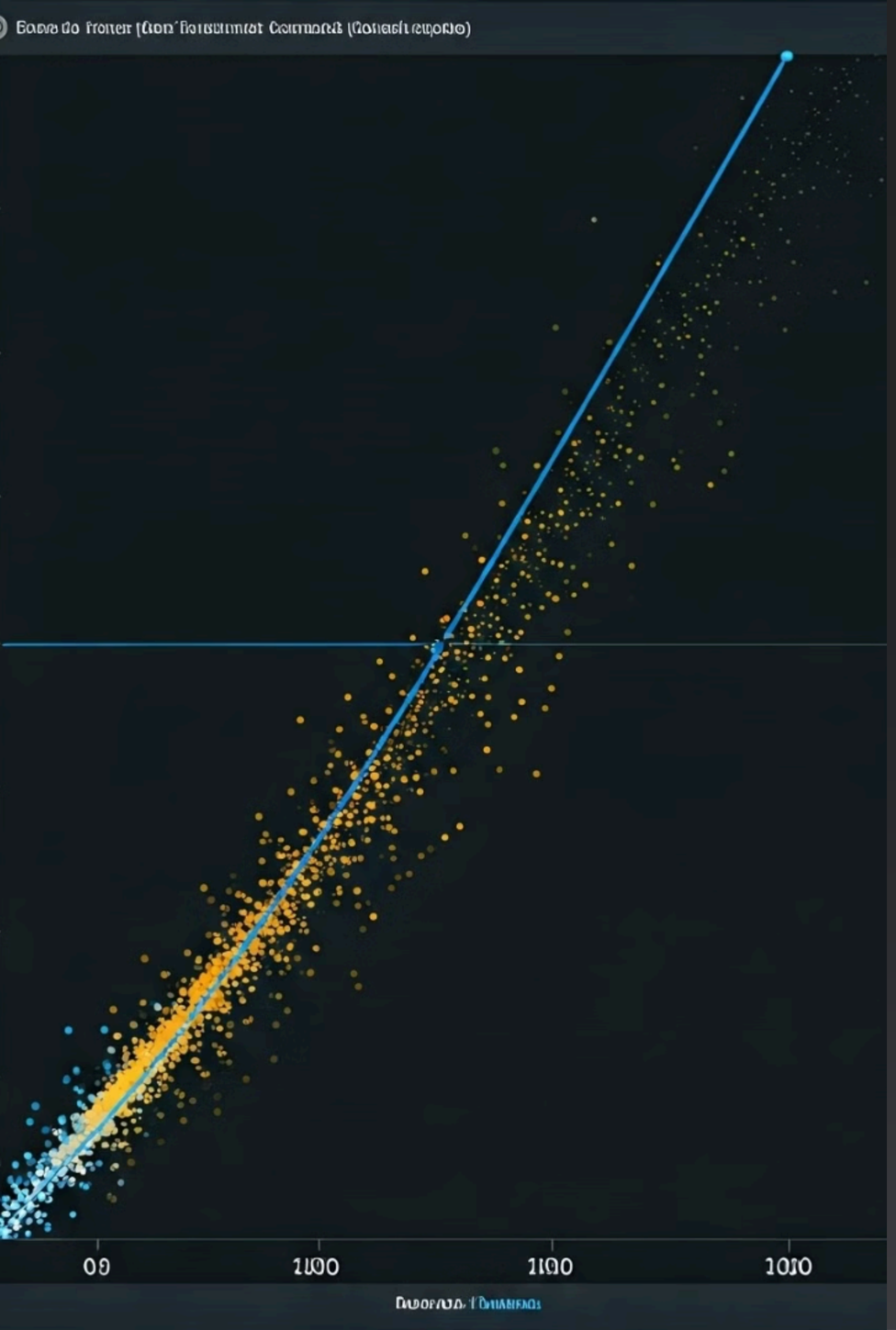
# K-Nearest Neighbors (KNN) Regression

K-Nearest Neighbors (KNN) Regression is a non-parametric supervised learning algorithm used for predicting continuous target variables. It operates based on the principle of similarity, predicting the value of a new data point based on the values of its nearest neighbors in the feature space.

**Manoj Kumar Sahoo**



# How does KNN perform regression?

1

## Step 1: Calculate distances

The algorithm computes the distances between the new data point and all existing data points in the training set.

2

## Step 2: Identify nearest neighbors

It identifies the 'k' nearest neighbors based on their calculated distances, where 'k' is a user-defined parameter.

3

## Step 3: Predict target value

The predicted value for the new data point is the average of the target values of its 'k' nearest neighbors.

# Role of the Parameter 'K' in KNN Regression

## Lower 'k'

A lower 'k' value leads to a more complex decision boundary, potentially prone to overfitting and capturing noise in the data.

## Higher 'k'

A higher 'k' value results in a smoother decision boundary, potentially underfitting and missing important patterns in the data.

## Finding the Optimal 'k'

Choosing an optimal 'k' value involves balancing the trade-off between bias and variance, typically using cross-validation techniques.

# Distance Metrics in KNN Regression

## Euclidean Distance

The most commonly used metric, calculates the straight-line distance between two points.

## Manhattan Distance

Calculates the distance by summing the absolute differences of coordinates along each dimension.

## Minkowski Distance

A generalization of Euclidean and Manhattan distances, allowing for different powers (p) to adjust the weighting of individual dimensions.

# Weighting Neighbors in KNN Regression

1 **Uniform Weighting**

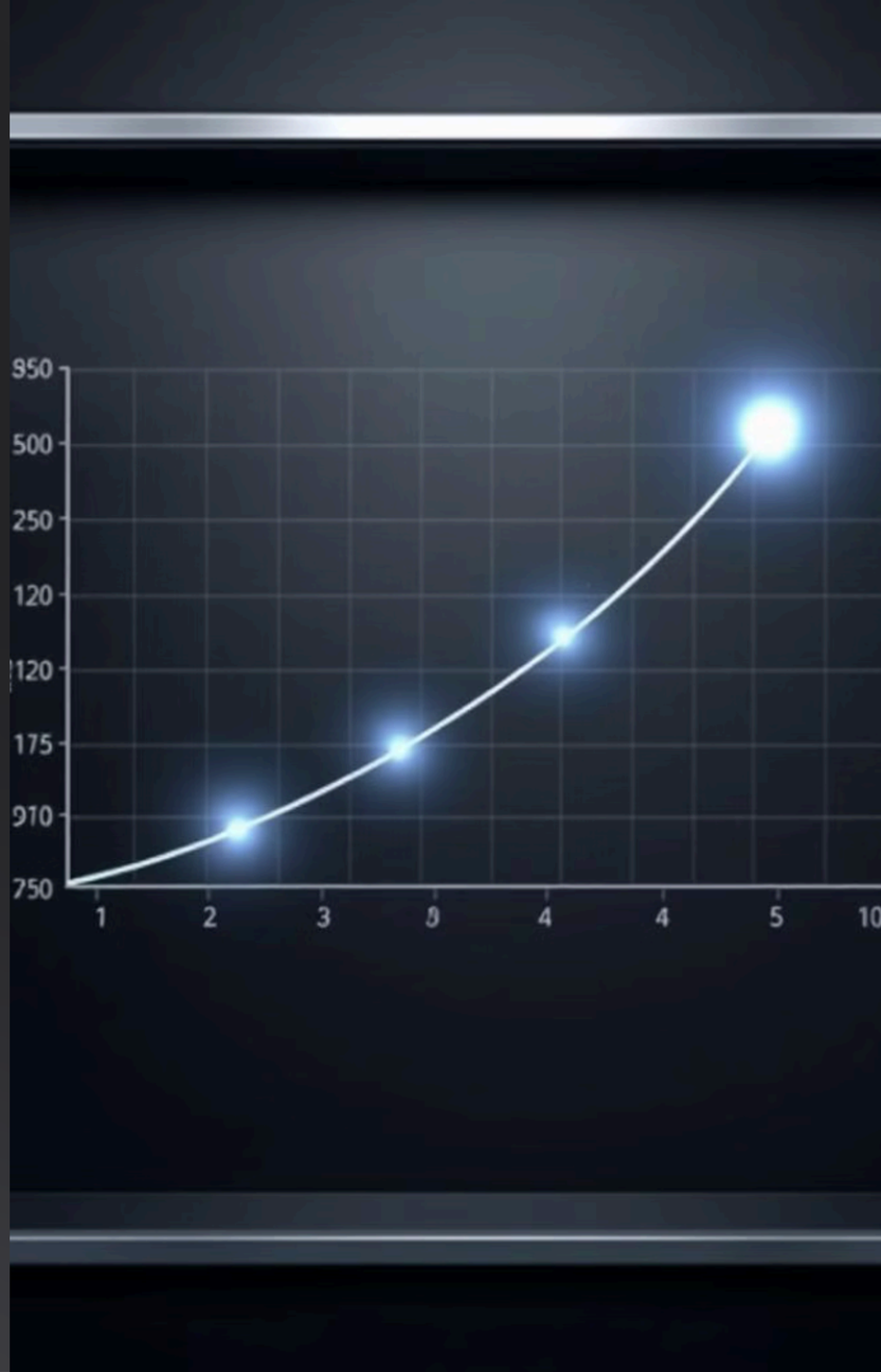
Each neighbor contributes equally to the prediction, regardless of its distance to the new data point.

2 **Distance-Weighted**

Neighbors closer to the new data point contribute more to the prediction than those further away.

3 **Inverse Distance Weighting**

Weighting is inversely proportional to the distance, meaning closer neighbors have a higher influence.



# Impact of Feature Scaling on KNN Regression

1

## Unscaled Features

Features with vastly different scales can distort distance calculations, leading to inaccurate predictions.

2

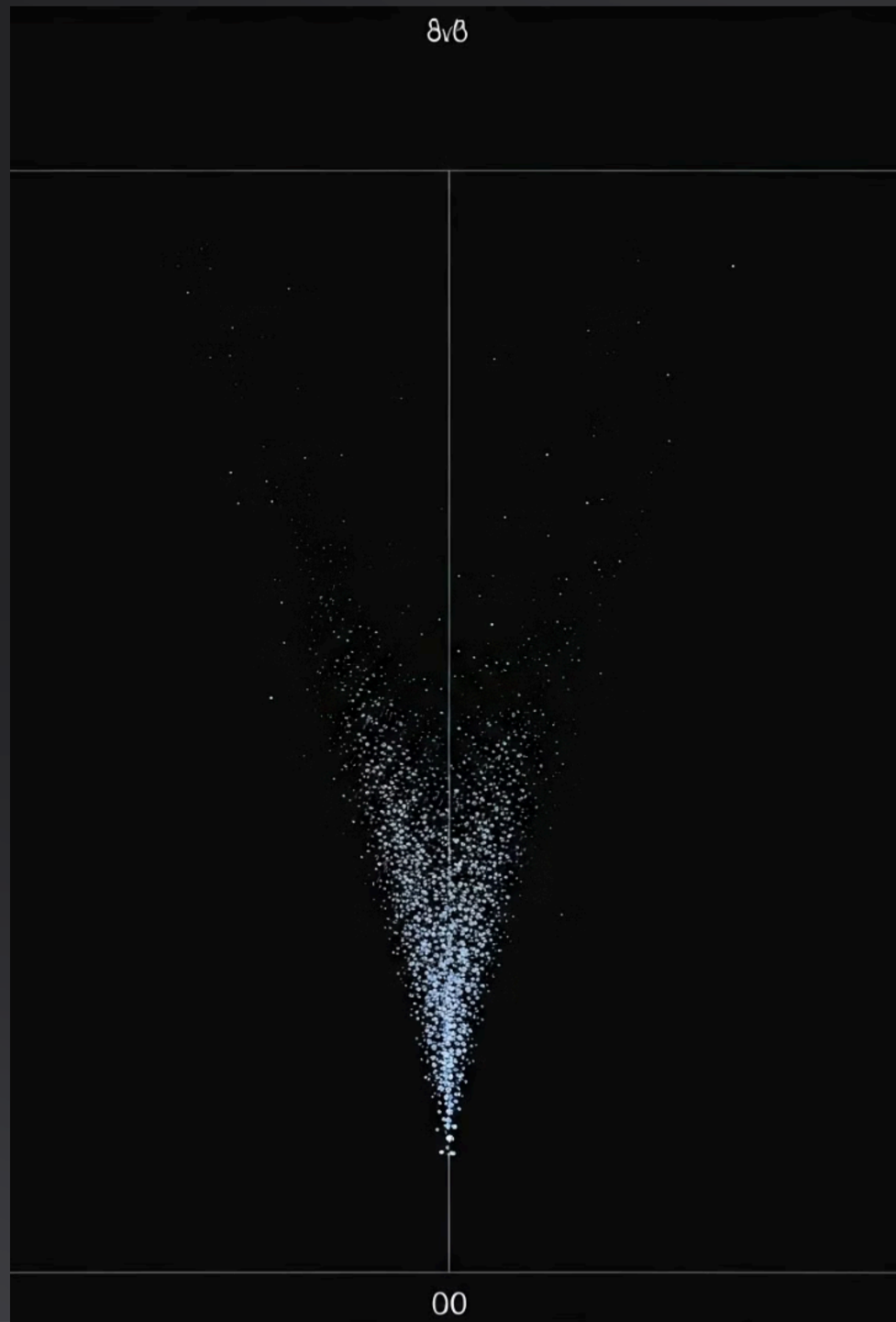
## Feature Scaling

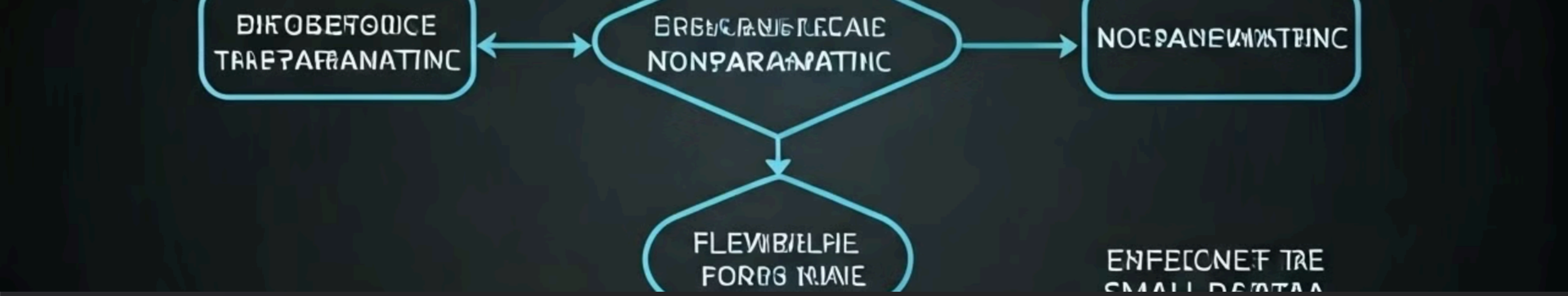
Scaling features to a common range ensures that all features contribute equally to the distance calculations.

3

## Improved Accuracy

By standardizing the scale, KNN can accurately assess the proximity of data points, resulting in improved model performance.





# Advantages of KNN Regression

## Ease of Implementation

KNN is relatively straightforward to implement and understand, making it accessible for beginners.

## Flexibility

KNN can handle complex relationships between features and target variables without making strong assumptions.

## Non-parametric Nature

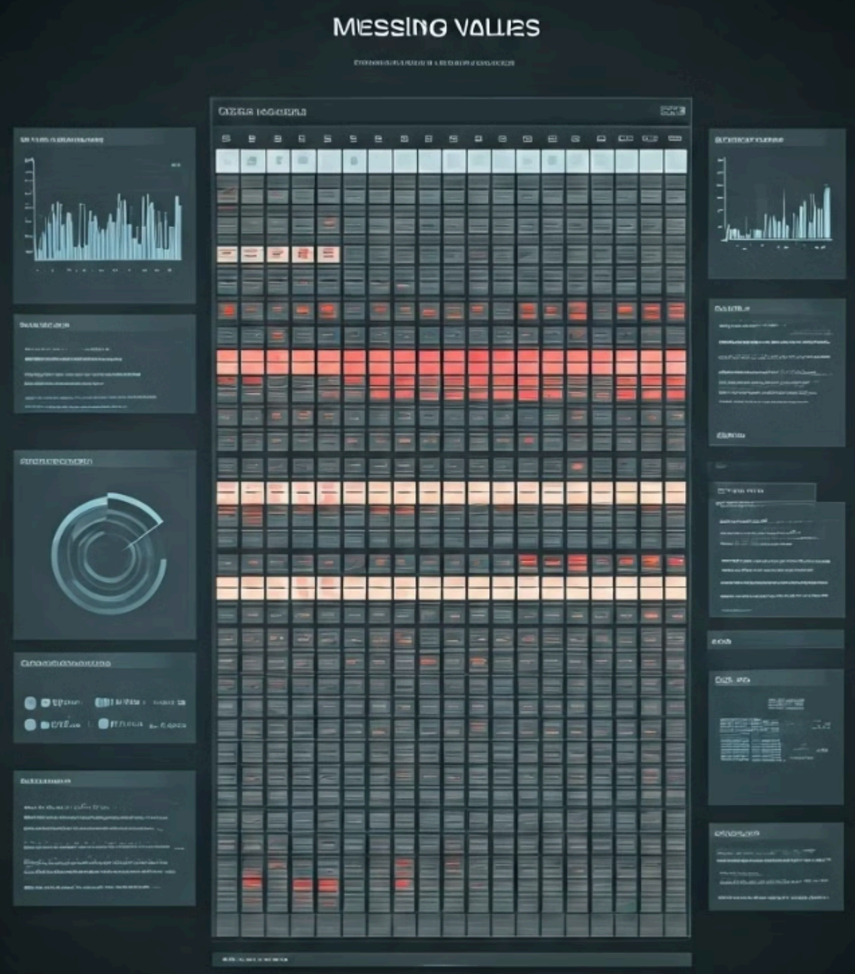
KNN does not require specifying a parametric model, making it adaptable to different data distributions.

## Efficiency for Small Datasets

KNN is particularly efficient for small datasets, as it does not require extensive training time.

# Disadvantages of KNN Regression

| | |
|---|---|
| High computational cost for large datasets | Sensitive to outliers and noisy data |
| Curse of dimensionality, performance degrades with increasing feature dimensions | Difficult to select the optimal 'k' value |




# Handling Missing Data in KNN Regression

1 **Imputation**

Replacing missing values with estimated values based on available data.

2 **Deletion**

Removing data points with missing values, potentially leading to data loss.

3 **Nearest Neighbor Imputation**

Imputing missing values based on the values of the nearest neighbors.

# Conclusion

KNN regression is a valuable tool for predicting continuous variables, particularly suitable for small datasets with complex relationships. While its simplicity and flexibility are advantageous, it is crucial to address its limitations, such as sensitivity to outliers and computational cost for large datasets.