MA515: Foundations of Data Science

Project Report

Submitted to- **Dr. Arun Kumar**

TA's- **Naman Krishna Pande,**
Monika Singh

Harshdeep (2021MCB1044)
Vrinda Dua (2021MCB1223)
Chinmay Thakur (2021MCB1233)
Mohit Soni (2021MCB1238)

26 nov, 2023

Table Of Contents

1. SUMMARY

The data analysis project is a comprehensive exploration of machine learning and dimensionality reduction techniques applied to different datasets. The project is structured into four tasks, each addressing distinct challenges and employing various methodologies:

Task 1: Predicting Default

- **Methods Used**: Decision Tree, Random Forests, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA).
- **Objective**: Predict whether a person will default on a task.
- **Analysis**: The findings from each method are compared, considering confusion matrices and accuracy. This task provides insights into the strengths and weaknesses of different classification algorithms.

Task 2: Dimensionality Reduction with PCA and SVD

- **Methods Used**: Principal Component Analysis (PCA), Singular Value Decomposition (SVD).
- **Objective**: Reduce the dimensionality of Data 2 to 5, 10, 15, and 20 columns using PCA and SVD.
- **Analysis**: The reduced datasets are examined, and exploratory data analysis (EDA) is performed on the SVD data with 5 columns. This task aims to understand the impact of dimensionality reduction on the structure and characteristics of the data.

Task 3: Predicting Crime Rate

- **Methods Used**: Linear Regression, Linear Discriminant Analysis (LDA).
- **Objective**: Predict whether a given suburb has a crime rate above or below the median.
- **Analysis**: The findings from linear regression and LDA are compared, using confusion matrices and accuracy. This task explores the effectiveness of regression and classification methods in predicting crime rates.

Task 4: Image Compression with SVD

- **Methods Used**: Singular Value Decomposition (SVD).
- **Objective**: Compress a given image to 25%, 65%, and 85% of the original size using SVD.
- **Analysis**: The project examines the trade-off between compression ratio and image quality. Visual inspection and quantitative metrics are used to evaluate the effectiveness of the compression technique.

---

2. PREDICTING DEFAULT

- *Here, we first describe the data, and see the correlation matrix corresponding to the Default data:*
- *Shape of the data: (10000, 5)*

	Unnamed: 0	balance	income
count	10000.000	10000.000	
mean	5000.500	835.374886	
std	2886.896	483.714985	
min	1.000	0.000	
25%	2500.750	481.731105	
50%	5000.500	823.636973	
75%	7500.250	1166.308386	
max	10000.000	2654.322576	

Fig. 1- Co-variance matrix

- *Now, we first split the given data in training data and testing data. Then, after feature scaling, we fit each of the*
  1. *Decision tree*
  2. *Random Forest*
  3. *K-nearest neighbors*
  4. *Linear Discriminant Analysis*

above mentioned models to our data and, the results obtained are same for each model, confusion matrix and accuracy is given below:

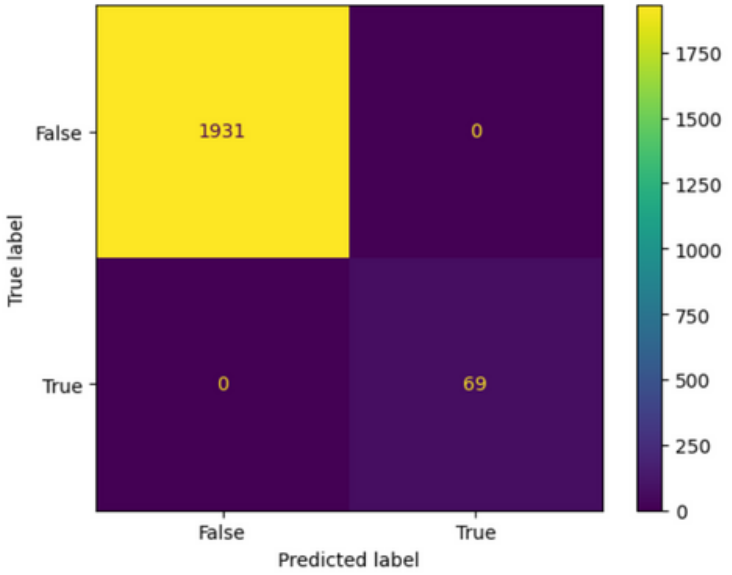

Fig. 2- Confusion matrix (same for all models)

Accuracy of

1. *Decision Tree- 100%*
2. *Random Forest- 100%*
3. *K-nearest neighbors- 100%*
4. *Linear Discriminant Analysis- 100%*

3. DIMENSION REDUCTION

3.1 Data Description

In this part, we are doing basic data exploratory data analysis, and plotting histograms corresponding to the data.
Shape of the data: (2501, 51)
Also, we check the correlation among the data in the correlation matrix:

	1	2	3	4	5	6	7	8	9	10	...
count	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	2500.000000	...
mean	19.436420	19.974677	20.207614	20.359998	20.034612	19.957323	19.431995	19.656091	20.055373	19.531695	...
std	20.114557	20.947014	20.748332	21.525254	20.606300	20.690533	20.148964	21.100771	21.475233	20.104812	...
min	-9.378292	-9.875271	-9.622198	-9.074825	-9.300815	-9.182675	-9.322297	-9.707890	-9.442658	-9.151600	...
25%	5.828970	6.043946	6.181921	5.707541	6.026202	6.136381	5.789500	5.711160	5.661808	6.088311	...
50%	14.289522	14.656679	14.907659	14.566654	14.886255	14.847056	14.268863	14.240981	14.130472	14.204394	...
75%	27.741491	28.217891	29.341570	29.230626	28.626721	28.501603	27.915492	28.385531	28.813072	28.642910	...
max	165.964596	210.494579	203.680246	164.252158	153.816566	226.742274	152.931236	166.618063	161.400935	157.045780	...

8 rows × 50 columns

Fig. 3-Description of the data for HD Data 2

Now, we plot histograms for the data features for better understanding. A few of them are:

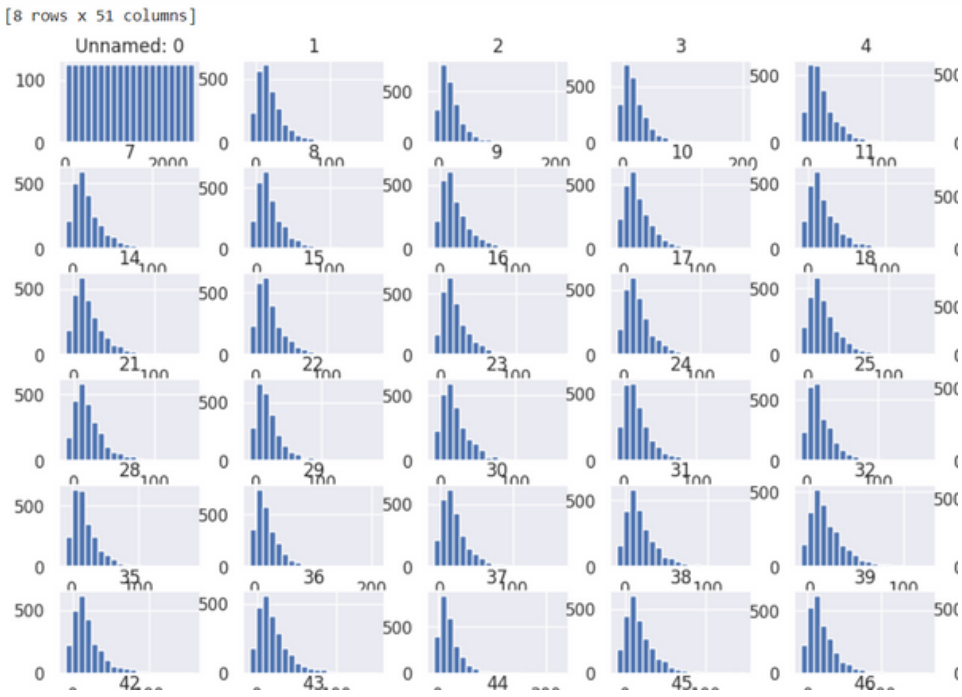

Fig.4 Histograms for a part of data

3.2 PCA, SVD on Data

Now, we standardize the data, and apply PCA on it with respect to 5, 10, 15, and 20 components using Singular Value Decomposition.

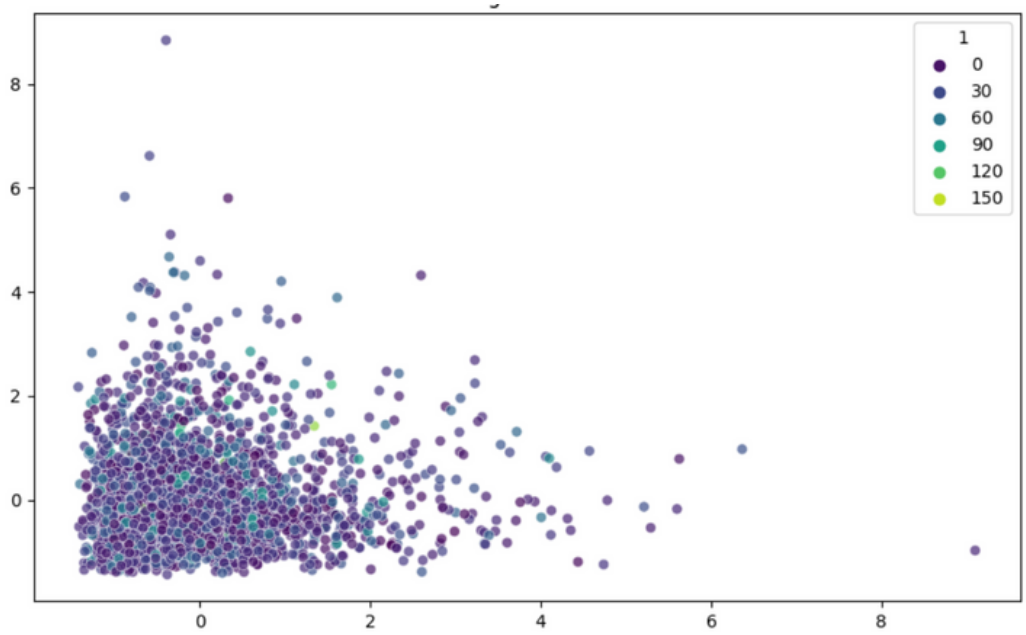

Fig. 5 - Original Data

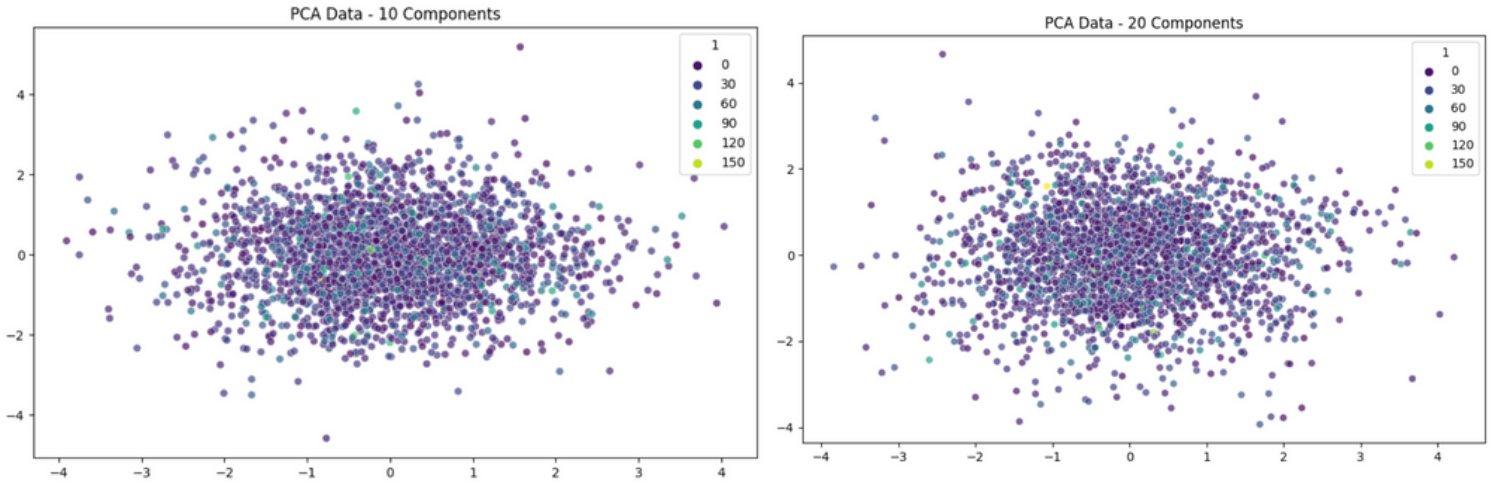

Fig. 6- PCA data with 10, 20 components

Note: Other graphs with PCA data having 5, 15 components are plotted in the main code file.

For the PCA with 5 components, the Exploratory Data analysis is given below:

Exploratory Data Analysis on SVD data with 5 columns:

	SVD_Component_1	SVD_Component_2	SVD_Component_3
count	2.500000e+03	2.500000e+03	2.500000e+03
mean	3.996803e-19	1.598721e-18	2.059464e-18
std	2.000400e-02	2.000400e-02	2.000400e-02
min	-7.788103e-02	-7.379958e-02	-8.499368e-02
25%	-1.269769e-02	-1.315262e-02	-1.257207e-02
50%	5.473256e-04	-2.886829e-05	6.841995e-05
75%	1.314972e-02	1.294197e-02	1.354836e-02
max	6.243884e-02	9.072575e-02	5.845942e-02

	SVD_Component_4	SVD_Component_5
count	2.500000e+03	2.500000e+03
mean	-1.054712e-18	-3.330669e-20
std	2.000400e-02	2.000400e-02
min	-8.452152e-02	-6.588902e-02
25%	-1.303313e-02	-1.396009e-02
50%	-3.783542e-04	-7.458752e-04
75%	1.262618e-02	1.286081e-02
max	7.074119e-02	8.144980e-02

4. CRIME-RATE PREDICTION

4.1 Data Description

In this part, we describe the data in the majorly used terms.
Shape of the data: (506, 14)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

Fig. 7- Description of Boston Data

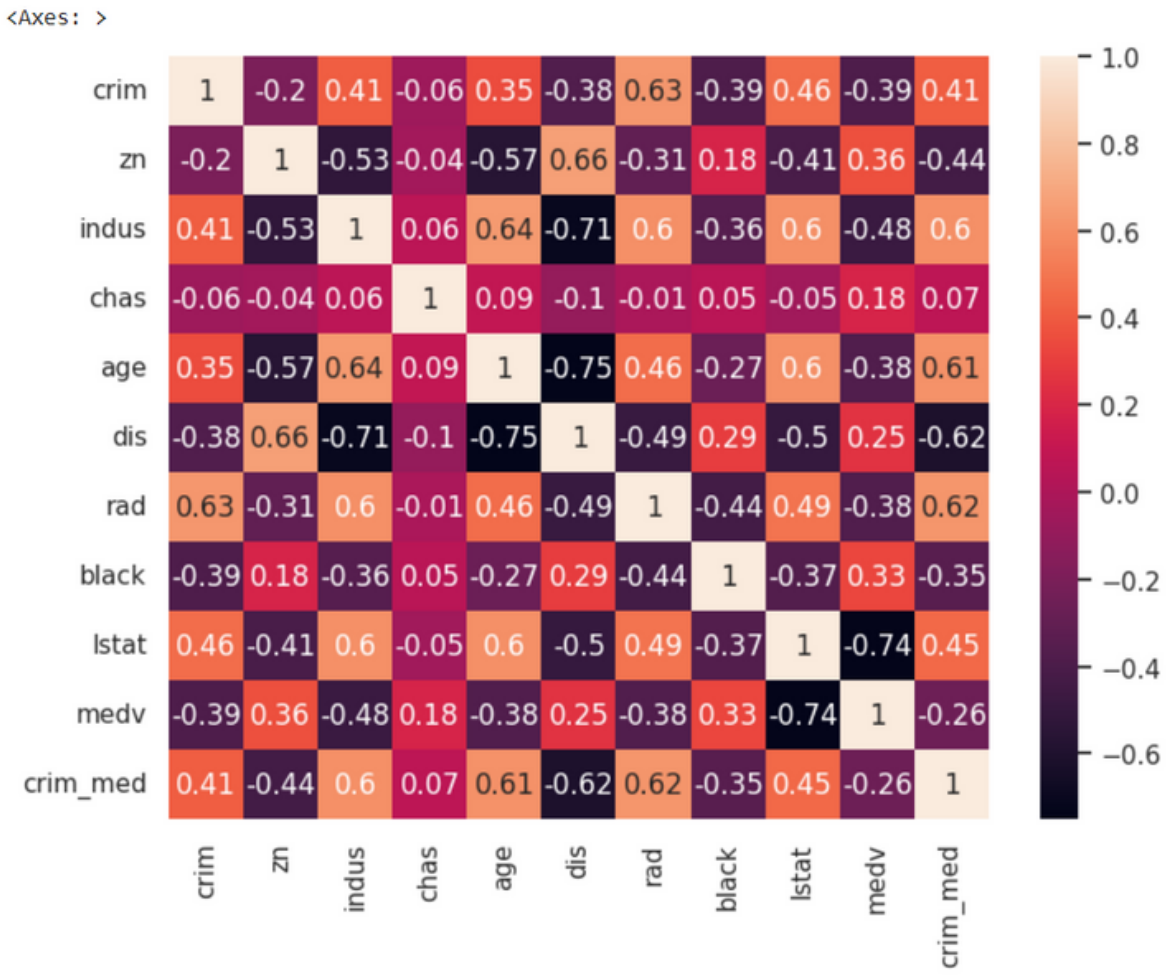

Fig. 8- Heatmap showing Correlation matrix

4.2 Data Pre-processing

Here, we are removing the columns which are having higher correlation with the "crim" column (since this is the column which we will work on).
So, we drop the columns "rm", "ptratio", "nox", and "tax" from the data.

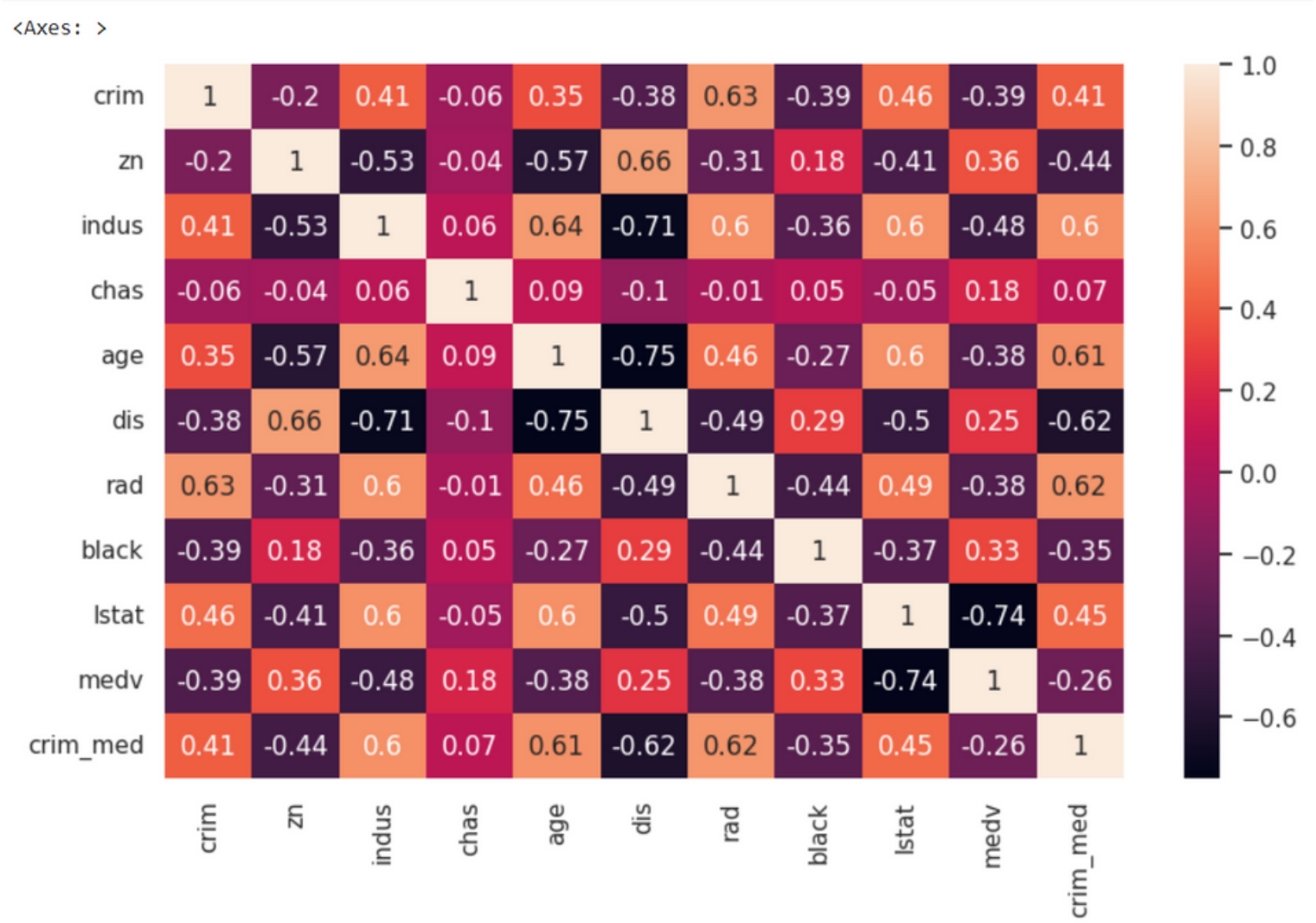

Fig. 9- Heatmap after removing highly correlated columns

4.3 LDA and Linear Regression

Now, we split the data-set into training and testing data, and corresponding to each data-set, we do linear regression and LDA, and the results are as follows:

In Linear Regression for test data-

- *R-squared: 0.25*
- *Mean squared error: 0.19*
- *The Linear Regression coefficients are: [0.00037981 -0.00094907 0.01280105 -0.00412964 0.00525867 -0.01168553 0.0195593 -0.00036874 0.00392519 0.00824843].*
- *The value of intercept is: -0.23925016148496625*

Talking about the results for Linear Regression and LDA, the results for the training data and the testing data are both same:

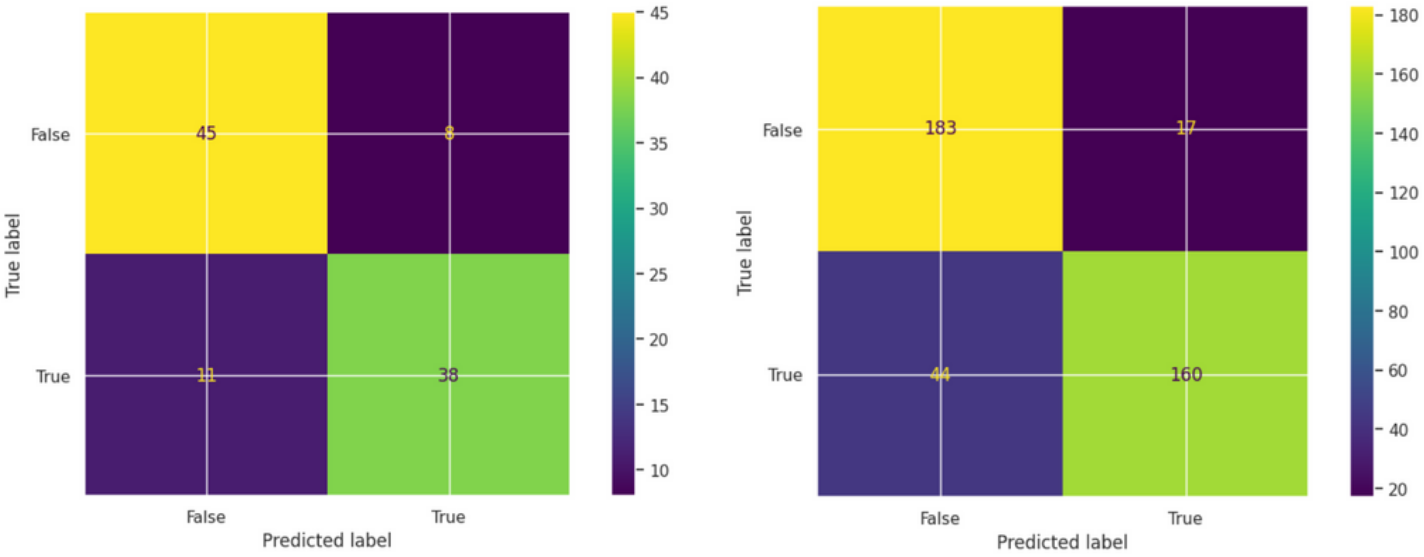

Fig. 10- Confusion matrix from LDA and Linear Regression for (a) Testing data (b) Training data

*For the training data, the accuracy attained in case of Linear Regression and Linear Discriminant Analysis is **84.90%.***

*For the testing data, the accuracy attained in case of Linear Regression and Linear Discriminant Analysis is **81.37%.***

5. IMAGE COMPRESSION (SVD)

Fig. 11 - Original image to be compressed

SVD was employed to decompose the original image matrix into three matrices: U, Σ (diagonal matrix of singular values), and V^T (transpose of matrix V).
Singular values were retained based on the 25%, 65%, and 85% compression ratios.
The reconstructed image was obtained by multiplying the selected components of U, Σ, and V^T matrices.

The image was first converted into greyscale image, and then was compressed.

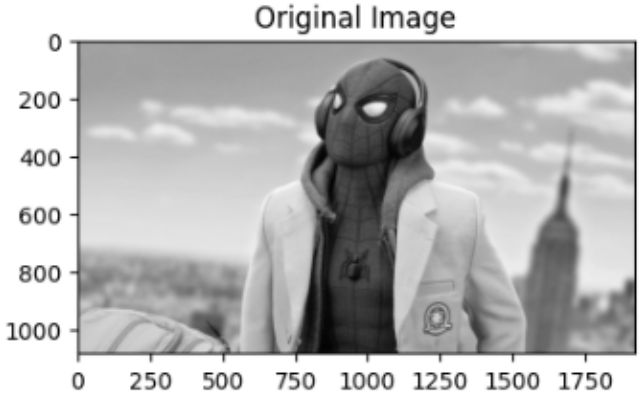

Fig. 12- Greyscale image

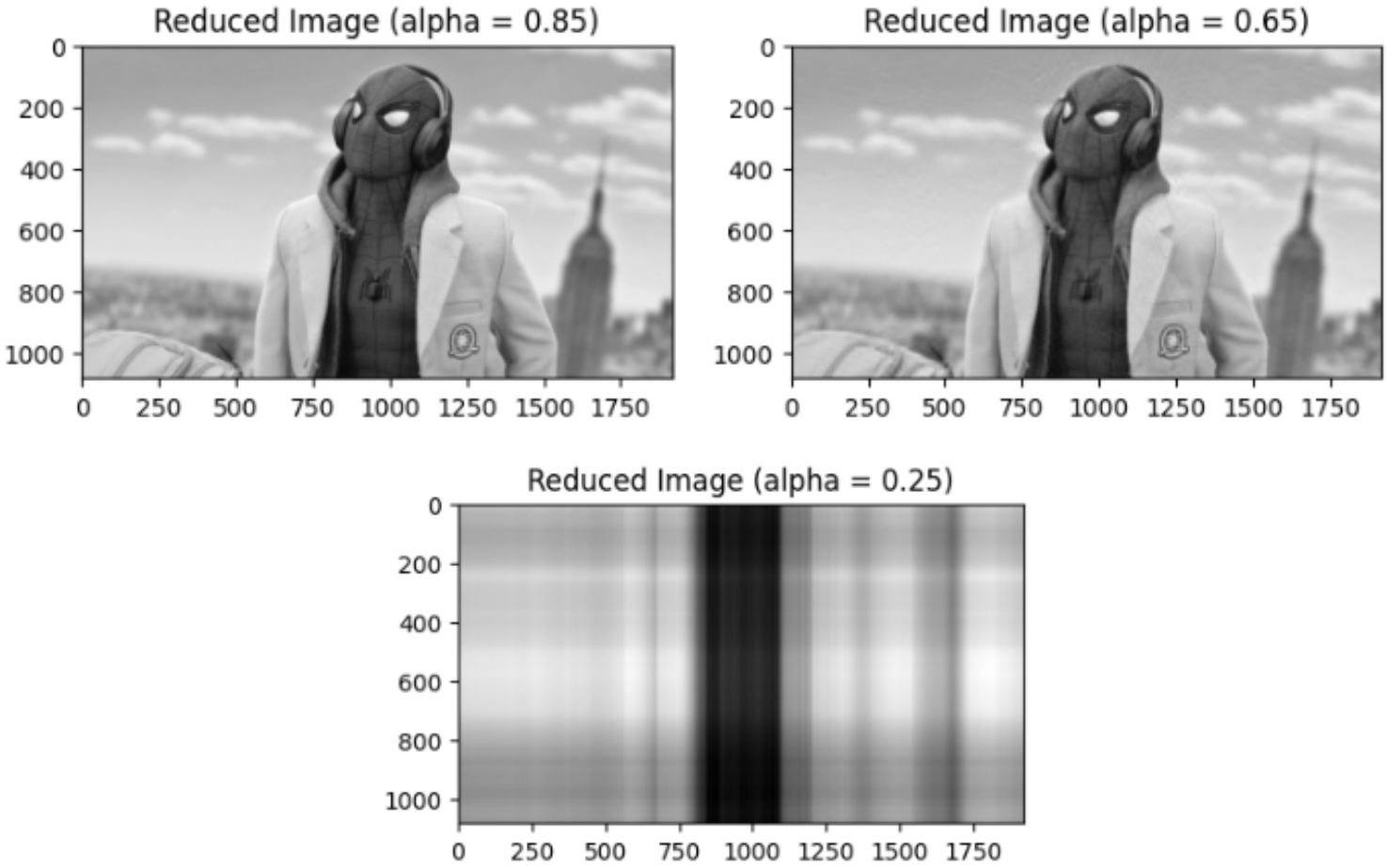

Fig. 13- Compressed images with 85%, 65%, and 25% of its size respectively

Observations-

- *At 25%, noticeable loss of detail was observed, especially in regions of fine texture and edges.*
- *The 65% compression ratio maintained a good balance between compression and image quality.*
- *The 85% compression retained almost all visual details, making it visually similar to the original.*

6. CONCLUSION

Task 1-

- Decision trees provide a straightforward understanding of the decision-making process, but their tendency to overfit might limit generalization to new data.
- Random Forests mitigate the overfitting issue of decision trees and generally offer improved predictive performance, making them a robust choice.
- KNN provides simplicity and flexibility, but its sensitivity to outliers might affect predictive accuracy.
- LDA is effective when classes are well-separated, but assumptions might limit its performance in certain scenarios.

*The accuracy for all the models for Default data is **100%.***

Task 2-

- The application of Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) to reduce the dimensionality of Data 2 has yielded valuable insights. The reduction was performed to obtain datasets with 5, 10, 15, and 20 columns, followed by an exploratory data analysis (EDA) specifically focused on the data reduced to 5 columns using SVD.
- The use of PCA and SVD for dimensionality reduction has proven effective in condensing the information in Data 2 while preserving its essential characteristics.

Task 3-

- The choice between Linear Regression and LDA depends on the underlying data characteristics and the trade-off between interpretability and predictive accuracy.
- Linear Regression offered an intuitive interpretation of the relationship between features and crime rates.
- LDA, by classifying into binary categories, provided a clearer understanding of the likelihood of crime rates falling above or below the median.

The accuracy of both the models for Boston test data is **84.90%.**

Task 4-

- *Compression to 85%: Retained almost all visual details, closely resembling original.*
- *Compression to 65%: Balanced compression, maintaining acceptable visual quality.*
- *Compression to 65%: Balanced compression, maintaining acceptable visual quality.*
- *The choice of compression ratio is crucial, balancing between achieving smaller file sizes and maintaining acceptable visual quality.*

6. INDIVIDUAL CONTRIBUTION

1. *Use decision tree, random forests, KNN and LDA to predict whether the person will default. Compare the findings from different methods. -* ***Vrinda***

2. *Use PCA and SVD to reduce the dimension of the data 2 so that the reduced data has 5, 10,15 and 20 columns. Give a exploratory data analysis on SVD data with 5 columns. -* ***Chinmay Thakur***

3. *Use linear regression and LDA on data 3 to predict whether a given suburb has a crime rate above or below the median. Compare the findings from different methods. -* ***Mohit Soni***

4. *Use the given image and compress it to 25, 65 and 85% of the original image using SVD. -* ***Harshdeep***

5. *Project Report- All members.*

7. REFERENCES

1. *Decision Tree*
2. *Random Forest*
3. *KNN algorithm*
4. *LDA*
5. *PCA*
6. *SVD*
7. *Linear Regression*