Lab1 Report

Mohamed Elmaghraby Mohamed

55

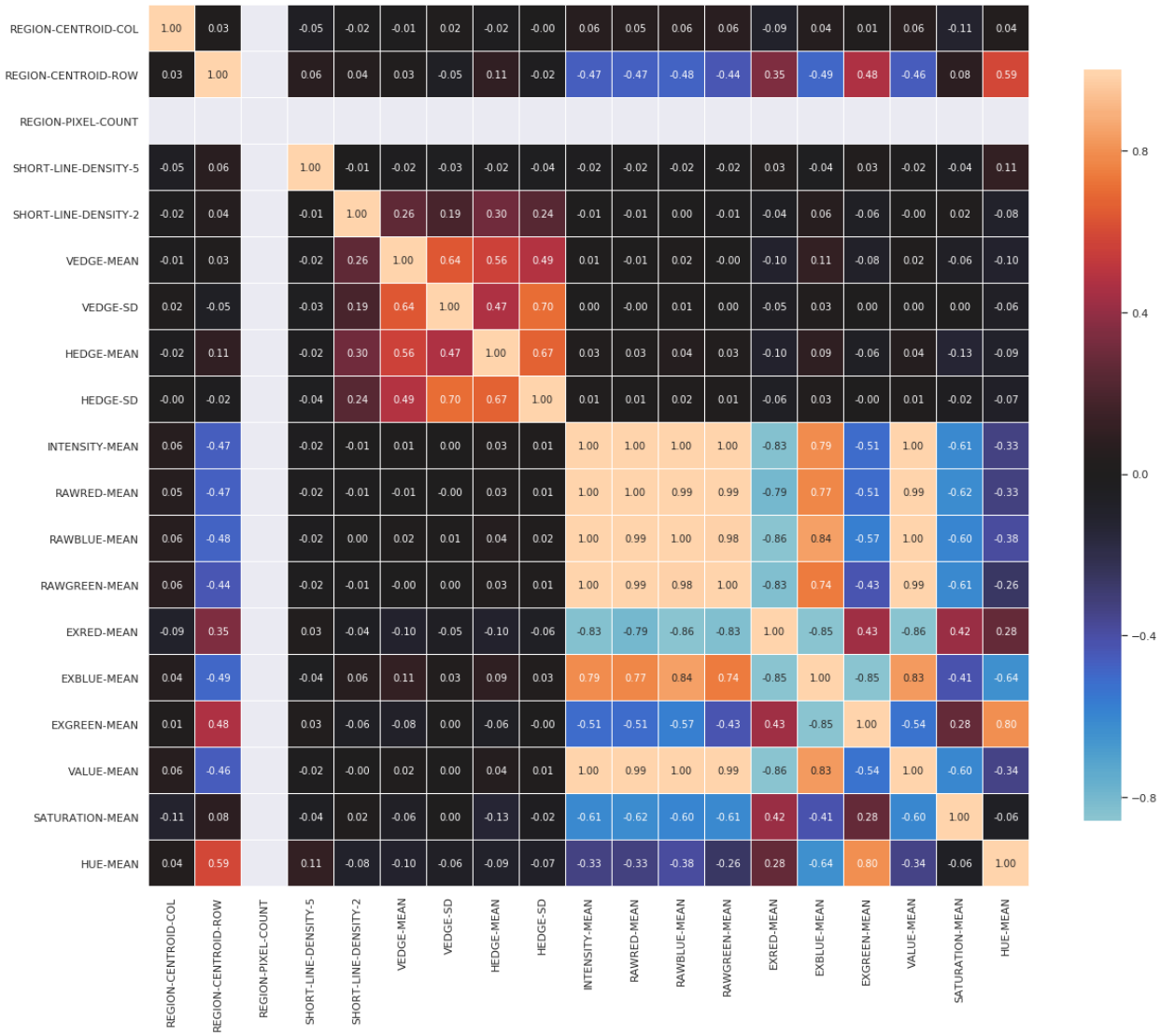
Prepare data:

- Import segmentation data and test then merged them.
- Skip the first two rows as they contain info about data.
- Merge both data sets.
- Add index column and class column.
- Explore data :
    - Number of instances = 2310
    - number of attributes = 19
    - class names : ['BRICKFACE' 'SKY' 'FOLIAGE' 'CEMENT' 'WINDOW' 'PATH' 'GRASS']
    - number of classes: 7

Data Exploration:

- Plot histogram with bins [1, 5, 10, 15]
    - Each figure contain 7 images, one for each class
    - Each image show all attributes with different colors
- Plot Box plot for the data
- Use `.corr(method = 'Pearson')` to get the Pearson's correlation between each 2 attributes.

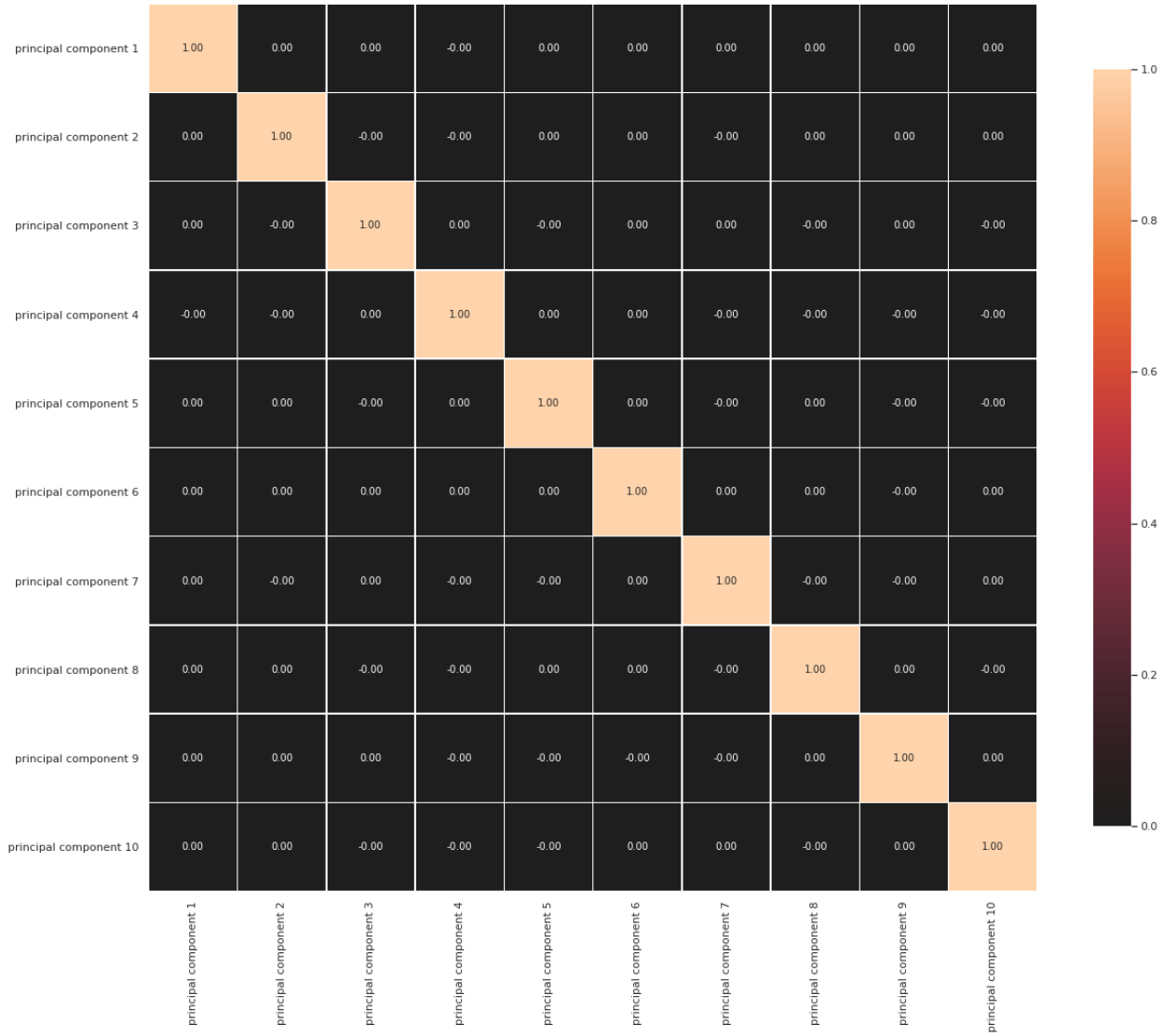
- Draw the correlation using heatmap.



Data Preprocessing:

- Normalization:
  - Min-max scalar:
    - Calculate min-max scalar using `preprocessing.MinMaxScaler() from the scikit-learn library`
    - Then plot box plot for resulting data
    - Conclusion: The result data spread on less scale, so values are more close to each other.
  - Z-score:
    - Calculate z-score using `zscore(x) from scipy library`
    - Then plot box plot for resulting data
    - Conclusion: The result data spread on less scale, so values are more close to each other.

- Dimensionality reduction
    - Feature Projection:
        - Use `PCA from the scikit-learn` library seeking to sum of variance ratio we need of 0.95.
        - Number of PCA components = 10
        - Variance ratio = [0.42341135, 0.16203649, 0.09959451, 0.05857283, 0.05197997, 0.05050372, 0.04041415, 0.03120143, 0.02999802, 0.02195028]
        - Plot correlation matrix for result data.



Conclusion

- Principal component analysis convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components
- less than half of the attributes can describe the data with 0.95 percent.

- Feature selection
  - Use `SelectKBest from the scikit-learn` library with k = 10.

- Plot correlation matrix for result data.

Conclusion

- Feature Selection selects those features which contribute most to your prediction variable or output in which you are interested in.
- These features are highly correlated.