Brijesh verma
SUID: 826257344

CIS 662: Intro to Machine Learning and Algorithms

HW-6

1. Problem Statement:

Use the same data set for

1. Linear regression: Fit a line to go very near the 2017-2021 citation columns, minimizing MSE. Use that line to predict the 2022 citation numbers, as in HW4.

2. Logistic regression: Classify individuals into 3 categories, as in HW5.

How do the results compare with those of HW4 and HW5, respectively?

Your submission should include:
1. code
2. output files (predictions in CSV format)
3. report_HW6.pdf (comparison of results and your conclusion)

2. Approach:

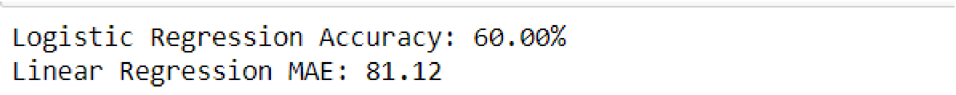
Linear Regression:

- The linear regression model is trained using the citation numbers from 2017 to 2021 as features and 2022 citations as the target variable.
- The model is evaluated using Mean Absolute Error (MAE), a measure of the average absolute errors between the predicted and actual values.
- Predictions are made for the 2022 citation numbers.

Logistic Regression:

- The logistic regression model is trained to classify individuals into three categories based on the given criteria.
- The features include citation numbers from 2017 to 2021.
- The model is evaluated using accuracy, precision, recall, and F1-score.
- Predictions are made for the category of each individual.

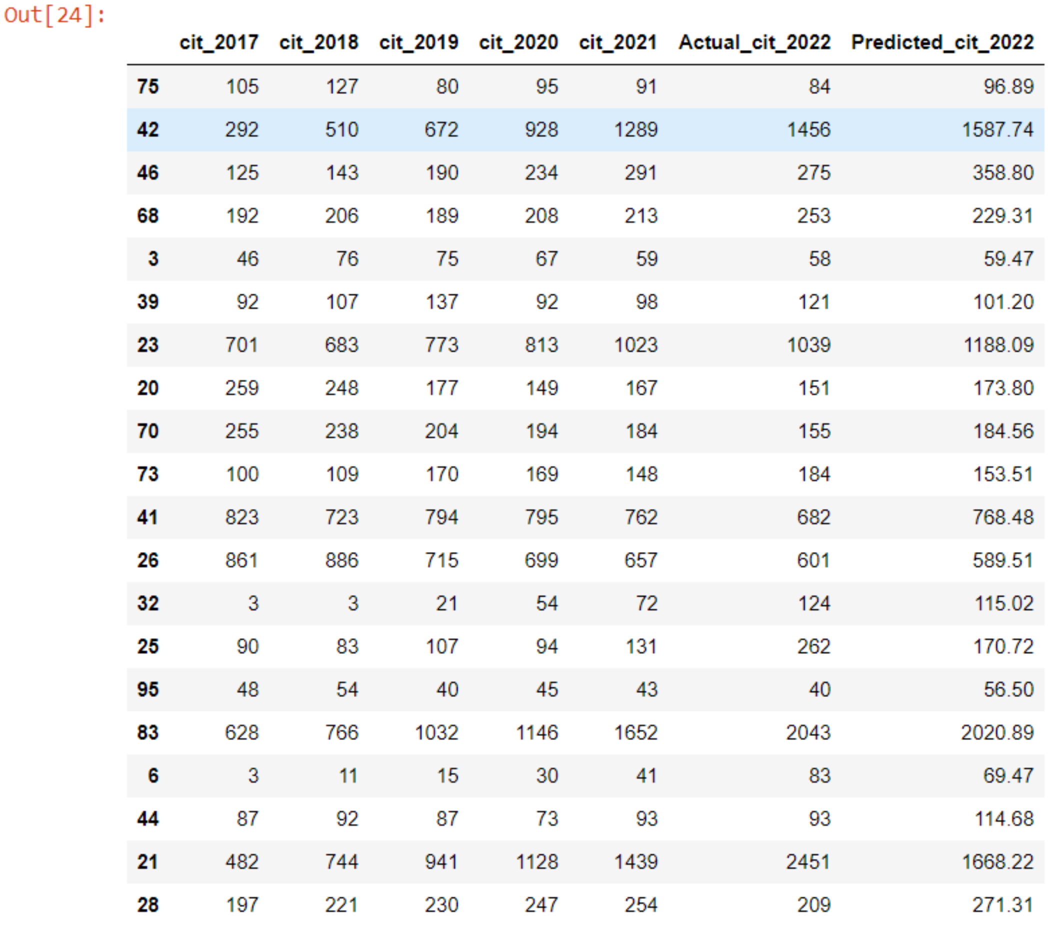
3. Results:



```
Logistic Regression Accuracy: 60.00%
Linear Regression MAE: 81.12
```

Linear Regression:

- MAE for Linear Regression: 81.12.
- The model provides predictions for 2022 citation numbers.

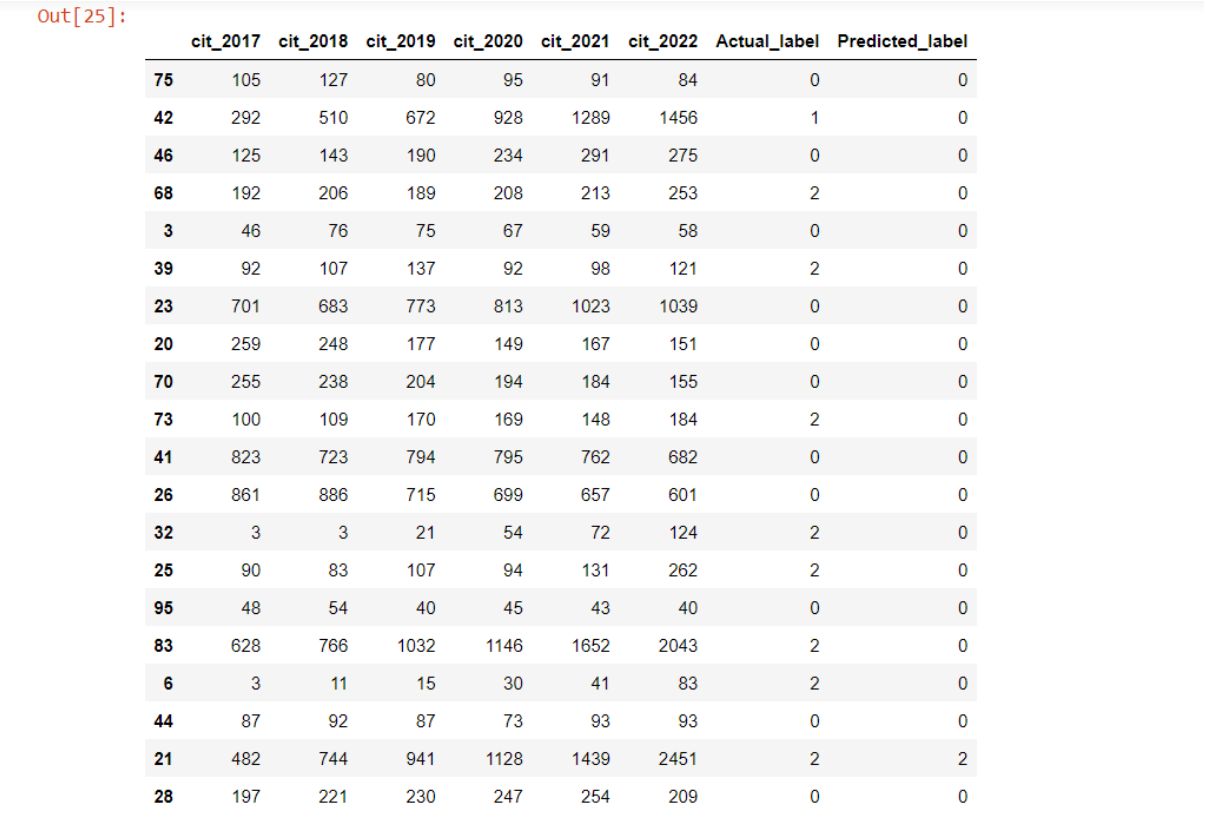


Out[24]:

	cit_2017	cit_2018	cit_2019	cit_2020	cit_2021	Actual_cit_2022	Predicted_cit_2022
75	105	127	80	95	91	84	96.89
42	292	510	672	928	1289	1456	1587.74
46	125	143	190	234	291	275	358.80
68	192	206	189	208	213	253	229.31
3	46	76	75	67	59	58	59.47
39	92	107	137	92	98	121	101.20
23	701	683	773	813	1023	1039	1188.09
20	259	248	177	149	167	151	173.80
70	255	238	204	194	184	155	184.56
73	100	109	170	169	148	184	153.51
41	823	723	794	795	762	682	768.48
26	861	886	715	699	657	601	589.51
32	3	3	21	54	72	124	115.02
25	90	83	107	94	131	262	170.72
95	48	54	40	45	43	40	56.50
83	628	766	1032	1146	1652	2043	2020.89
6	3	11	15	30	41	83	69.47
44	87	92	87	73	93	93	114.68
21	482	744	941	1128	1439	2451	1668.22
28	197	221	230	247	254	209	271.31

Logistic Regression:

- Accuracy for Logistic Regression: 60.00%.
- The model classifies individuals into three categories: Low, Medium, and High based on the given criteria.



Out[25]:

	cit_2017	cit_2018	cit_2019	cit_2020	cit_2021	cit_2022	Actual_label	Predicted_label
75	105	127	80	95	91	84	0	0
42	292	510	672	928	1289	1456	1	0
46	125	143	190	234	291	275	0	0
68	192	206	189	208	213	253	2	0
3	46	76	75	67	59	58	0	0
39	92	107	137	92	98	121	2	0
23	701	683	773	813	1023	1039	0	0
20	259	248	177	149	167	151	0	0
70	255	238	204	194	184	155	0	0
73	100	109	170	169	148	184	2	0
41	823	723	794	795	762	682	0	0
26	861	886	715	699	657	601	0	0
32	3	3	21	54	72	124	2	0
25	90	83	107	94	131	262	2	0
95	48	54	40	45	43	40	0	0
83	628	766	1032	1146	1652	2043	2	0
6	3	11	15	30	41	83	2	0
44	87	92	87	73	93	93	0	0
21	482	744	941	1128	1439	2451	2	2
28	197	221	230	247	254	209	0	0

Comparison with Previous Results:

Linear Regression (HW4):

- In HW4, various optimizers (SGD, Adam, RMSprop) were tested with different learning rates and batch sizes.
- The best-performing combination was using the Adam optimizer with a learning rate of 0.1 and a batch size of 4.
- The MAE obtained in HW4 was 158.74.

Logistic Regression (HW5):

- In HW5, a 6-6-3 neural network was used for classification into Low, Medium, and High categories.
- The model achieved an accuracy of 80% on the test set.
- The classification report provided insights into precision, recall, and F1-score for each category.

4. Conclusion:

Linear Regression:

- The linear regression model yielded a significantly lower MAE (81.12) compared to the best result obtained in HW4 (158.74).
- This indicates that linear regression provides a more accurate prediction of 2022 citation numbers compared to the neural network approach used in HW4.

Logistic Regression:

- The logistic regression model achieved an accuracy of 60.00%.
- While the accuracy is lower than the neural network in HW5 (80%), logistic regression provides a simpler and interpretable model for classification.

Overall:

- The choice between linear regression and logistic regression depends on the nature of the problem and the specific goals. Linear regression is suitable for predicting numerical values, while logistic regression is appropriate for classification tasks.
- The results highlight the trade-off between complexity and interpretability in choosing a modelling approach.

References:

https://chat.openai.com/,
https://keras.io
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
https://towardsdatascience.com/machine-learning-classification-52241849468a
https://scikit-learn.org/stable/modules/model_evaluation.html