Deakin University

SIG720- OnTrack Submission

Task C1

Submitted by
[Guruprasadh J P]
[223939745]
[Attempt # 1]
Date 23 July 2023

Target Grade : Credit

Task Details –

Part - I

Datafile: SCADI.csv

Data Description: This dataset contains 206 attributes of 70 children with physical and motordisability based on ICF-CY. For more information click this link.

1. Determine the number of subgroups from the dataset using attributes 3 to 205 i.e.,exclude attributes 1, 2 and 206.

29 subgroups from the dataset using attributes 3 to 205 are inferred.

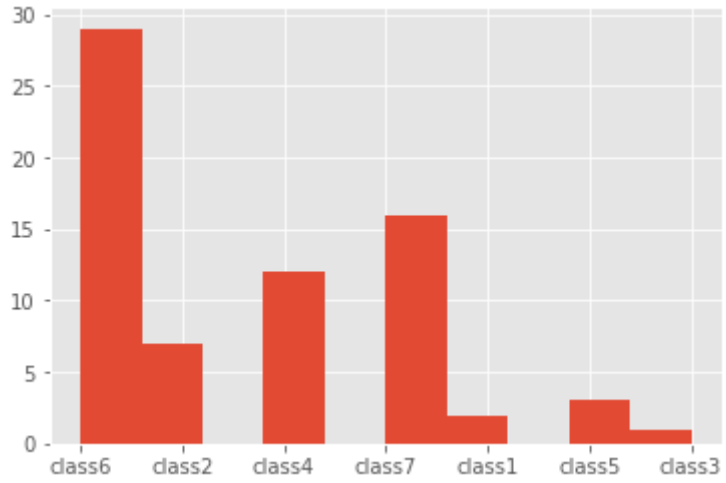
Those 29 subgroups names are {'5302', '5600', '5404', '5402', '5101', '5400', '5200', '5102', '5100', '5203', '5403', '5401', '5201', '5700', '53000', '53001', '57022', '53010', '57020', '57021', '5602', '5501', '53011', '5500', '5205', '571', '5701', '5202', '5204'}

Twenty-nine activities in seven categories are mentioned below are considered as self-care activities,

1. Washing oneself
2. Caring for body parts
3. Toileting
4. Dressing
5. Eating
6. Drinking
7. Looking after one’s health and looking after one’s safety.

Is this number the same as the number of classes presented by attribute 206? Explain and justify your findings.

7 classes present in attribute 206 are ['class6' 'class2' 'class4' 'class7' 'class1' 'class5' 'class3']



29 activities and further split 7 subgroups to determine the seven-class criteria. mentioned below

{'5100': 7, '5101': 7, '5102': 7, '5200': 7, '5201': 7, '5202': 7, '5203': 7, '5204': 7, '5205': 7, '53000': 7, '53001': 7, '53010': 7, '53011': 7, '5302': 7, '5400': 7, '5401': 7, '5402': 7, '5403': 7, '5404': 7, '5500': 7, '5501': 7, '5600': 7, '5602': 7, '5700': 7, '5701': 7, '57020': 7, '57021': 7, '57022': 7, '571': 7}

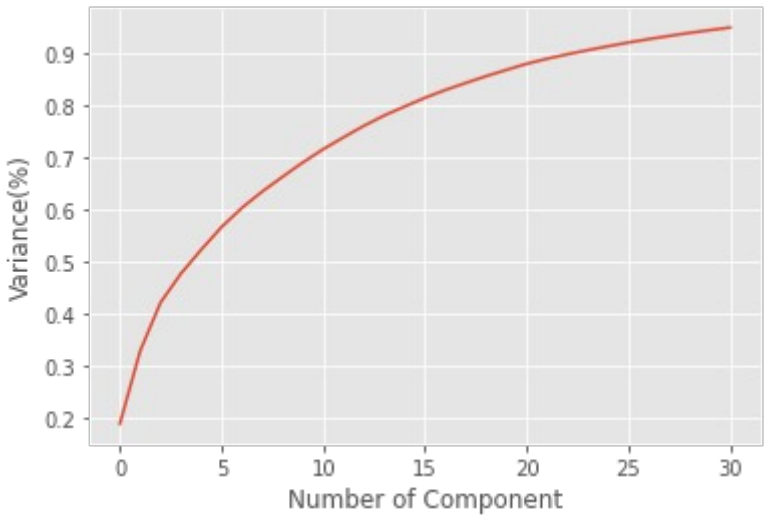
Those 29 activites in self-care activities are measured using the below mentioned seven-class criteria is same as the number of classes presented by attribute 206.

1. NO impairment
2. 2MILD impairment
3. MODERATE impairment
4. SEVERE impairment
5. COMPLETE impairment
6. NOT Specified
7. NOT Applicable

2. Is this data facing the curse of dimensionality? If so, then how to solve this problem.

Yes. Data is facing the curse of dimensionality.

Explain with a two-dimensional plot and report relevant loss of information.

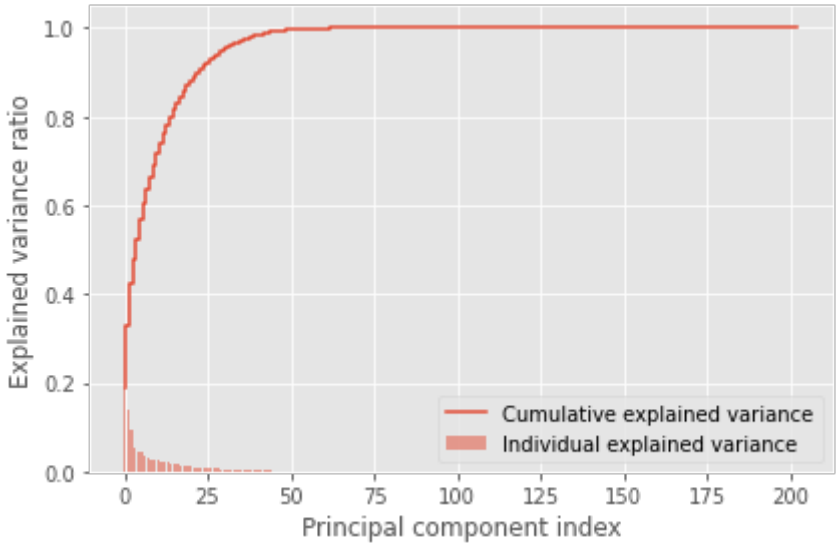


Number of component > 30, the variance is almost reaching 1. Thus the remaining components of features 30 to 203 are mere loss of information.

3. After applying principal component analysis (PCA) on a given dataset, it was found that the percentage of variance for the first N components is X%.

26.82% of variance for the first component

How is this percentage of variance computed?



- Determine the covariance matrix
- Determine eigenvalues and eigenvectors
- Determine explained variance
  - **Compute percentage of variance:**

        - **Percentage of Variance of PC_i = (Eigenvalue of PC_i) / (Sum of all Eigenvalues)**
- Plot the explained variance against cumulative explained variance
    - Cumulative percentage of variance is often calculated by summing up the percentage of variance for each principal component from the first to the current one. This helps in understanding how much total variance is explained by considering the first k principal components.

Part – II

Dataset filename: obesity_levels.csv

Dataset description: This dataset includes data for the estimation of obesity levels in individuals
based on their eating habits and physical condition. The data contains 17 attributes and 2111
records.

Features and labels: The attribute names are listed below. The description of the attributes can be
found in this article (web-link).

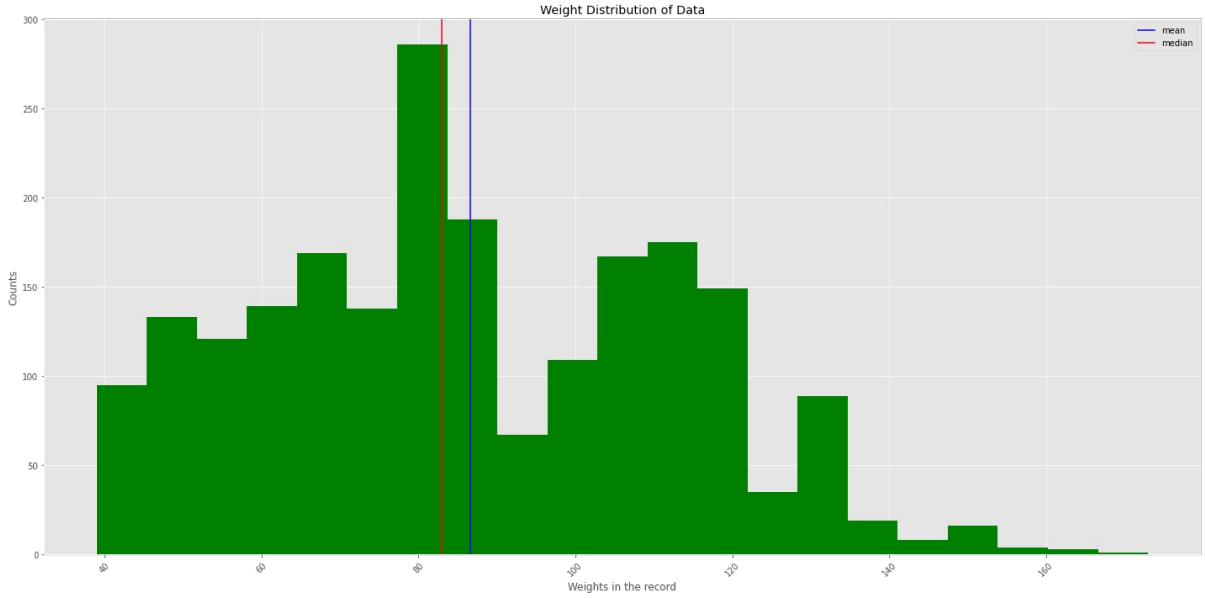
4. Create a machine learning (ML) model for predicting “weight” using all features except “NObeyesdad” and report observed performance. Explain your results based on following criteria:

a. What model have you selected for solving this problem and why?

I have chosen linear regression model since it's simple one to evaluate. But by seeing the results **r2_score is only 56%**. It turn out to be worst model I have chosen for this type of problem.

b. Have you made any assumption for the target variable? If so, then why?

Yes. The above histplot distribution seems to be bi-model type. The mean and median are shifted to left because larger number of people above 250 counts are weighing 80 Kg.



c. What have you done with text variables? Explain.

All the categorical data or test variables are converted numerical data using label encoder technique.

d. Have you optimized any model parameters? What is the benefit of this action?

Yes. I used ridge regression techniques to see any improvement on accuracy. But it turns out to worst prediction with r2_score of 51%.

e. Have you applied any steps for handling overfitting or underfitting issues? What is that?

No, I haven't tried for this. Since I have to consider different ML model technique. For that model I will work on handling overfitting or underfitting issue.