Danmarks Tekniske Universitet

Introduction to Machine Learning and Data Mining

Georgios Arvanitidis

Project 1: Data's Feature Extraction and Visualisation

Alexandre del Barco (s232518)
Jakub Oszczak (s233577)

November 26, 2023

Contents

List of Figures

List of Tables

Student ID	Section 1	Section 2	Section 3	Section 4	Exam Questions
s232518	60%	60%	40%	40%	50%
s233577	40%	40%	60%	60%	50%

1 Description of the data set

The data set has been extracted from the *UC Irvine Machine Learning Repositiry*[1], a repository where is possible to donate and find datasets to the machine learning community. The dataset is named *Glass Identification* and was donated in 1987 by Vina Spiehler, Ph.D., *DABFT Diagnostic Products Corporationand* from the *USA Forensic Science Service*. The overall problem of interest is to classify the glass types based on the given attributes, while the glass types are defined in terms of their oxide content (weight percent in corresponding oxide.

The original source papers for this data set are not available. However, the data set has been used in several studies. Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis. The goal was to determine whether the glass was a type of "float" glass or not. The study was motivated by criminological investigation. Previous analysis of this data has been done by Jason Brownlee1 where he explored how to develop and evaluate a model for the imbalanced multi-class classification. In the classification task, we hope to learn how to predict the glass type based on the given attributes. In the regression task, we hope to learn how to predict refractive index based on other attributes. It should not be necessary to transform the data in order to accomplish those tasks.

The dataset shows 6 types of glass defined in terms of their oxide content, multiple observations of each type of glass are taken, and for each observation the composition of their oxides is decomposed by presenting the corresponding percentage.

It is a multivariate dataset with 9 features and 214 instances. The features correspond to 8 oxides possibly composing the glass and also the refractive index. Each of these features correspond to one column, to which is added a tenth column corresponding to the target, thus the glass' identification, and a final column with the ID.

Variable Name	**Role**	**Type**	**Description**
ID	ID	Integer	Identification number from 1 to 214
RI	Feature	Continuous	Refractive Index
Na	Feature	Continuous	Sodium Percentage
Mg	Feature	Continuous	Magnesium Percentage
Al	Feature	Continuous	Aluminium Percentage
Si	Feature	Continuous	Silicon Percentage
K	Feature	Continuous	Potassium Percentage
Ca	Feature	Continuous	Calcium Percentage
Ba	Feature	Continuous	Barium Percentage
Fe	Feature	Continuous	Iron Percentage
Type_of_glass	Target	Categorical	From 1 to 7 corresponding to each type of glass

Table 1: Description of variables in the dataset

As mention, the types of glass are categorised by giving them a number between 1 and 7, with the following meaning describet in Table 2:

Number attributed	**Type of glass**
1	Building windows float processed
2	Building windows non float processed
3	Vehicle windows float processed
4	Vehicle windows non float processed
5	Containers
6	Tableware
7	Headlamps

Table 2: Description of types of glass' categorisation

The source states that in the dataset there is no representation of any glass type corresponding to number 4 or *Vehicle windows non float processed*.

2 Attributes of the data

All nine attributes corresponding to the eight composition of the oxides and the refractive index are continuous and ratio variables. The reason they are considered ratio attributes is because they have a clear definition of zero, e.g. it could happen a zero percentage of Sodium in the observation; and the difference between two values is meaningful, for example a glass observation with 2% Sodium has twice as much Sodium as a glass observation with 1% Sodium. Whereas the type of glass, thus the target, is indeed categorical and nominal.

Once the dataset us imported and reviewed, basic statistic for each attribute can be performed with the output shown in Table 3:

Column	Mean	Median	Min	Max	Standard Deviation
RI	1.51	1.51	1.51	1.53	0.00
Na	13.40	13.30	10.73	17.38	0.81
Mg	2.68	3.48	0.00	4.49	1.43
Al	1.44	1.36	0.29	3.50	0.49
Si	72.65	72.79	69.81	75.41	0.77
K	0.49	0.55	0.00	6.21	0.65
Ca	8.95	8.60	5.43	16.19	1.41
Ba	0.17	0.00	0.00	3.15	0.49
Fe	0.05	0.00	0.00	0.51	0.09
Type_of_glass	2.78	2.00	1.00	7.00	2.09

Table 3: Basic summary statistics of the attributes

From Table 3 it is possible to see that not all glass observations contain all oxides, as there are oxides that minimum amount found is 0%. Whereas the maximum percentage of a oxide in an observation is 75%, corresponding to silicon, this oxide is the predominant one as the minimum amount found in an observation is 69%. On the other hand, the standard deviations for all cases are relatively low, what means that the percentage of each oxide in each observation is quite constant; in other words, there is not a high variation between observations. One reason for this small variation between oxides' composition is that there is not much variation in types of glass, so Figure 1 is plotted.

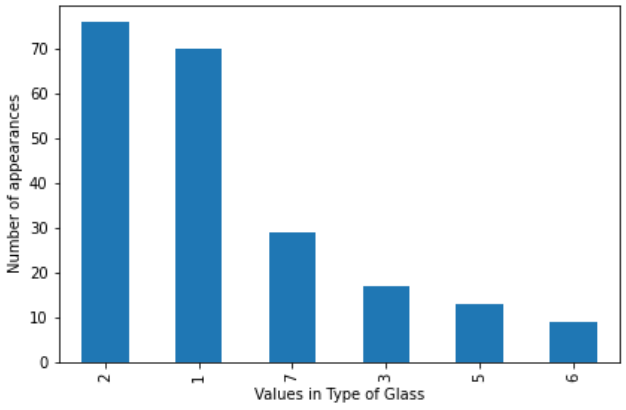

Figure 1: Histograms for each type of glass

Seeing the result of Figure 1 it might be said then then the observations under experiment are not equally distributed by types of glass, as type of glass 1 and 2, so *Building windows float processed* and *Building windows non float processed*, are way more present in the dataset than the others.

Regarding data issues, it's important to mention that the source stands the dataset has not any missing value [1]. The source does not mention anything about corrupted data nor outliers

3 Data visualization

3.1 Data distribution

Lets begin with histograms. We created those for each attribute of dataset:

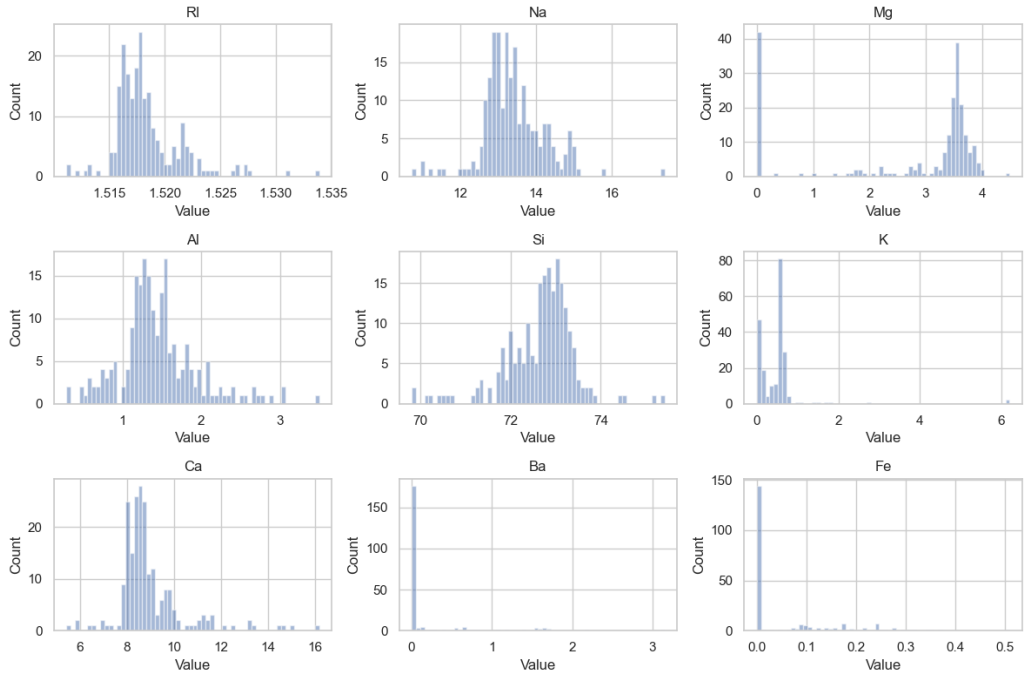

Figure 2: Histograms for each of 9 attributes.

Most of the attributes seem to be more or less normally distributed. The distributions are not perfectly symmetrical. Magnesium (Mg), Potassium (K), Iron (Fe) and Barium (Ba) are not normally distributed. In case of Magnesium if it was not for the 0 values the distribution would have been normal. Barium and Iron are almost non existent in the glass.

The graph represents standardized box plots. Standardized means each attribute has a subtracted mean and standard deviation equal to 1. As one can see from the graphs our data has quite a bit of outliers. But those are not extreme outliers, maybe except one in K attribute. Given my knowledge of the dataset, it seems to me that outliers should not be removed for the classification task.

3.2 Correlation

Next we performed a visualisation of correlation. Generated correlation matrix looks as below:

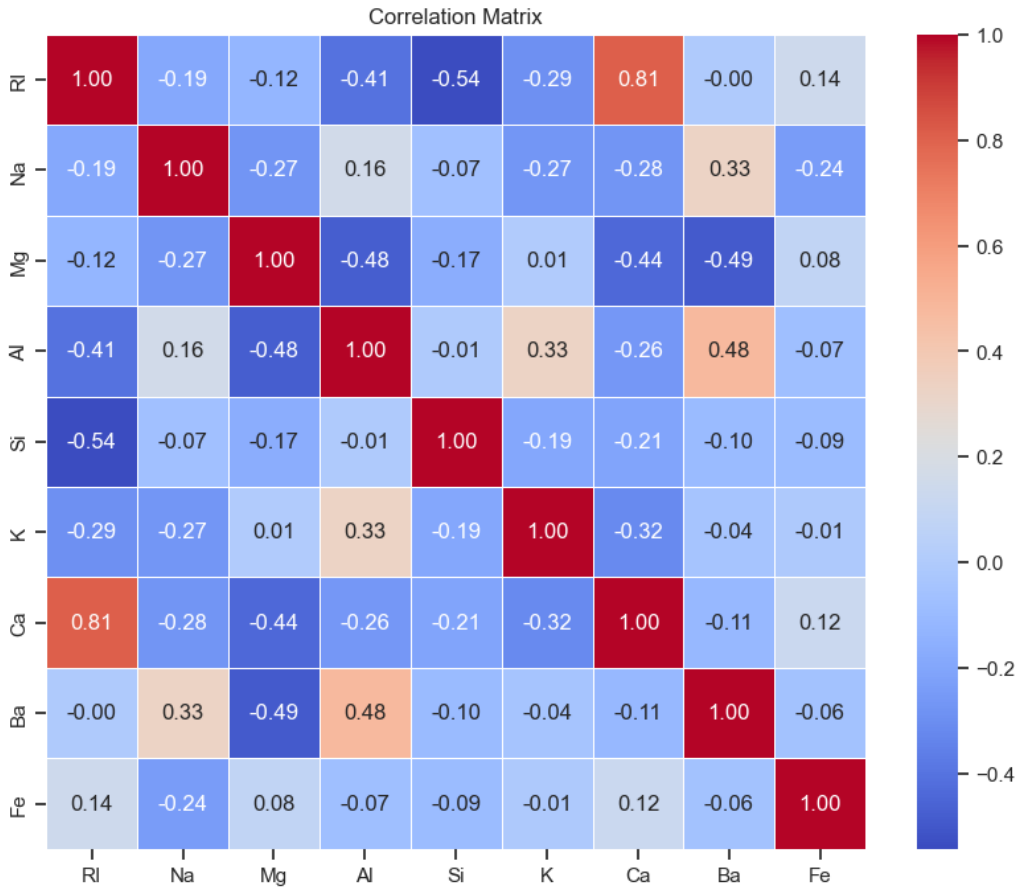

Figure 3: Correlation matrix for all attributes.

Attributes do not seem to be highly correlated, except Calcium (Ca) and Refractive index (RI). We can see a strong positive correlation between the amount of Calcium oxide in glass and it's refractive index. It means the more of Ca in the glass the higher the RI. There is also a medium strength negative correlation between the amount of Silicone (Si) and refractive index (RI). So the silicone sort of counteracts the Calcium in terms of refractive index. Also there is no correlation between the amount of Calcium and Silicone in the glass. We have also performed a pair plots for attributes to better visualize the correlation in a form of graphs. On the diagonal there are univariate distribution plots.

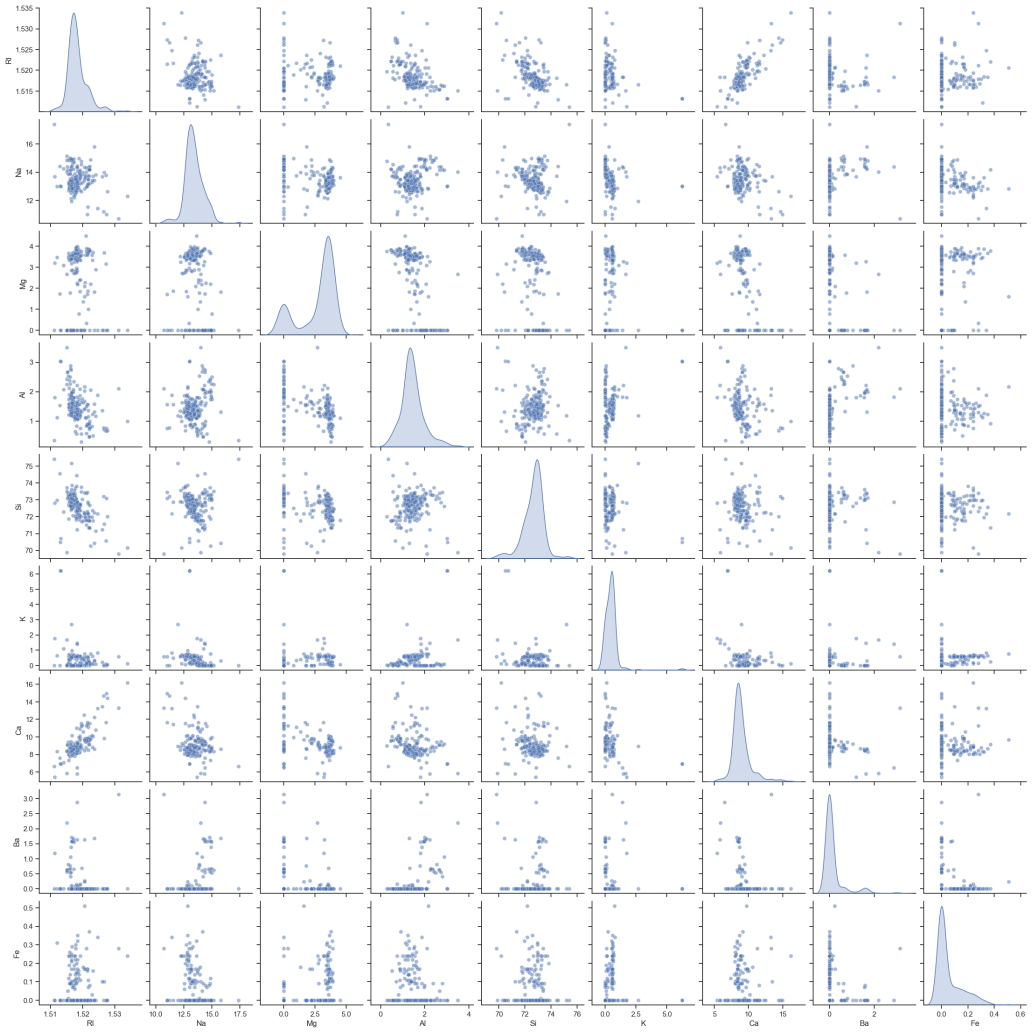

Figure 4: Pair plots of attributes and a univariate distribution plot on the diagonal.

As one can see there are a lot of straight vertical and horizontal patterns created from data points which means there is absolutely no correlation.

3.3 Principal component analysis

One important practice in data analysis is a Principal Component Analysis (PCA).

It serves a purpose of **Dimensionality Reduction** - PCA simplifies high-dimensional data, enabling easier visualization and exploration. By projecting data onto a lower-dimensional space, patterns and relationships become more apparent. But that is not the only use of PCA it also helps with **Feature Selection** - PCA helps identify the most influential features by analyzing their loadings on principal components. This aids in selecting essential attributes while discarding less informative ones. To perform the PCA, we must first standardize the data by subtracting the mean and dividing by the standard deviation for each attribute. Our result of PCA analysis is shown below:

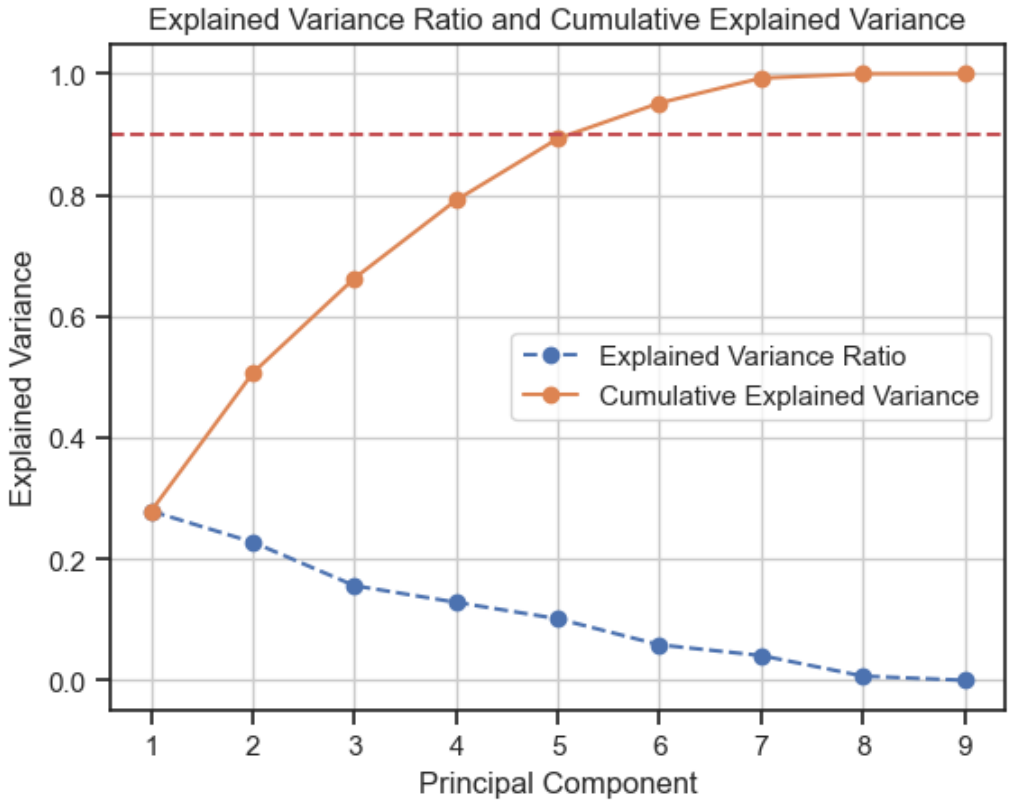

Figure 5: Graph of variance explained by every principal component. (Red dashed line is a 0.9 threshold)

It can be seen that first five components explains almost 90% of variance. So we can say that the variance is quite evenly distributed between each principal component - the slope of blue line is not steep and pretty even. The orange line shows cumulative variance explained by each next attribute.

Now lets look at the loadings, gathered in Table 3.3.

	PCA1	**PCA2**	**PCA3**	**PCA4**	**PCA5**	**PCA6**	**PCA7**	**PCA8**	**PCA9**
RI	0.545	0.285	0.086	0.147	-0.073	0.115	0.081	-0.752	0.025
Na	-0.258	0.270	-0.384	0.491	0.153	-0.558	0.148	-0.127	-0.311
Mg	0.110	-0.593	0.008	0.378	0.123	0.308	-0.206	-0.076	-0.577
Al	-0.428	0.295	0.329	-0.137	0.014	-0.018	-0.699	-0.274	-0.192
Si	-0.228	-0.155	-0.458	-0.652	0.008	0.086	0.216	-0.379	-0.298
K	-0.219	-0.153	0.662	-0.038	-0.307	-0.243	0.504	-0.109	-0.260
Ca	0.492	0.345	-0.000	-0.276	-0.188	-0.148	-0.099	0.398	-0.579
Ba	-0.250	0.484	0.074	0.133	0.251	0.657	0.351	0.144	-0.198
Fe	0.185	-0.062	0.284	-0.230	0.873	-0.243	0.073	-0.016	-0.014

Table 4: Loading of the principal components

As we can see from PCA1 loadings the RI, Al and Ca attributes are quite important ones. High values of RI and Ca will have significant positive projection on PCA1 and Al will have a negative projection.

The plot shown below pictures the data projected onto the first two PCA components. Since first two components account for only about 50% of the data's variance, this visualisation looses quite a bit of information. But despite that there is still some noticeable grouping of data occurring for different classes.

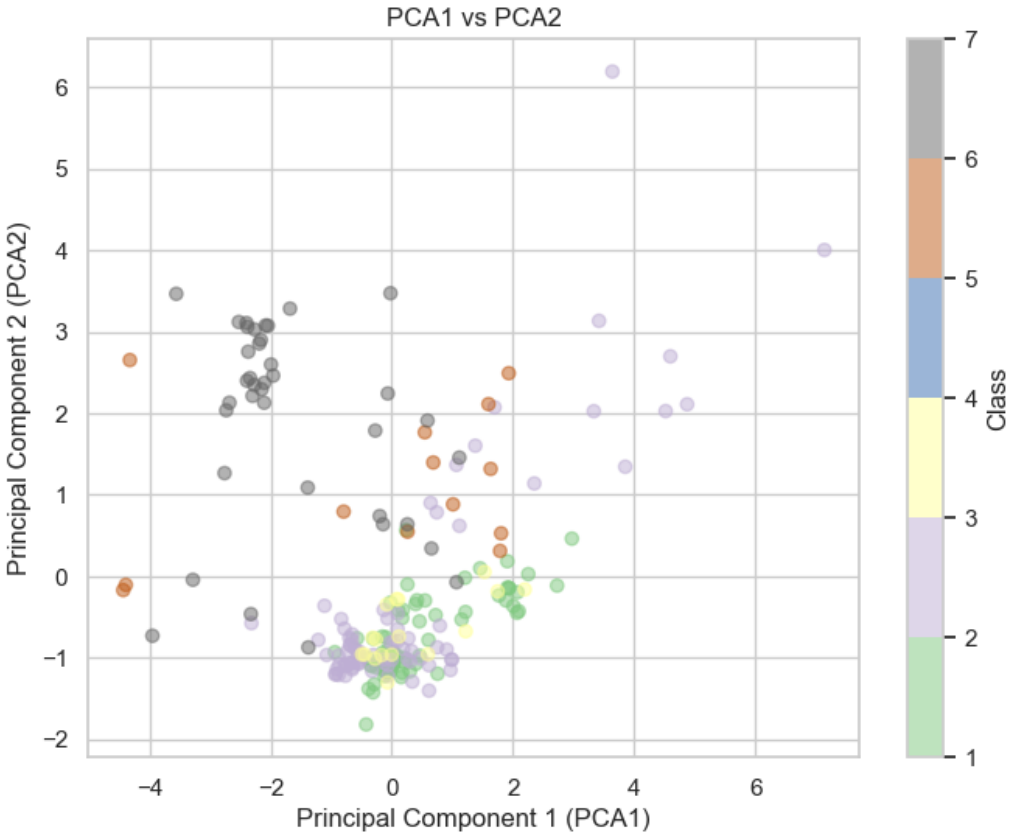

Figure 6: Projection of data points onto first two principal components.

4 Discussion

The data is gathering of observations of six types of glass with their respective attributes, in this case the attributes are percentage in composition of 9 oxides and also the refractive index of the glass. Some of the type of glass are classified as a *float processed*, the goal of the initial experiment was to classify the glass types based on the given attributes and also to determine whether the glass was a type of "float" glass or not, as some of the glass types are classified as *float processed*. Based on this, is expected to learn how to predict refractive index based on the other attributes and predict glass type based on the given attributes.

From the first inspection carried out in this report, it has been seen some outliers but not extreme ones, it has been decided then to carry on without removing any of them for the future classification task. Moreover, despite most of the attributes do not seem to be highly correlates, it looks like the refractive index does show a positive correlation with Calcium and a negative correlation with Silicon.

5 Exam questions

Q1 - answer A:

The reasoning is that attributes $x_2 - x_7$ are **ratio** as the value 0 means absence of occurrences. The attribute y is **ordinal** because it can be ordered, 1 is less of a congestion than 3, but the distance can't be measured between 1-4 levels of congestion. Lastly, the attribute x_1 is **Nominal**. Even though it may look like it is interval because it is time, so there is no absence of time (no 0 value) and we can measure difference in time, in this specific situation we don't really care about the time itself but if some road events happened in the same interval

of time, therefore I would say it is Nominal.

Q2 - answer A:

Reasoning is following:
Below there is a formula for p-distance:

- General Minkowsky distance (p-distance)

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}$$

- Max-norm distance ($p = \infty$)

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_M - y_M|\}$$

Figure 7: Formula for p-distance dissimilarity measure.

M is the length of vector. We used this formula and got a correct answer for case A (p=∞), so we did not even do further calculations. (|26-19| = 7 and |2-0|=2 but 7 is bigger)

Q3 - answer A:

We use following formula:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\boldsymbol{X}\|_F^2} = \frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

Figure 8: Formula for variance explained by first K number of principal components.

Where: K - number of PCs taken into consideration, M - number of all PCs.
Calculations:

$$\frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.867 \tag{1}$$

Q4 - answer D:

The sign of the coefficients corresponding to each of the variables influences the projection onto the component. Time of the day has a negative coefficient, what means a low value will have a positive projection, a broken truck has a positive coefficient so a high value would impact positively, and so on.

Q5 - answer A: We have the formula for Jaccard similarity:

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{11}}{K - f_{00}}$$

Figure 9: Formula for Jaccard coefficient, the measure of similarity.

Where K is the size of vocabulary which is M=20000, the f_{11} stands for the attributes that occur in both documents ($x_k = y_k = 0$). There are two of those "the" and "words". The f_{00} stands for attributes that do not occur in any of the documents. There are 15 unique words in both sentences, two are the same so 15-2=13 then 20000-13=19 987. Now we can calculate the answer which is 2/13=0.153846.

Q6 - answer D: The probability of $x_2 = 0$ when y=2 is 0.81 + 0.03. But we have to keep in mind that probability for y=2 occurring is 0.23. So the probability for such situation that there is light congestion (y=2) and $x_2 = 0$ is 0.84*0.23 = 0.1932.

References

[1] https://archive.ics.uci.edu/dataset/42/glass+identification. *Glass Identification dataset*. 1987. URL: https://archive.ics.uci.edu/dataset/42/glass+identification. (Accessed: 30.09.2023).