Group 1

1. Alireza Faghihi Moghaddam
2. Adam Rokah
3. Alexander Verdieck

TASK 1: Reading the data

What data type have you assigned to attribute *id?*

The assigned data type for ID is int64.

What do you think is the practical consequence of setting this data type?

Since IDs are usually integers and not floating numbers, it makes sense to cast them also this way. This allows for slightly fast operations, because floating point operations are more complex, than integer operations.

What are the average length of sepals (sl) and their standard deviation?

Average sepal length (sl): -5.705507692307693
Standard deviation (sl): 303.7889483450809

TASK 2: database preprocessing

How many instances are there for each class?

Virginica: 2998

Setosa: 2996

Versicolor: 500

TASK 3: data cleaning

Why is it important to let the system know which values are missing?

As mentioned in the course content, poor data quality can affect many data processing efforts. One of which is the missing values in our data. Having missing values in data introduces inaccuracies in clustering and classification and endangers the validity of analyses. According to [1], the proportion of missing values in a dataset can play a significant role in how we select our appropriate method. While less than 5% is still manageable, missing values between 5%-15% require more sophisticated approaches for handling missing data.

What are the average length of sepals (sl) and their standard deviation after declaring missing values (3.1)?

Before declaring the missing values, we had the length and standard deviation as follows:

Average sepal length (sl): -5.705507692307693
Standard deviation (sl): 303.7889483450809

For the average length of speals, we calculated the mean, as it gives a standard average length of the data.

Average sepal length (sl): 3.5275947028025865
Standard deviation (sl): 2.1024922333854033

What is the average length of sepals (sl) and their standard deviation after removing outliers (3.2)?

Average sepal length (sl) after removing outliers: 3.5202833821038038
Standard deviation (sl) after removing outliers: 2.018405706390843

Do you think the outliers you have removed were noise (that is, wrong measurements) or unusual but correct observations?

According to the instructions, Missing values have been coded using -9999 in the input file, and after removing all of them from each feature, we can the dataset length drops from 6500 rows to 6494.
After calculating the z-score for all the features, we removed the extreme values as they basically deviated too much.

Original shape: (6494, 6)
After removing outliers: (6487, 6)

After removing the outliers, we can see that only approximately 500 values were removed among the 4 features, so this can't be an error in gathering the data, but simply unusual but correct observations.

Would you first handle missing data and then remove outliers, or the other way round? Why?

Handling the missing values first is more beneficial, as they have a direct impact on distort statistics (mean, std) that are used to detect outliers. If we compute the z-score while some values were coded as -999, the mean and std would be completely wrong.

Once the dataset is clean, we can measure the central tendency and spread of our data more reliably.

Assume your observations (records) represent people in a social network, and one variable stores their degree centrality. Would you remove outliers in this case? why?

In this case, the degree centrality can be an important factor that depends on the person. There can be a variety of connections in social media, based on the function of their account, which means that data is also important and cannot be overlooked or labelled as outliers; however, if the data is too skewed, it can impact the analysis.

The best practice is to preprocess the data so that the values are more centred.

TASK 4: data transformation

What are the average length and standard deviation of sepals after min-max normalization?

SL mean before MinMax Normalization = 3.5275947028025865
SL Standard Deviation before MinMax Normalization= 2.1024922333854033
SL mean After MinMax Normalization = 0.05433455583272086
SL Standard Deviation After MinMax Normalization= 0.04188231540608299

What are the average length and standard deviation of sepals after standardization?
SL mean before Standardization = 3.5275947028025865
SL Standard Deviation before Standardization= 2.1024922333854033
SL mean After Standardization = 0.05433455583272086
SL Standard Deviation After Standardization= 0.04188231540608299

How many components have been selected after 4.3?

After running the PCA on our data, we see that by keeping the first component, we will be able to explain about 91% of the data, adding the second PCA, can increase our explainability to approximately 95.3%. Keeping the first two components is enough to keep the majority of the information.

How much variance is captured by the first two components?
PC1 = 0.91329008
PC2 = 0.04017737
Which is approximately 0.95346745 or 95.3%.

How is the first component defined as a combination of the original attributes?
Generally, each Principal Component is a linear combination of the original attributes
PC1=a1.sl + a2.sw + a3.pl + a4.pw
The coefficients are then chosen to explain the maximum variance.
These coefficients are chosen in such a way that:

1. PC1 explains the maximum variance possible in the dataset.

2. The vector of coefficients is constrained to have unit length (to avoid trivial scaling).

3. Subsequent components (PC2, PC3, …) are defined similarly but are required to be orthogonal to the previous ones while still capturing as much remaining variance as possible.

Thus, PC1 is the single direction (linear combination of the original features) along which the data shows the greatest variability.

How many components would have been selected after 4.4 (that is, with an attribute expressed on a larger range)?
The 4.4 Section shows the average absolute loadings per feature ratio. The Coefficients tell how strong each feature contribute to that PC.
The following is the result of 4.4 :
Sepal L 0.430356
Sepal W 0.353662
Petal L 0.441237

Petal W 0.385186

But the decision to delete components should be based on the variance ratio; this, with the first two components, PC1, 2, give us 97%.

How many components would have been selected after 4.5 (that is, with an outlier)?

TASK 5:

	Simple sampling	Bootstrapping	Stratified (5.3)	Stratified (5.4)
Number of iris versicolor	12	7	250	50
Number of Iris setosa	76	64	1498	50
Number of iris virginica	62	79	1499	50
Are there repeated identifiers?	No	Yes	No	No
Does the number of iris versicolor included in the sample change if you change the local random seed?	yes	yes	No	No

REFERENCES

[1] Shafiq Alam, Muhammad Sohaib Ayub, Sakshi Arora, Muhammad Asad Khan, An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity, Decision Analytics Journal, Volume 9, 2023, 100341,ISSN 2772-6622,https://doi.org/10.1016/j.dajour.2023.100341.