Massive Data Processing

Final Assignment

Group **Shoppe**
La Gia Hiep, Nguyen Duy Anh, Pham Long Duy Tien, Pham Quang Tung, Pham Tuyen

Ton Duc Thang University

May 8, 2023

Student List

Name	Student ID	Task(s) Assigned	Contribution%
La Gia Hiep	**521K0133**	Task 2, Task 1	100%
Nguyen Duy Anh	**521K0126**	Task 3 and code refinements	100%
Pham Long Duy Tien	**520K0220**	Slides, Task 2 refinements, Task 5	100%
Pham Quang Tung	**520K0265**	Task 4	100%
Pham Tuyen	**520K0337**	Task 2 and code refactors	100%

Table: Student List

Note

Appending **@student.tdtu.edu.vn** will give you the email.

Outline

Outline

Illustration: What did we do?

- Using `PySpark.sql` and `matplotlib.pylot` libraries, we are able to construct a dataframe.
- With this dataframe, we are able to illustrate the MNIST dataset visually.

Illustration: How the program worked

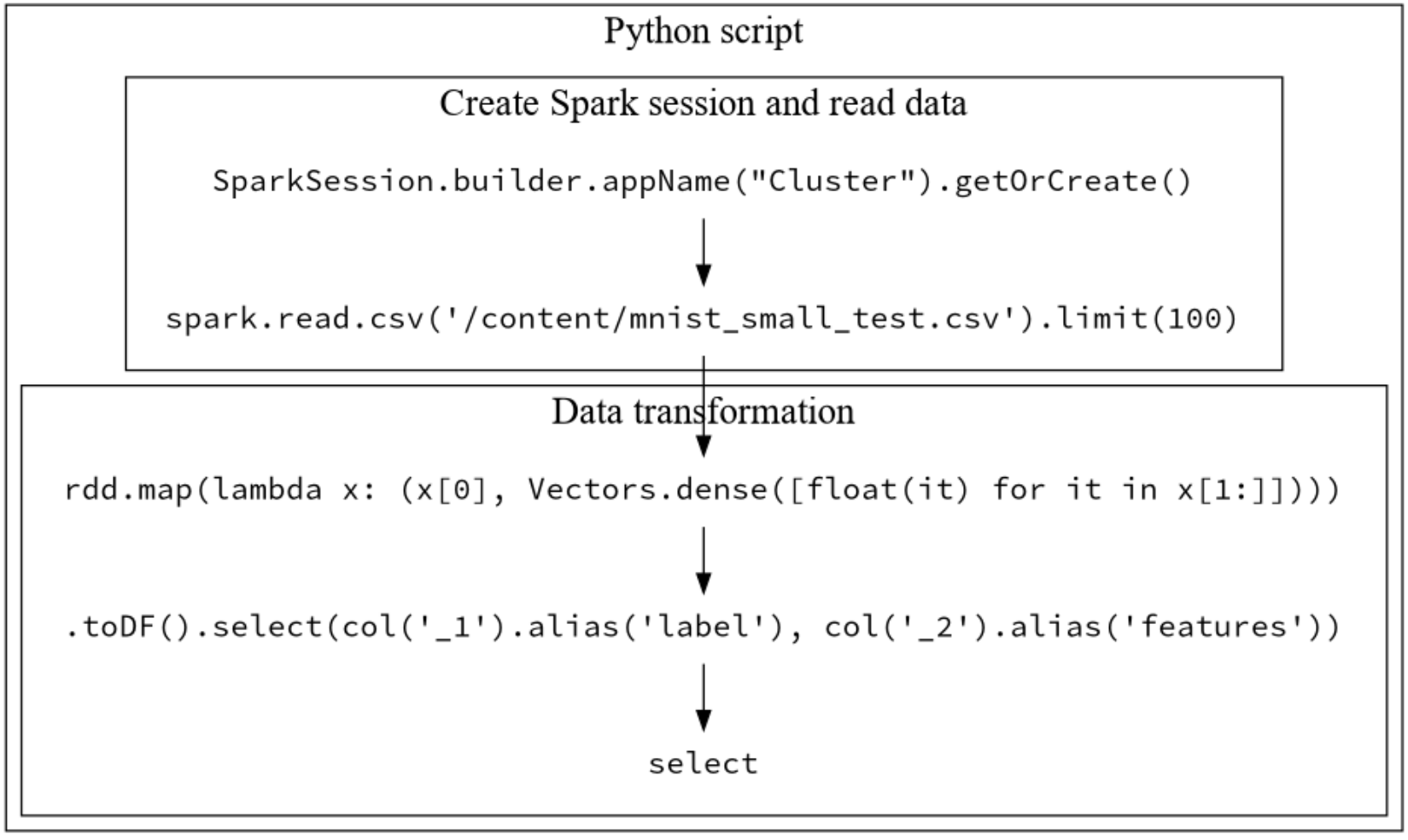
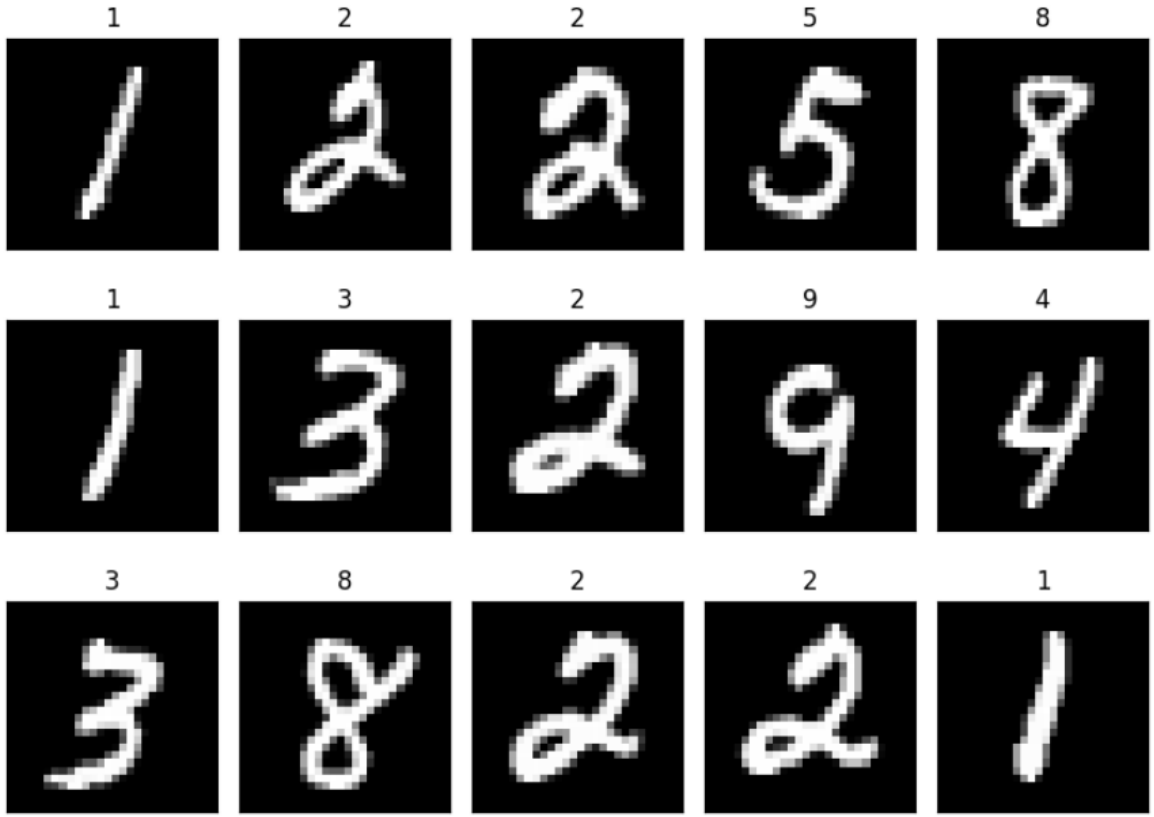


Illustration: Results



Clustering: What did we do?

- We training three models for K-means, each for K-values of 15, 10 and 5.
- After training, we used these models to place a centroid.
- Using the points to form the centroid, we used the summation of distance to the centroid.

K-means advantages

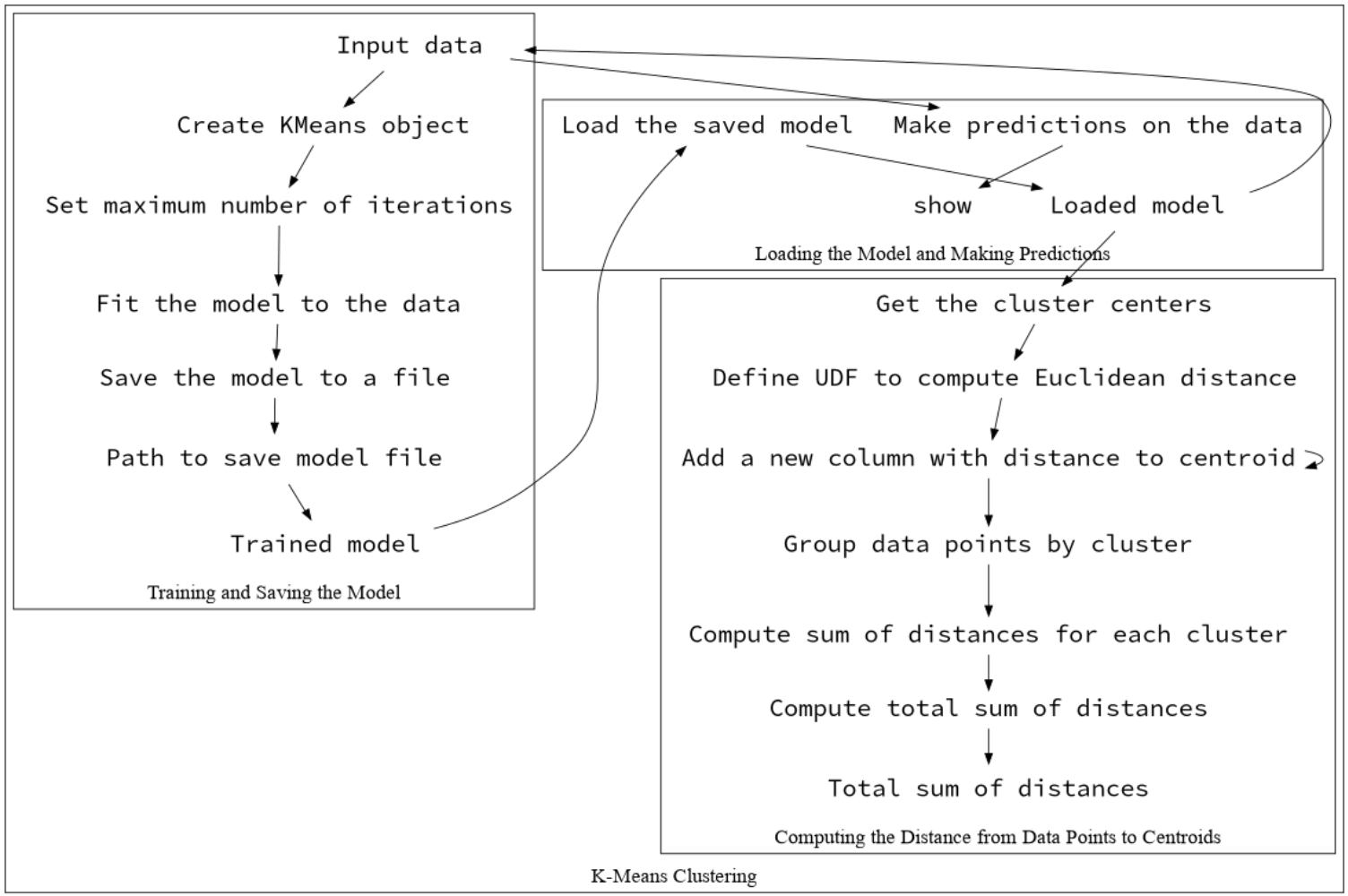
- Simple and widely used.
- Computationally efficient.
- Flexible.

K-means disadvantages

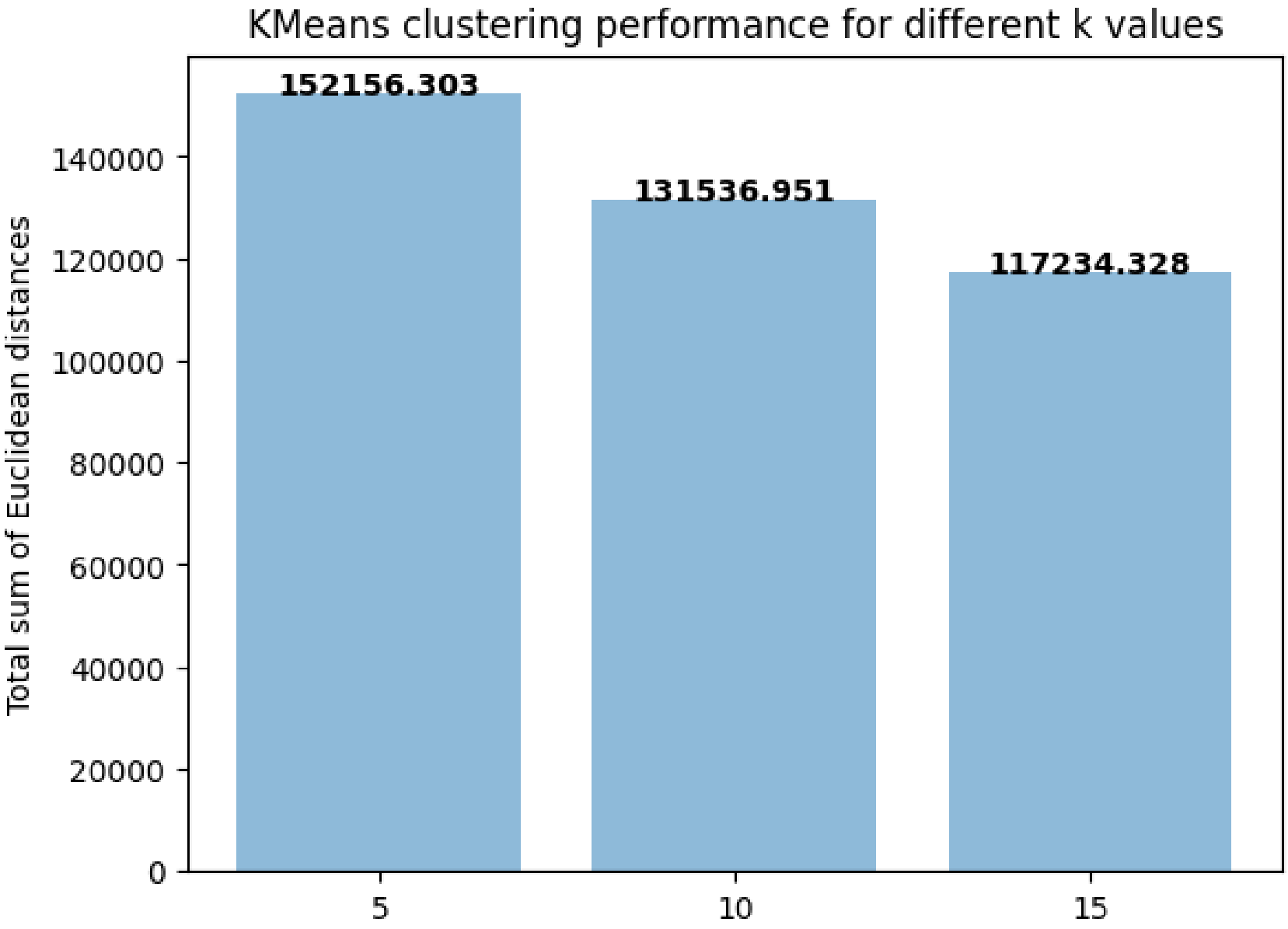
- K needs to be specified in advance.
- Sensitive to initial placement of centroid.
- Perceptible to outliers.

Clustering: How the program worked

Note: $k = 15$ shown, similar work done for $k = 10$, $k = 5$.



Clustering: Results



Outline

Dimensionality Reduction: What did we do?

- We utilized SVD operations to compute the matices U, S and V.
- PySpark was used for additional computation methods.

Dimensional Reduction Advantages

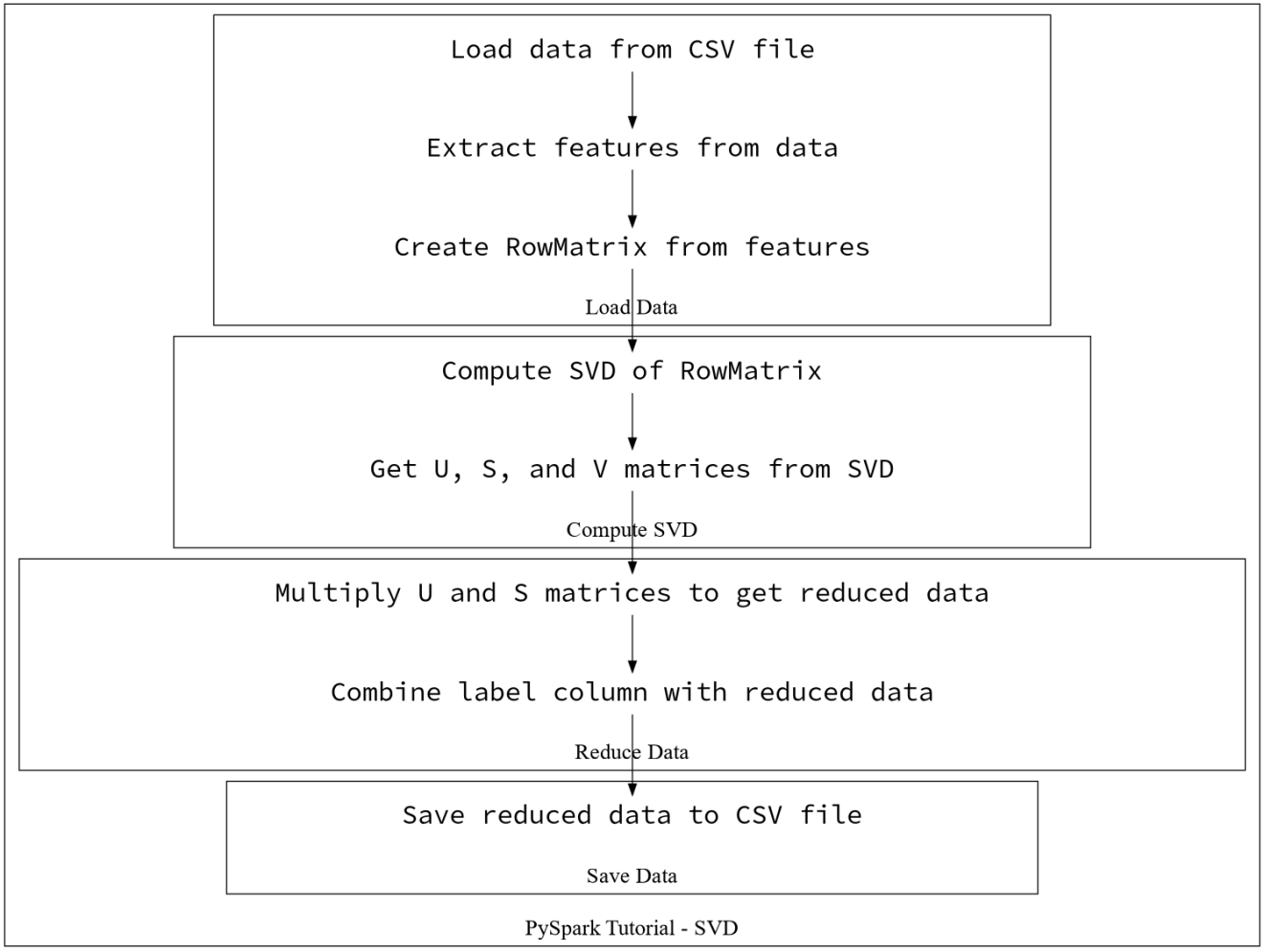
- Allows to evaluate how a model generalizes to new, unseen data.
- Helps selecting the best model.
- Prevents overfitting.
- Easier benchmarking.

Dimensional Reduction Disadvantages

- Limited data use.
- Likely to overfit.
- Influence of data distribution.

Dimensionality Reduction: How the program worked

Showing training set, the same methods were conducted for the test set.



Dimensionality Reduction: Results for Training Set

	0	1	2	3	4	5	6	7
0	7.0	-1055.005321	248.244718	-598.132258	-0.717797	385.242274	572.808262	267.941212
1	2.0	-1393.004031	-133.681947	1018.332204	81.291235	-277.464165	-392.264988	293.510614
2	1.0	-648.809758	549.281895	487.342624	288.599795	-79.728968	224.800368	188.461602
3	0.0	-2246.001287	-900.448759	-35.891545	188.103162	617.892539	-212.266997	548.128325
4	4.0	-1106.075886	-451.020604	-814.494302	40.166073	-410.897242	8.945943	-268.763500
...	...	...	...	...	...	...	...	...
6995	4.0	-1361.527432	301.652469	-394.134595	607.092674	98.815843	-175.863989	-576.632165
6996	9.0	-1326.202228	712.617455	-506.453987	183.673109	316.255873	35.445115	7.813972
6997	4.0	-1368.608628	804.714173	110.261574	187.989534	226.395297	43.804819	-280.780714
6998	9.0	-1661.566897	547.111909	-643.050432	385.335951	574.413094	22.622496	-210.931038
6999	4.0	-1284.795657	781.550433	-131.658866	391.221679	66.947658	-42.162357	-406.627145

7000 rows × 197 columns

Dimensionality Reduction: Results for Test Set

	0	1	2	3	4	5	6	7
0	1.0	-719.336823	-517.676027	-560.968574	-244.345334	211.322198	107.687917	379.441332
1	2.0	-1705.390328	-266.338804	-332.129837	-623.960334	-2.108565	28.689877	19.115734
2	2.0	-2218.073320	47.072433	-21.154111	-184.794019	-117.336745	-264.275664	-170.836506
3	5.0	-1752.024057	169.540720	-629.262360	702.153920	148.800581	630.382267	-371.175394
4	8.0	-1591.575614	-676.191137	-423.445397	443.421250	-306.867662	100.827341	-172.311313
...	...	...	...	...	...	...	...	...
2995	2.0	-2209.420379	-390.750455	-614.307996	-8.128519	-650.550997	-780.804666	532.855222
2996	3.0	-1917.403906	241.552251	-699.503750	909.245345	-624.867327	303.093117	545.209046
2997	4.0	-1971.165299	-721.780966	690.089932	-212.919595	22.253756	5.356001	-356.491504
2998	5.0	-1578.276037	-233.523411	-23.952617	-253.942674	575.375673	557.814322	61.171699
2999	6.0	-2234.848065	1139.486930	-150.589793	-970.006129	-431.570655	406.721657	269.208301

3000 rows × 197 columns

Outline

What did we do?

- Importing data into a DataFrame, sort them by userID.
- Copied data from user #71 and Item #401 to a new test DataFrame for evaluation.

Advantages

- Personalized Recommendations.
- Scalability and Flexibility.
- Solving the Cold Sart Problem.

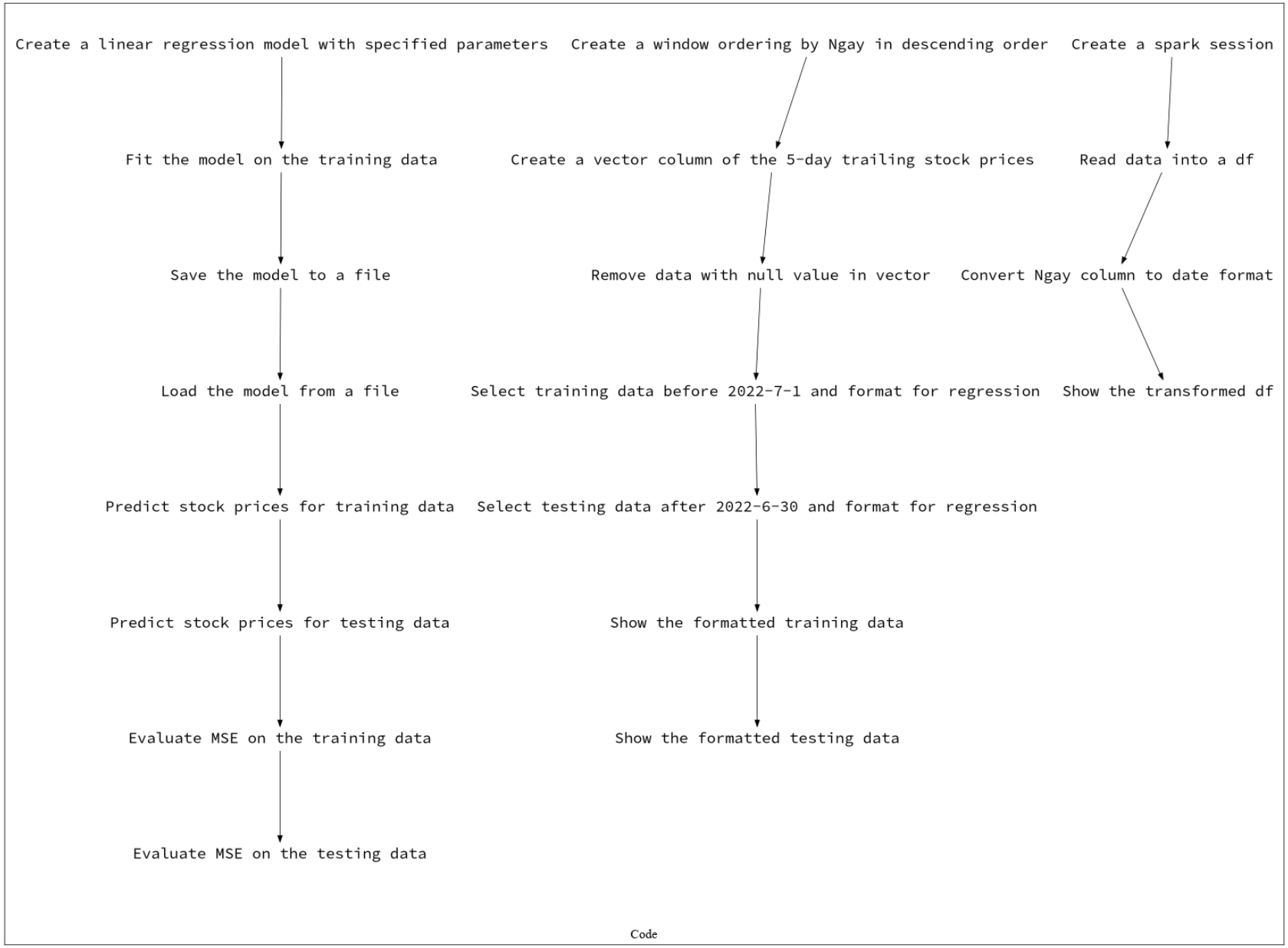
Disadvantages

- Lack of diverse data.
- Computationally intense.
- Cold start for new users.

Examples

- Shopping recommendation systems (Shoppee).
- Content recommendation (YouTube, TikTok).
- Targeted ads and demographics.

How the program worked



Results

```
Mean Squared Error = 0.0076
+-----+----+----+------+----------+
|index|user|item|rating|prediction|
+-----+----+----+------+----------+
| 2208|  72| 451|   4.0|  4.050446|
| 1898|  72| 436|   4.0|  4.005865|
|  843|  72| 406|   2.0|  2.192101|
| 1585|  72| 412|   3.0|  2.938842|
| 1208|  73| 412|   3.0| 3.0924602|
|  324|  72| 460|   3.0| 3.0378344|
| 1788|  72| 417|   2.0|  2.012427|
|  299|  73| 417|   3.0|  3.007647|
|  157|  72| 444|   3.0| 3.0443978|
|   41|  72| 435|   4.0| 4.0202446|
|  694|  72| 440|   3.0| 3.2422335|
| 1228|  72| 432|   5.0| 5.1234264|
| 2043|  72| 452|   4.0|  3.949532|
| 2132|  72| 425|   3.0| 3.0209284|
|  200|  72| 447|   4.0|  4.050446|
| 1897|  72| 462|   4.0| 3.9052527|
| 1035|  72| 437|   5.0| 4.7810526|
| 1701|  72| 456|   5.0|   4.81373|
|  797|  72| 446|   2.0| 1.9957384|
| 1587|  72| 453|   4.0|  4.050446|
+-----+----+----+------+----------+
only showing top 20 rows
```

Outline

What did we do?

- DataFrames to read to he data in descending order.
- `lead()` to get stock price of the 5 previous dates and put them into a vector.
- Use a Linear Regression model to predict stock price.
- Mean Square Error (MSE) used for further evaluation.

Advantages

- Simple.
- Interpretable.
- Low computational overhead.

Disadvantages

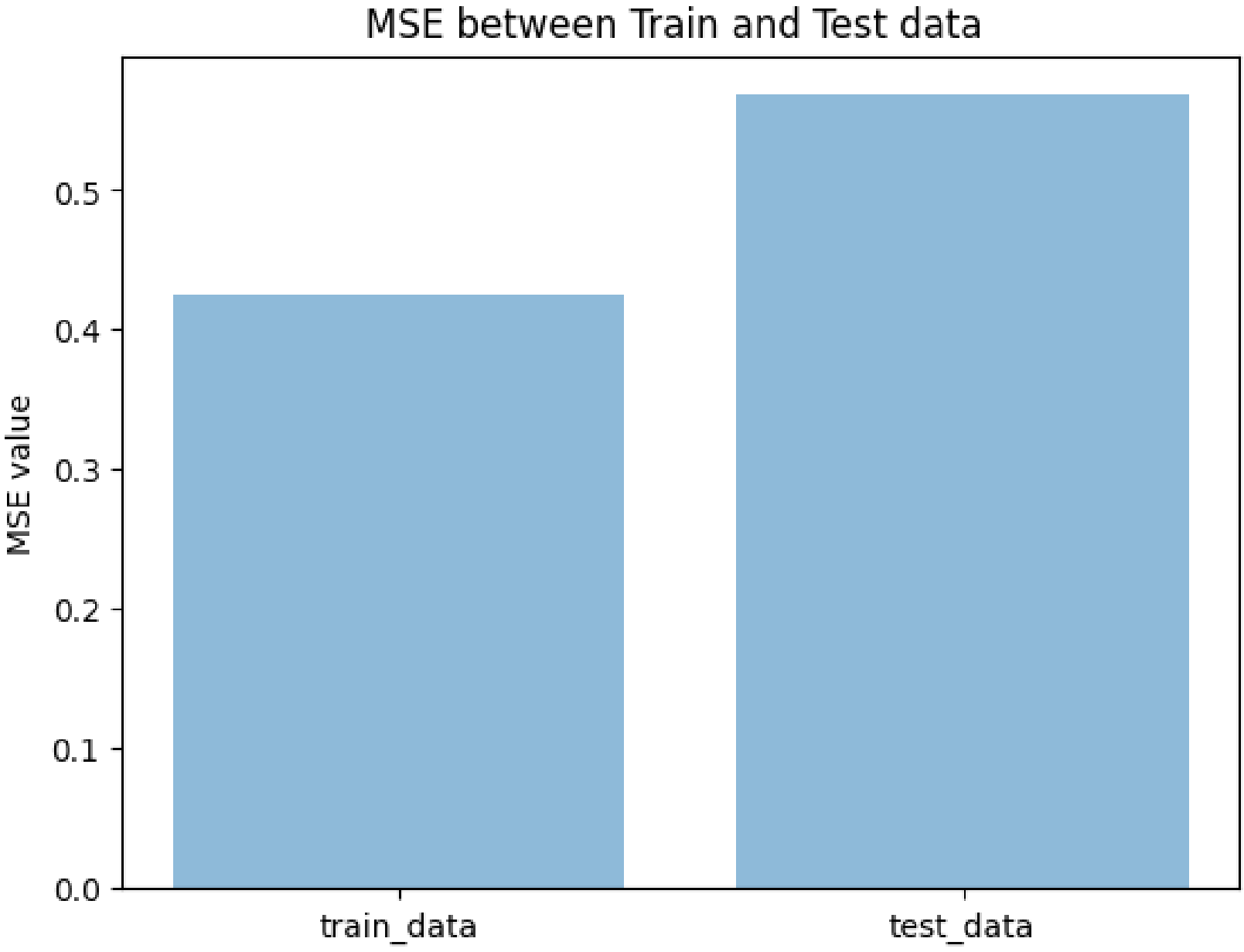
- Limited Complexity
- Sensitive to outliers.
- Limited Feature Engineering.

Examples

- Stock price prediction.
- Predicting trends.
- Marketing research.

How the program worked

Results


MSE between Train and Test data
MSE value
0.0
0.1
0.2
0.3
0.4
0.5
train_data
test_data

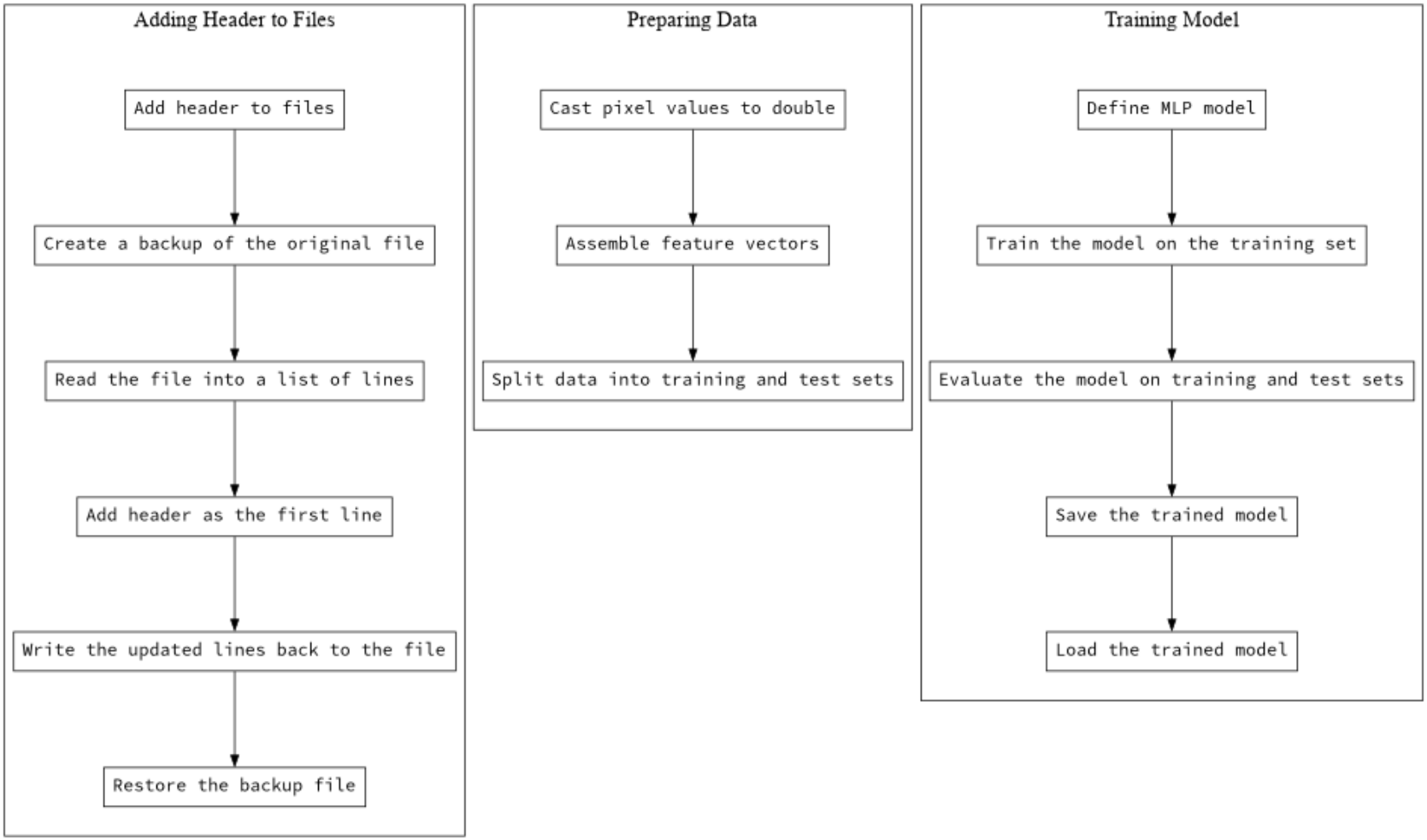
Outline

What did we do?

- Used Multi-Layer Perceptrons (MLP), Random Forest (RF), Linear Support Vector Machines (LSVM).
- Data was pre-processed.
- All models were trained with the same parameters and training conditions, all six models took an hour to train all together.

How the program worked

Note: Showing Multi-Layer Perceptron on MNIST 784-dimension dataset, RF and LSVM are not shown for brevity.



Advantages

- MLP: Flexible and usable in multiple input datatypes.
- RF: Accurate and robust.
- LSVM: Well-documented.

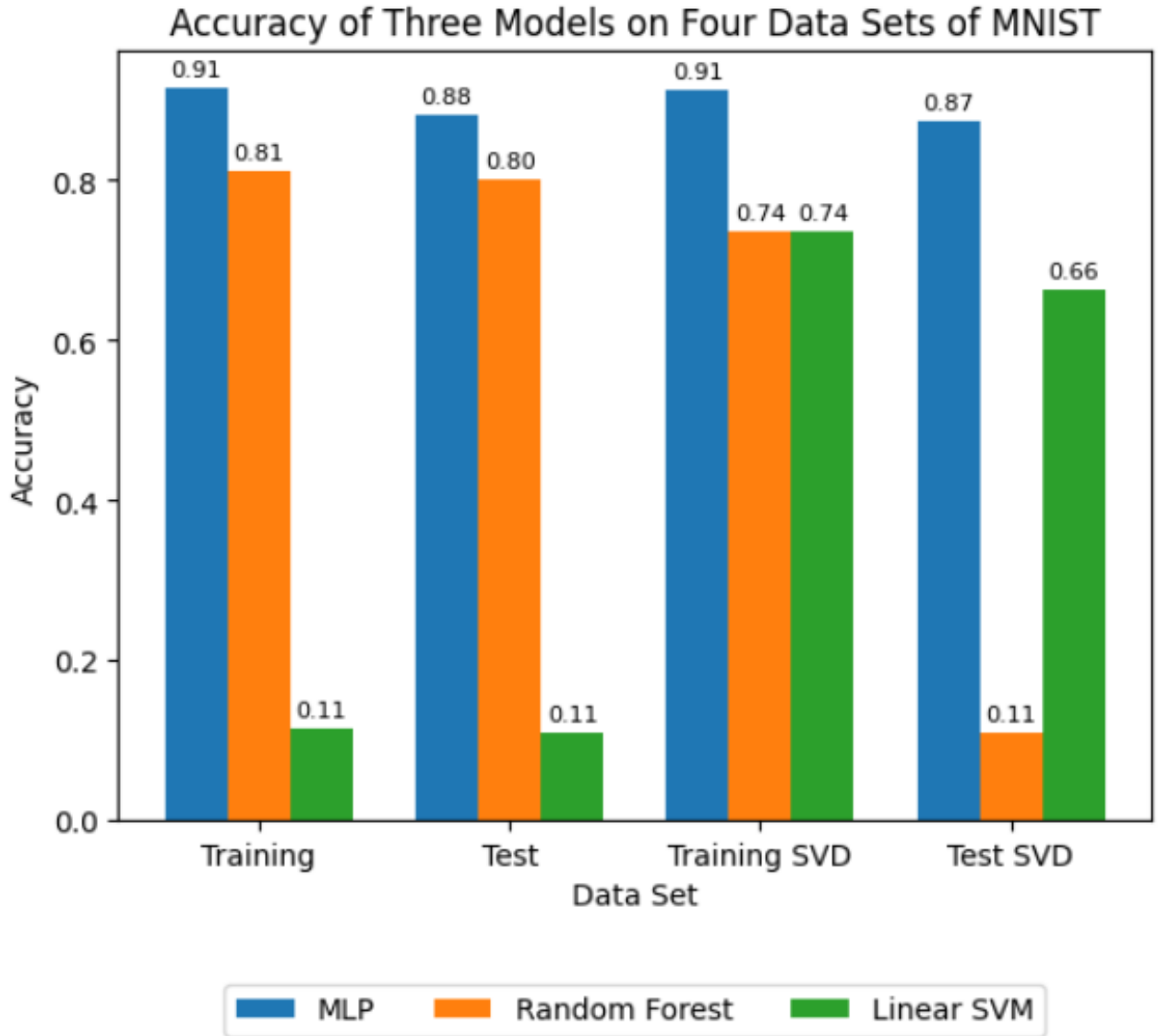
Disadvantages

- MLP: Prone to overfit.
- RF: Computationally expensive to train.
- LSVM: Sensitive to configuration settings.

Examples

- Natural Language Processing (NLP).
- Identifying disease biomarkers (cancer cells, tumors)
- Stock prediction.

Results


Accuracy of Three Models on Four Data Sets of MNIST
Accuracy
0.0
0.2
0.4
0.6
0.8
0.91
0.81
0.11
0.88
0.80
0.11
0.91
0.74
0.74
0.87
0.11
0.66
Training
Test
Training SVD
Test SVD
Data Set
MLP
Random Forest
Linear SVM

What did we learn?

- Collaboration
- Google Colab should not be Google Collab
  - Merge conflicts
  - Unable to see what the other person is doing.
- Learning more about Apache Spark and PySpark.
- How to use machine learning models.

End of slides.