plotted the cumulative explained variance against the number of components.

Observations

- The more components we add, the more of the data's variance we keep.
- With just **13 principal components**, we already capture around **89.12%** of the total variance.
- Adding one more (up to 14 components) pushes that number to about **94.66%**.

Interpretation

- The dashed horizontal lines in the plot represent common thresholds for acceptable variance retention:
  - **85%** is the minimum for many use cases.
  - **90–95%** is preferred when accuracy really matters.
- Based on this, picking **13 or 14 components** is a smart balance — we reduce complexity but still keep the important patterns.

Conclusion:

Choosing 13 or 14 components helps shrink the dataset without losing too much valuable information.

2.1.3 Visualize data distribution in 2D

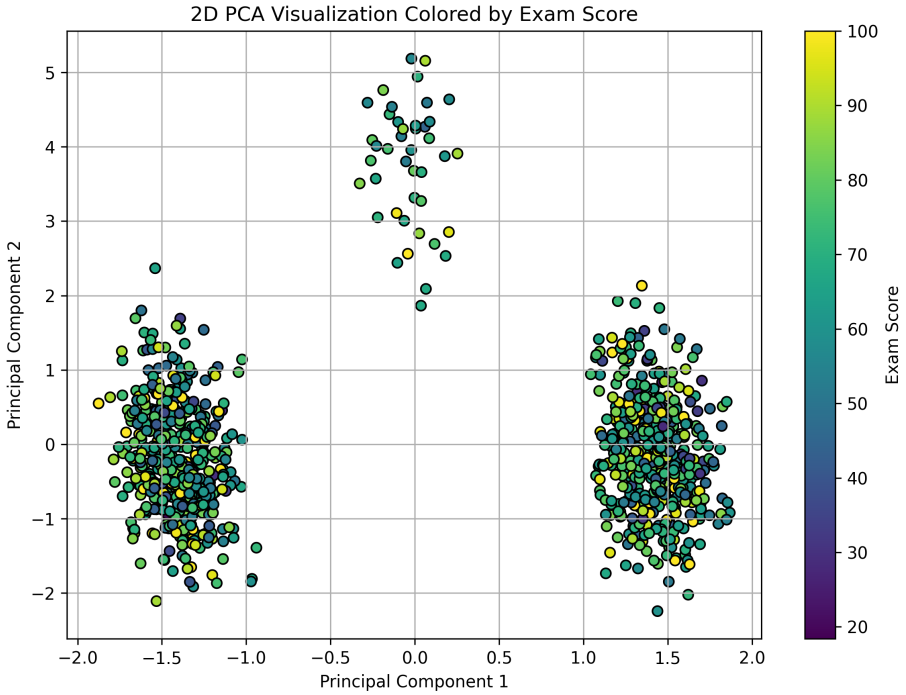

Figure 2.3: 2D PCA Visualization of Dataset 1

Observations

- While there's some visible separation by color, the data does not split into sharply distinct clusters. This suggests that student performance may follow a continuous gradient rather than discrete categories.
- There is a gradual color transition in the PCA space, but no strong directional clustering is observed. High and low exam scores are spread across the plot, indicating a continuous rather than segmented distribution of student performance.
- The 2D PCA plot reveals a smooth gradient in exam scores, indicating a continuous relationship between features and performance. Although no distinct clusters are observed, PCA has preserved meaningful variance. However, more components should be used for accurate regression modeling.

Conclusion

2D PCA is effective for visualizing and understanding the dataset, but it should not be used as the only input for predicting exam scores. Instead, using 13 to 14 principal components is recommended to reduce dimensionality while preserving most of the important information

2.2 Dataset 2: Insurance Premium

2.2.1 Apply Principal Components Analysis

Before applying PCA, we need to encode the categorical features and standardize the data. PCA relies on variance to identify the principal components.

Select the principal components

To select the number of principal components, we analyzed the `explained_variance_ratio_` and visualized it through a Scree Plot. We observed that the first few components retain the majority of the variance. Based on the "elbow method" and the cumulative variance curve, we selected k components, which retained approximately XX% of the total variance in the dataset.

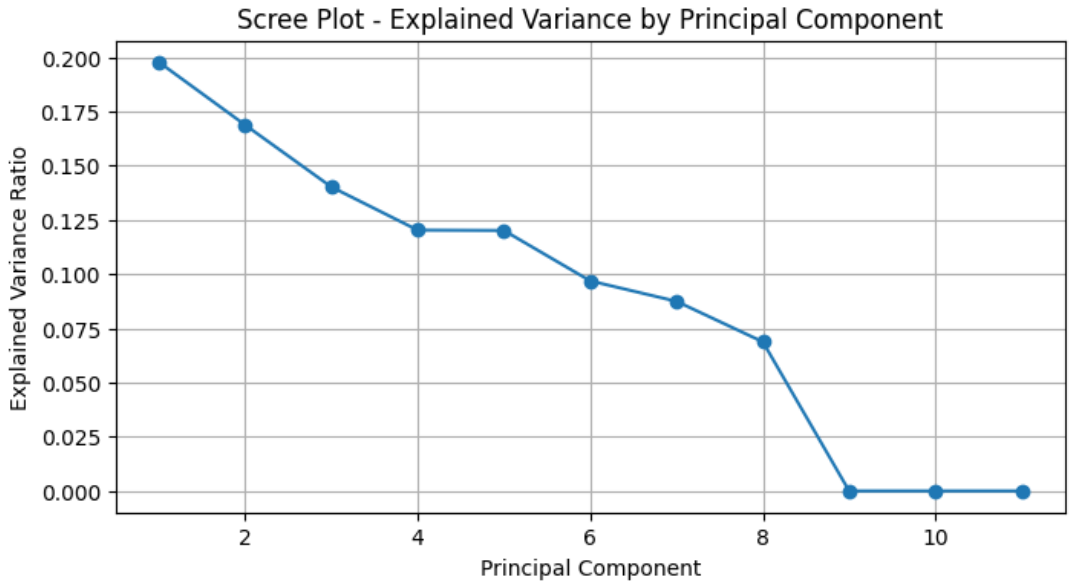

Figure 2.4: The Scree Plot of Dataset 2

Select the first 6 principal components with a total explained variance of 84%.

Difference while using principal components with highest and lowest values

- When we use components with high variance, like PC1 and PC2, we see clear patterns and well-defined clusters in the data. These components contain meaningful information that helps distinguish between different regions in the dataset.
- When we use components with low variance, the plot just looks like random noise — there's no useful structure.

Conclusion: The components with the highest explained variance are far more useful for dimensionality reduction, visualization, and interpretation. The components with the lowest explained variance do not contribute much meaningful information and can be safely discarded without significantly affecting the data representation.

2.2.2 Vary the number of used principal components

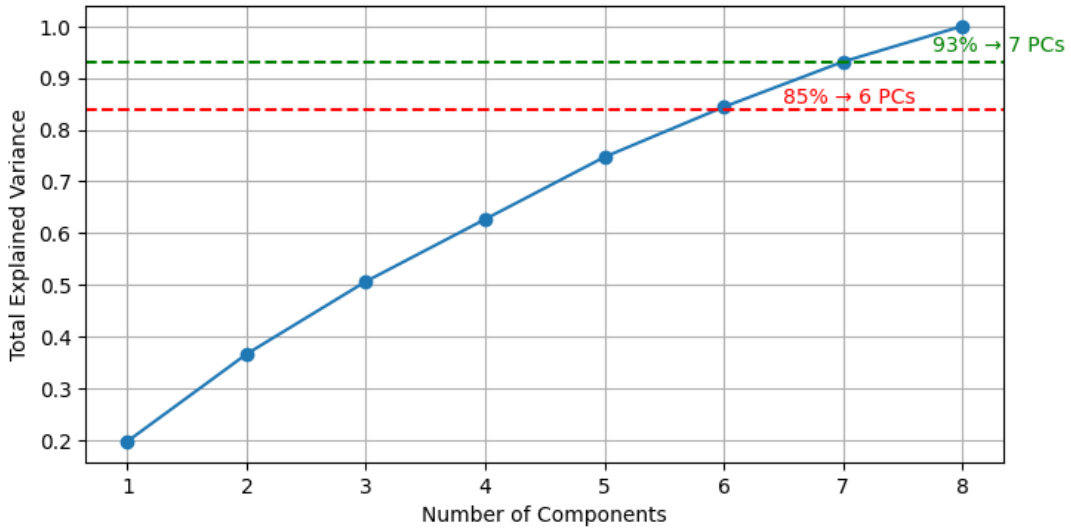

Figure 2.5: Cumulative Explained Variance vs Number of Principal Components of Dataset 2

To evaluate how much information is retained as we vary the number of principal components, we plotted the cumulative explained variance against the number of components.

Observations

- The more components we add, the more of the data's variance we keep.
- With just **6 principal components**, we already capture around **85%** of the total variance.
- Adding one more (up to 7 components) pushes that number to about **93%**.

Interpretation

- The dashed horizontal lines in the plot represent common thresholds for acceptable variance retention:
  - **85%** is the minimum for many use cases.
  - **90–95%** is preferred when accuracy really matters.
- Based on this, picking **6 or 7 components** is a smart balance — we reduce complexity but still keep the important patterns.

Conclusion:

Choosing 6 or 7 components helps shrink the dataset without losing too much valuable information.

2.2.3 Visualize data distribution in 2D

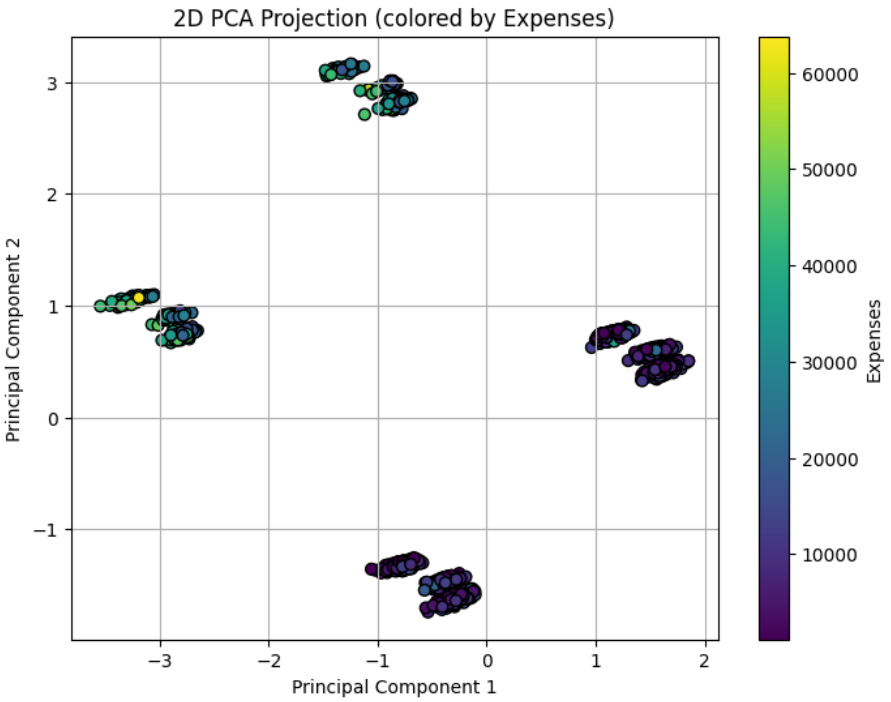

Figure 2.6: 2D PCA Visualization of Dataset 2

Observations

- The data naturally falls into 5 to 6 clear groups, showing that PCA does a good job capturing the hidden structure in the dataset.
- These clusters are well-separated, showing that PCA effectively reduces dimensionality while preserving meaningful variation.
- The color gradient reveals that individuals with high insurance costs (in yellow) tend to concentrate in the left side of the plot (PC1 less than 0), while those with lower costs (in purple) are grouped in other areas.
- *PC1 appears to carry more discriminative power than PC2, as the color changes more significantly along the PC1 axis.*

Conclusion

PCA gives us a clear picture of how features relate to insurance costs. The clusters that show up suggest that things like smoking, BMI, or age might have a big impact on how much people pay.