FEATURE SCALING

Individual assignment: Data Exploration or Pre-processing Techniques(s)

Presented by Thorung Boonkaew 6510545454

CONTENT

01

MEANING AND IMPORTANCE OF FEATURE SCALING

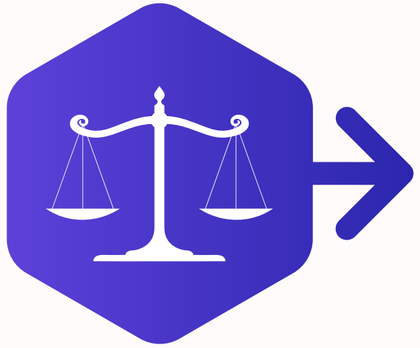
WHAT IS FEATURE SCALING?

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

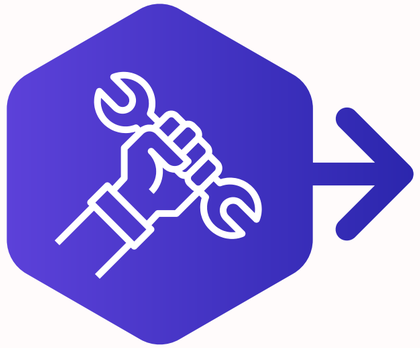
CREDIT: GUPTA_OMG, M. (N.D.). FEATURE ENGINEERING: SCALING, NORMALIZATION, AND STANDARDIZATION. RETRIEVED FROM GEEKSFORGEEKS: HTTPS://WWW.GEEKSFORGEEKS.ORG/ML-FEATURE-SCALING-PART-2/

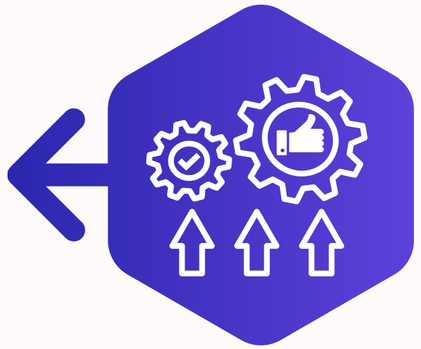
IMPORTANCE OF FEATURE SCALING

ALL FEATURES ARE ON A COMPARABLE SCALE



PREVENTING NUMERICAL INSTABILITY





ALGORITHM PERFORMANCE IMPROVEMENT

REMOVING BIAS

02

FEATURE SCALING TECHNIQUES

FEATURE SCALING TECHNIQUES

IN THIS SUDY, WE WILL FOCUS ON 6 TECHNIQUES

- Standardization (Z-score Scaling)
- Min-Max Scaling
- Max Absolute Scaling
- Robust Scaling
- Mean Normalization
- Unit Vector Scaling

STANDARDIZATION (Z-SCORE SCALING)

Standardization, also known as Z-score normalization, is a method for rescaling values by subtracting the mean and dividing by the standard deviation. This process transforms features to have a mean of 0 and a standard deviation of 1

$$x' = \frac{x - \bar{x}}{\sigma}$$

CREDIT: FEATURE SCALING. (2024, FEBRUARY 16). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/FEATURE_SCALING

MIN-MAX SCALING

Min-Max Scaling, also known as normalization, rescales features to a fixed range, typically between 0 and 1. This technique subtracts the minimum value from each feature and then divides it by the range (maximum value minus minimum value).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

CREDIT: FEATURE SCALING. (2024, FEBRUARY 16). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/FEATURE_SCALING

MAX ABSOLUTE SCALING

Max Absolute Scaling scales features by dividing each feature by its maximum absolute value. This technique preserves the sign of the data while scaling it to the range [-1, 1]. Max Absolute Scaling is useful when the presence of outliers is expected and when preserving the sign of the data is important.

$$x_{scaled} = \frac{x}{max(|x|)}$$

CREDIT: VASHISHT, R. (2021, JANUARY 06). WHEN TO PERFORM A FEATURE SCALING? RETRIEVED FROM ATOTI: HTTPS://WWW.ATOTI.IO/ARTICLES/WHEN-TO-PERFORM-A-FEATURE-SCALING/

ROBUST SCALING

Robust Scaling, as the name suggests, is robust to outliers and extreme values in the dataset. Instead of using the mean and standard deviation like standardization, robust scaling employs the median and the interquartile range (IQR). It subtracts the median from each feature and divides by the IQR.

$$x_{rs} = \frac{x_i - Q_2(x)}{Q_3(x) - Q_1(x)}$$

CREDIT: FEATURE SCALING. (2024, FEBRUARY 16). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/FEATURE_SCALING

MEAN NORMALIZATION

Mean Normalization centers the data around zero by subtracting the mean of each feature from its values. This technique does not scale the range of features to a specific interval but ensures that the mean of each feature is zero.

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

CREDIT: FEATURE SCALING. (2024, FEBRUARY 16). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/FEATURE_SCALING

UNIT VECTOR SCALING

Unit Vector Scaling, also known as vector normalization, scales each feature vector to have a unit norm. This technique is particularly useful when the magnitude of features is important, but their direction is irrelevant. By scaling each feature vector to have a length of 1,

$$x' = \frac{x}{\|x\|}$$

$\|x\|$ Is the Euclidean length of the Feature Vector.

CREDIT: VASHISHT, R. (2021, JANUARY 06). WHEN TO PERFORM A FEATURE SCALING? RETRIEVED FROM ATOTI: HTTPS://WWW.ATOTI.IO/ARTICLES/WHEN-TO-PERFORM-A-FEATURE-SCALING/

03

METHODOLOGY OUTLINE

DATASET

DAVID LAPP · UPDATED 5 YEARS AGO

959 | New Notebook | Download (6 kB)

Heart Disease Dataset

Public Health Dataset

CREDIT: LAPP, D. (2019). HEART DISEASE DATASET. RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/DATASETS/JOHNSMITH88/HEART-DISEASE-DATASET/DATA?SELECT=HEART.CSV

ATTRIBUTE INFORMATION

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

CREDIT: LAPP, D. (2019). HEART DISEASE DATASET. RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/DATASETS/JOHNSMITH88/HEART-DISEASE-DATASET/DATA?SELECT=HEART.CSV

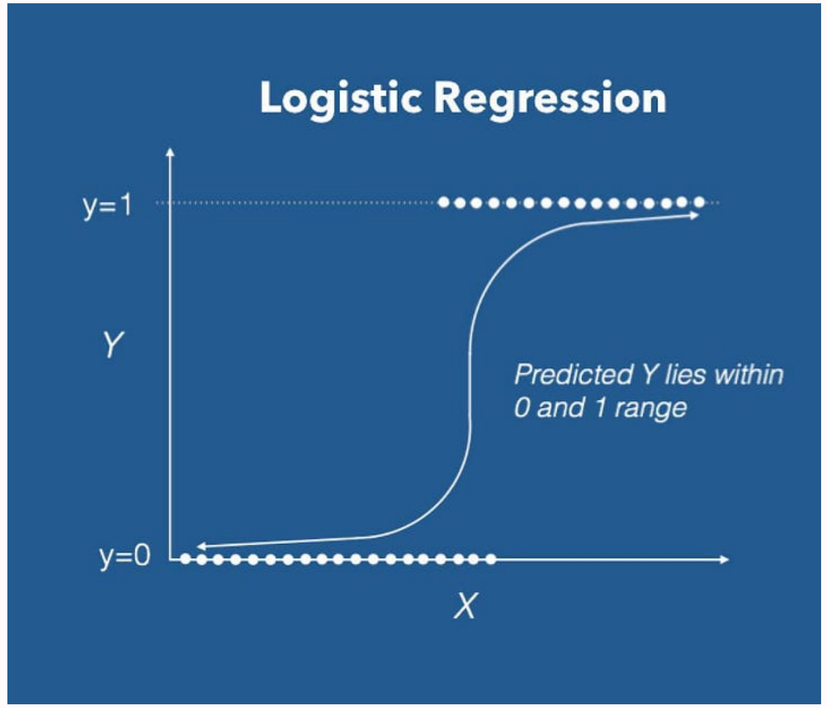
DATASET BEFORE PRE-PROCESSED

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

CREDIT: LAPP, D. (2019). HEART DISEASE DATASET. RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/DATASETS/JOHNSMITH88/HEART-DISEASE-DATASET/DATA?SELECT=HEART.CSV

LOGISTIC REGRESSION

Logistic regression is a statistical method used in regression analysis to estimate the parameters of a logistic model for binary classification tasks. It involves predicting the probability of an event occurring based on a given set of independent variables. The logistic function, with its distinctive S-shaped curve, transforms a linear combination of input features into a probability value between 0 and 1, making it suitable for binary classification tasks like spam email detection or disease diagnosis.



CREDIT: HTTPS://TOWARDSDATASCIENCE.COM/INTRODUCTION-TO-LOGISTIC-REGRESSION-66248243C148

METHODOLOGY

1. Dataset Exploration and Preparation
2. Feature Scaling
3. Model Training
4. Model Evaluation

EVALUATION METRICS

To assess the effectiveness of each feature scaling technique, we employ logistic regression as our chosen classification algorithm. Logistic regression is well-suited for binary classification tasks and provides insights into the predictive performance of our models. Specifically, we utilize the ‘classification_report’ function to generate a comprehensive summary of key evaluation metrics, including precision, recall, F1-score, and support, for each class in the dataset.

04

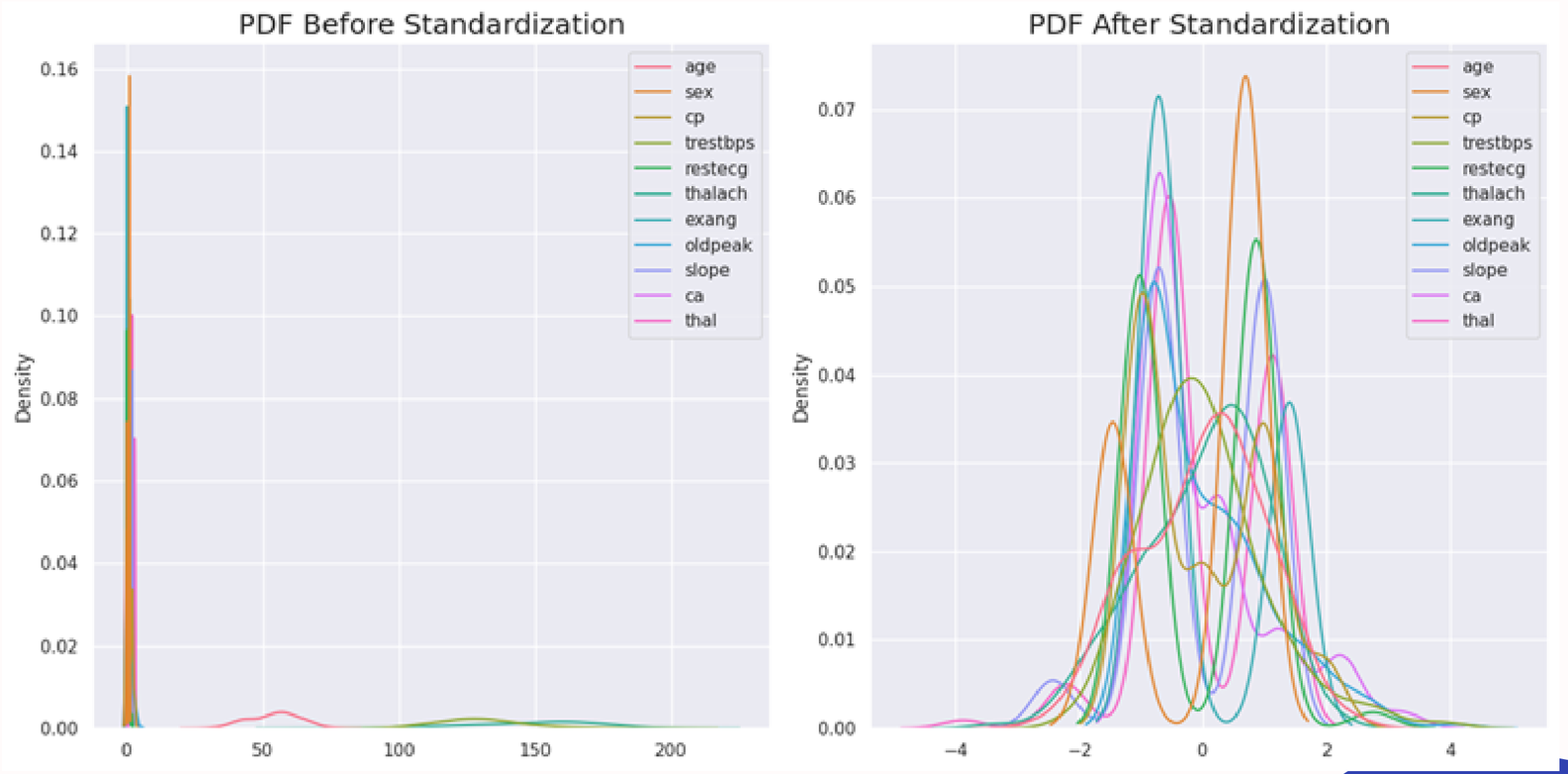
PRACTICAL IMPLEMENTATION

05

INTERPRETATION AND DISCUSSION

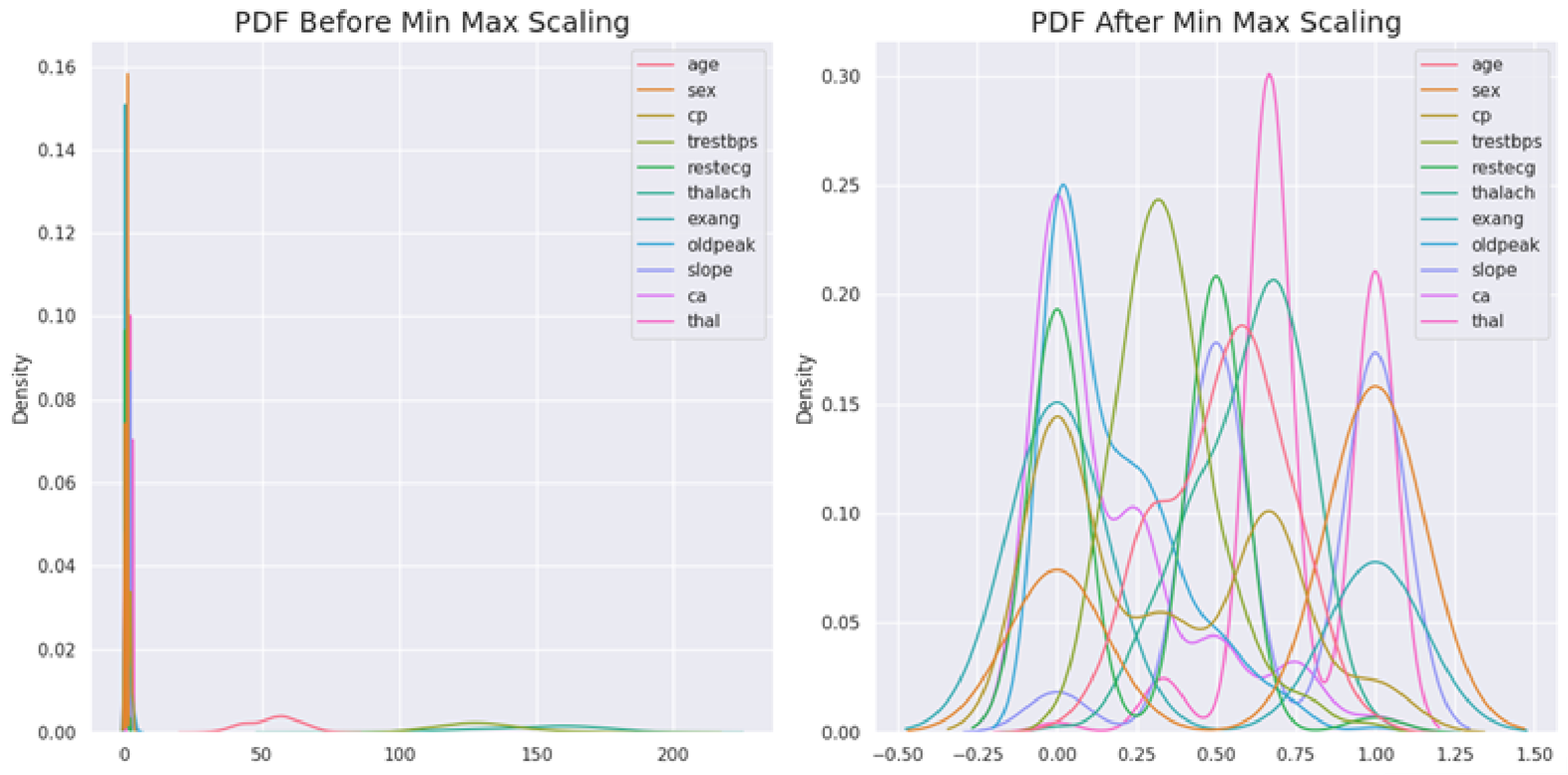
COMPARATIVE ANALYSIS

STANDARDIZATION



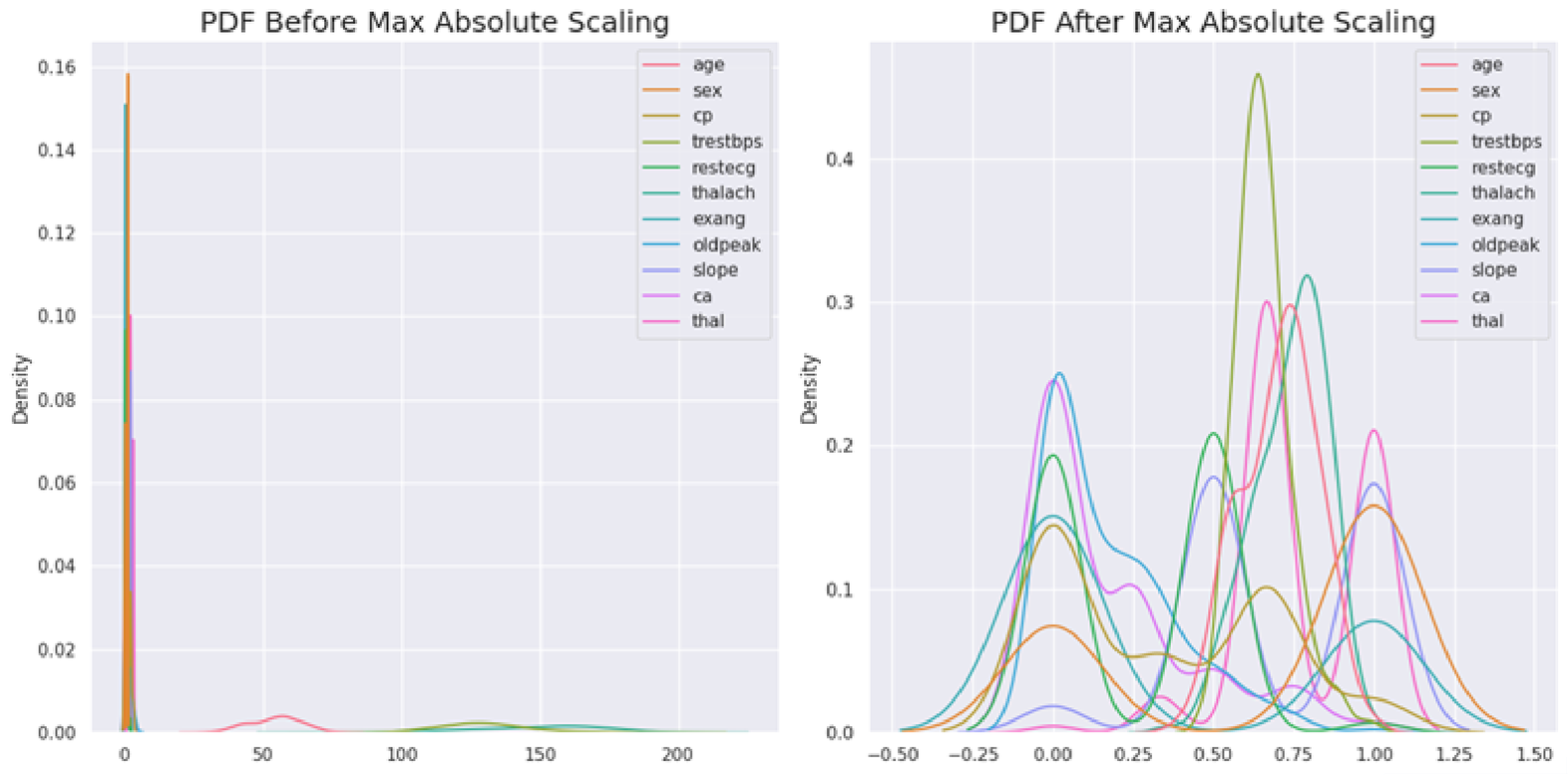
COMPARATIVE ANALYSIS

MIN-MAX SCALING



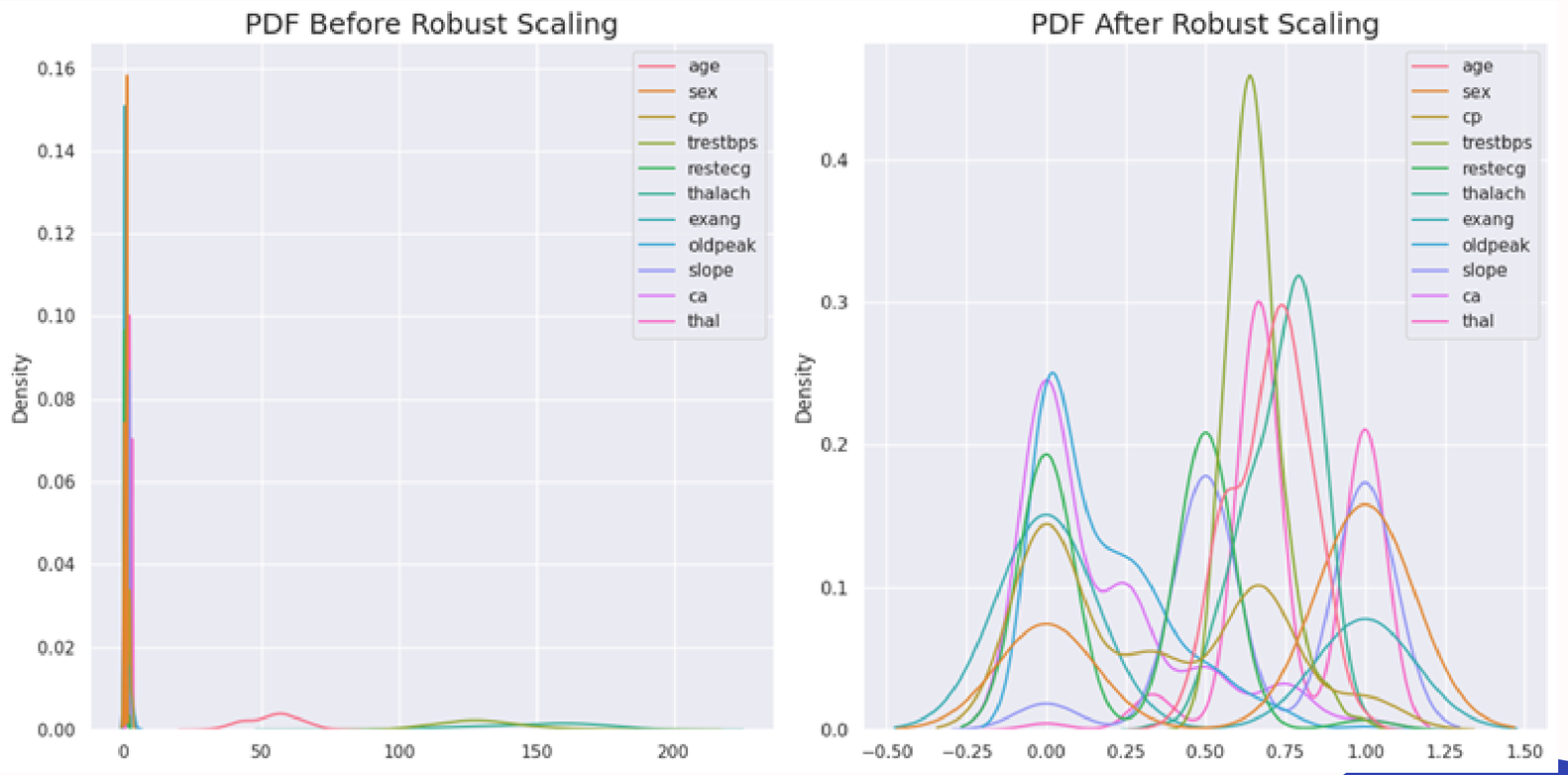
COMPARATIVE ANALYSIS

MAX ABSOLUTE SCALING



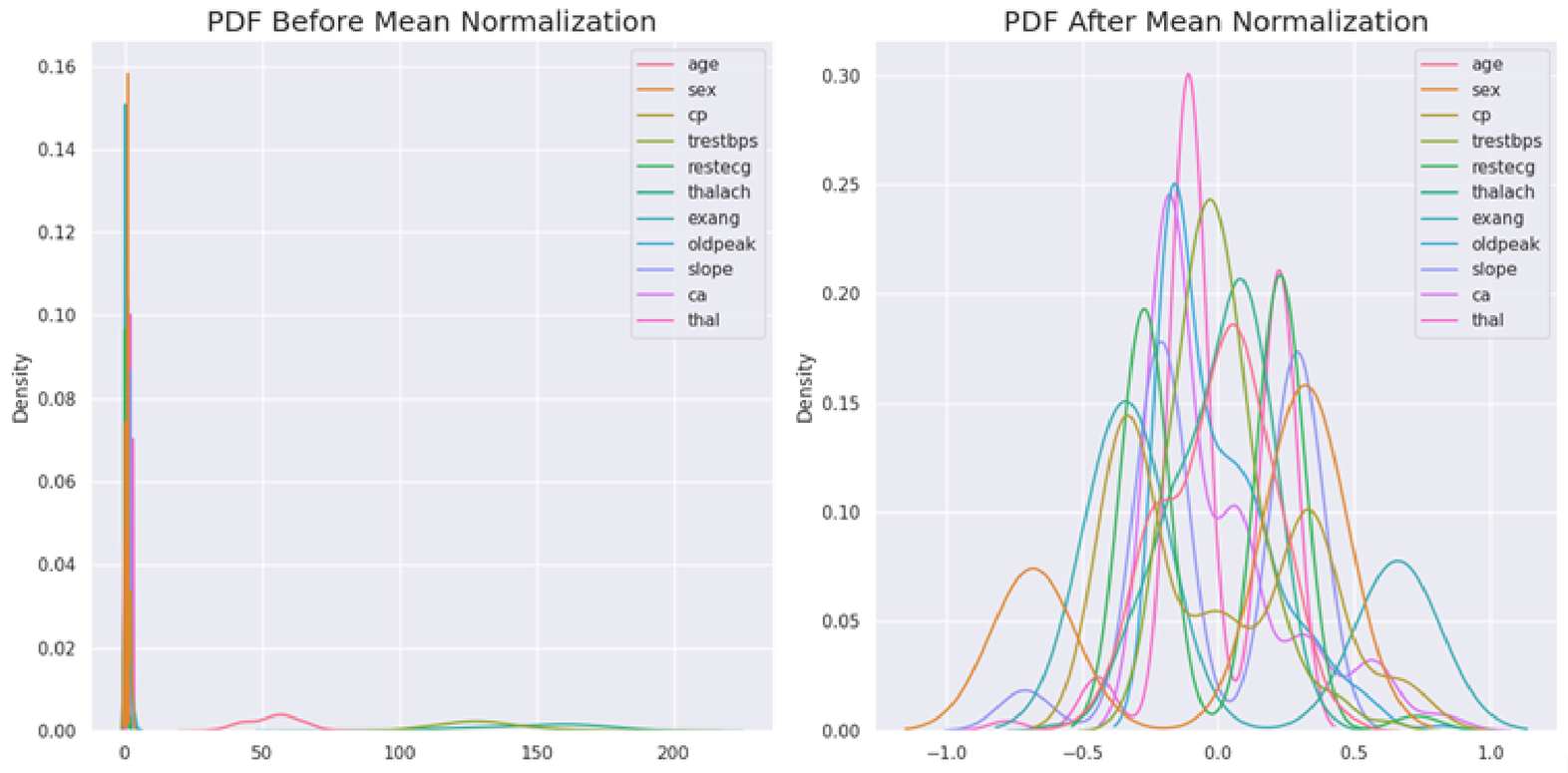
COMPARATIVE ANALYSIS

ROBUST SCALING



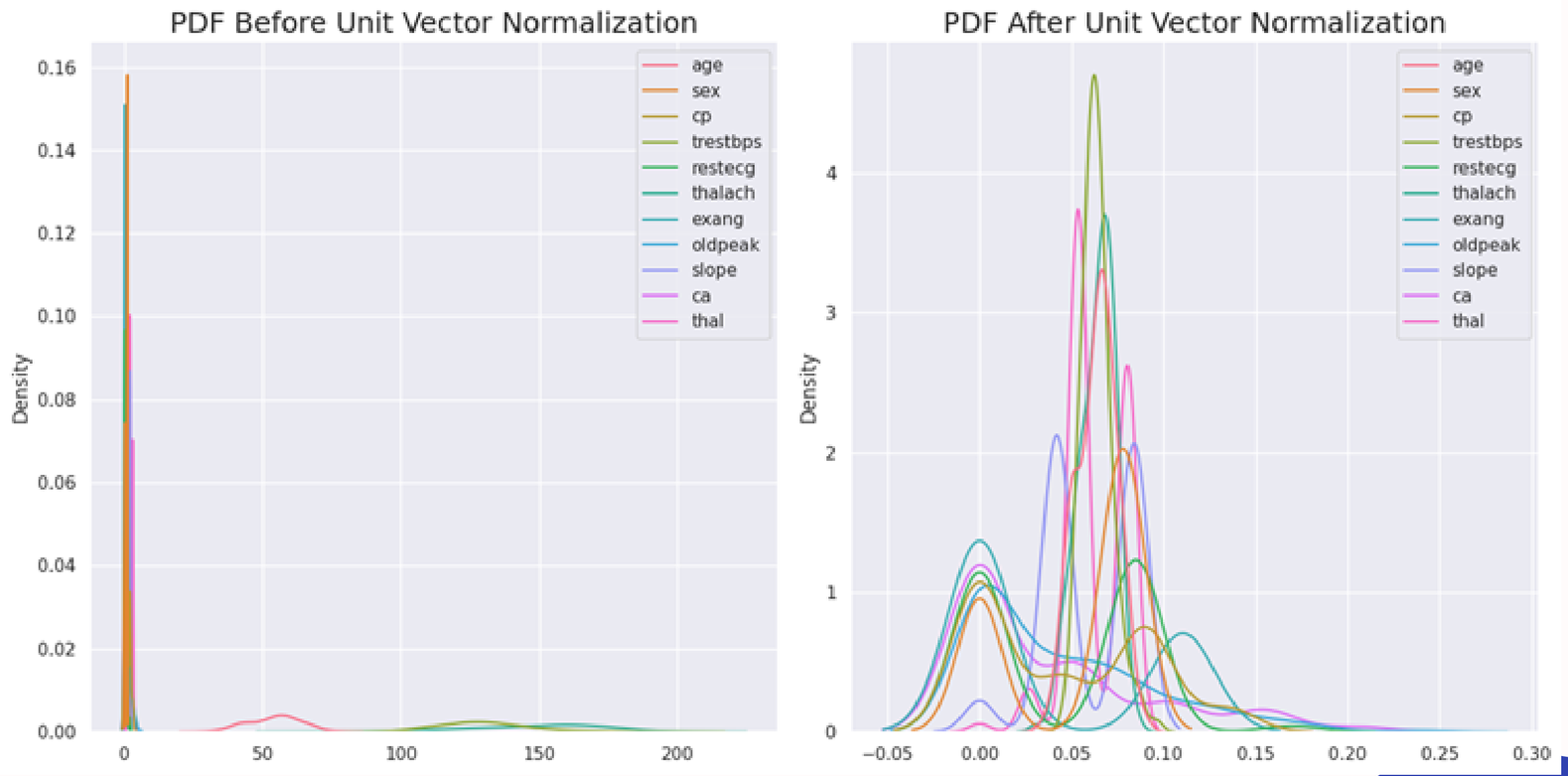
COMPARATIVE ANALYSIS

MEAN NORMALIZATION



COMPARATIVE ANALYSIS

UNIT VECTOR NORMALIZATION



CLASSIFICATION PERFORMANCE

Techniques	Accuracy	Precision	Recall	F1-Score
Without Feature Scaling	0.80	0.82	0.81	0.80
Standardization	0.77	0.81	0.78	0.77
Min-Max Scaling	0.75	0.78	0.76	0.75
Max Absolute Scaling	0.77	0.78	0.77	0.77
Robust Scaling	0.77	0.78	0.77	0.77
Mean Normalization	0.75	0.78	0.76	0.75
Unit Vector Normalization	0.80	0.80	0.80	0.80

CLASSIFICATION PERFORMANCE

Techniques	Accuracy	Precision	Recall	F1-Score
Without Feature Scaling	0.80	0.82	0.81	0.80
Standardization	0.77	0.81	0.78	0.77
Min-Max Scaling	0.75	0.78	0.76	0.75
Max Absolute Scaling	0.77	0.78	0.77	0.77
Robust Scaling	0.77	0.78	0.77	0.77
Mean Normalization	0.75	0.78	0.76	0.75
Unit Vector Normalization	0.80	0.80	0.80	0.80

CLASSIFICATION PERFORMANCE

Techniques	Accuracy	Precision	Recall	F1-Score
Without Feature Scaling	0.80	0.82	0.81	0.80
Standardization	0.77	0.81	0.78	0.77
Min-Max Scaling	0.75	0.78	0.76	0.75
Max Absolute Scaling	0.77	0.78	0.77	0.77
Robust Scaling	0.77	0.78	0.77	0.77
Mean Normalization	0.75	0.78	0.76	0.75
Unit Vector Normalization	0.80	0.80	0.80	0.80

DISCUSSION

FROM THE CLASSIFICATION PERFORMANCE, STANDARDIZATION, MIN-MAX SCALING, AND MEAN NORMALIZATION DO NOT SIGNIFICANTLY IMPROVE MODEL PERFORMANCE COMPARED TO THE UNSCALED MODEL. MAX ABSOLUTE SCALING AND ROBUST SCALING MAINTAIN PERFORMANCE SIMILAR TO THE UNSCALED MODEL, MAKING THEM SUITABLE ALTERNATIVES FOR PRESERVING MODEL ACCURACY. UNIT VECTOR NORMALIZATION PROVIDES CONSISTENT PERFORMANCE WITH THE UNSCALED MODEL, INDICATING ITS EFFECTIVENESS IN MAINTAINING MODEL STABILITY.

COMPARE ADVANTAGE & DISADVANTAGE

Techniques	Advantages	Disadvantages
Without Feature Scaling	The model achieved relatively high precision and recall for both classes without feature scaling, indicating that the original feature values were already conducive to model performance.	The model's performance could potentially be further improved with feature scaling, especially in scenarios where features have different scales or distributions.
Standardization	Standardization ensures that features have a mean of 0 and a standard deviation of 1, making it robust to outliers and suitable for algorithms assuming Gaussian distributions.	In this study, standardization led to a decrease in precision for class 0, suggesting that it may not be the optimal scaling technique for this specific dataset and model.

COMPARE ADVANTAGE & DISADVANTAGE

Techniques	Advantages	Disadvantages
Min-Max Scaling	Min-Max Scaling bounds feature values between 0 and 1, preserving relationships between data points and making it suitable for algorithms relying on distance measures.	Although Min-Max Scaling improved recall for class 1, it led to a decrease in precision for both classes, indicating a trade-off between precision and recall.
Max Absolute Scaling	Max Absolute Scaling preserves the sign of the data while scaling it, making it useful when the sign is significant.	Similar to Min-Max Scaling, Max Absolute Scaling resulted in a decrease in precision for both classes, indicating potential trade-offs between precision and recall.

COMPARE ADVANTAGE & DISADVANTAGE

Techniques	Advantages	Disadvantages
Robust Scaling	Robust Scaling utilizes robust statistics such as the median and interquartile range, making it resilient to outliers and suitable for datasets with skewed distributions.	While Robust Scaling maintained precision and recall for class 1, it led to a decrease in precision for class 0, suggesting potential limitations in preserving class boundaries.
Mean Normalization	Mean Normalization centers data around zero, preserving relative distances between data points and ensuring equal contribution from each feature.	Similar to other scaling techniques, Mean Normalization resulted in a decrease in precision for both classes, indicating potential trade-offs between precision and recall.
Unit Vector Scaling	Unit Vector Normalization scales each feature vector to have a unit norm, ensuring equal contribution from each feature regardless of magnitude.	While Unit Vector Normalization maintained precision and recall for both classes, it did not significantly improve model performance compared to other techniques.

LIMITATIONS

- THIS ANALYSIS ONLY CONSIDERED A SINGLE DATASET AND MAY NOT GENERALIZE TO ALL CLASSIFICATION PROBLEMS.
- OTHER FACTORS BEYOND SCALING, SUCH AS HYPERPARAMETER TUNING OR FEATURE SELECTION, COULD ALSO INFLUENCE THE MODEL'S PERFORMANCE.

FUTURE WORK

Future study could explore additional feature scaling techniques and evaluate their performance on different datasets. Additionally, the impact of feature scaling on various machine learning algorithms and tasks could be investigated to further refine best practices in data preprocessing and model training.

06

CONCLUSION

CONCLUSION

Feature scaling is a critical preprocessing step in machine learning that improves model convergence, performance, stability, and interpretability. Understanding and appropriately applying feature scaling techniques can significantly impact the success of machine learning models across various domains.

In summary, this report provides an in-depth analysis of feature scaling, its importance, common techniques, practical implementation, and recommendations for best practices in machine learning workflows.

REFERENCE

- DEY, A. (2021). COMPLETE GUIDE TO FEATURE SCALING. RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/CODE/AIMACK/COMPLETE-GUIDE-TO-FEATURE-SCALING
- FEATURE SCALING. (2024, FEBRUARY 16). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/FEATURE_SCALING
- GUPTA_OMG, M. (N.D.). FEATURE ENGINEERING: SCALING, NORMALIZATION, AND STANDARDIZATION. RETRIEVED FROM GEEKSFORGEEKS: HTTPS://WWW.GEEKSFORGEEKS.ORG/ML-FEATURE-SCALING-PART-2/
- LAPP, D. (2019). HEART DISEASE DATASET. RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/DATASETS/JOHNSMITH88/HEART-DISEASE-DATASET/DATA?SELECT=HEART.CSV
- LOGISTIC REGRESSION. (N.D.). RETRIEVED FROM WIKIPEDIA: HTTPS://EN.WIKIPEDIA.ORG/WIKI/LOGISTIC_REGRESSION
- PREPROCESSING DATA. (N.D.). RETRIEVED FROM SCIKIT-LEARN: HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/PREPROCESSING.HTML
- VASHISHT, R. (2021, JANUARY 06). WHEN TO PERFORM A FEATURE SCALING? RETRIEVED FROM ATOTI: HTTPS://WWW.ATOTI.IO/ARTICLES/WHEN-TO-PERFORM-A-FEATURE-SCALING/
- YADAV, P. (2023, NOVEMBER 9). DIFFERENCE BETWEEN STANDARDIZATION AND NORMALIZATION. RETRIEVED FROM MEDIUM: HTTPS://MEDIUM.COM/@PAWAN329/DIFFERENCE-BETWEEN-STANDARDIZATION-AND-NORMALIZATION-52C2B2313193


THANK YOU
FOR WATCHING