K Nearest Neighbor vs K Means: Where's the difference?

Pakeeza Saleem · Follow

5 min read · Aug 16, 2023

Listen Share

Many people confuse the K nearest neighbor (KNN) with the K means clustering algorithm. However, KNN is a supervised machine learning algorithm used for classification, while K means is an unsupervised machine learning algorithm used with the intention of clustering data points.

Sounds too complicated? Let us break down both algorithms into easy terms and note the differences.

Photo by Andy Kelly on Unsplash

Table of Contents

· What is K nearest Neighbor (KNN)?
◦ KNN Algorithm Visualization
◦ Working of KNN Algorithm
· What is K means clustering?
◦ K means Algorithm Visualization
◦ Working of K means Algorithm
· K means clustering vs K nearest neighbor: What is the difference?
· The Take Away
· FAQs

What is K nearest Neighbor (KNN)?

> *As per the name, in KNN, we are looking for a specific number of points (k) that are the nearest neighbors to a particular point in terms of distance. It is a supervised learning algorithm, meaning that all the "neighbors" are* ***labeled*** *inputs.*

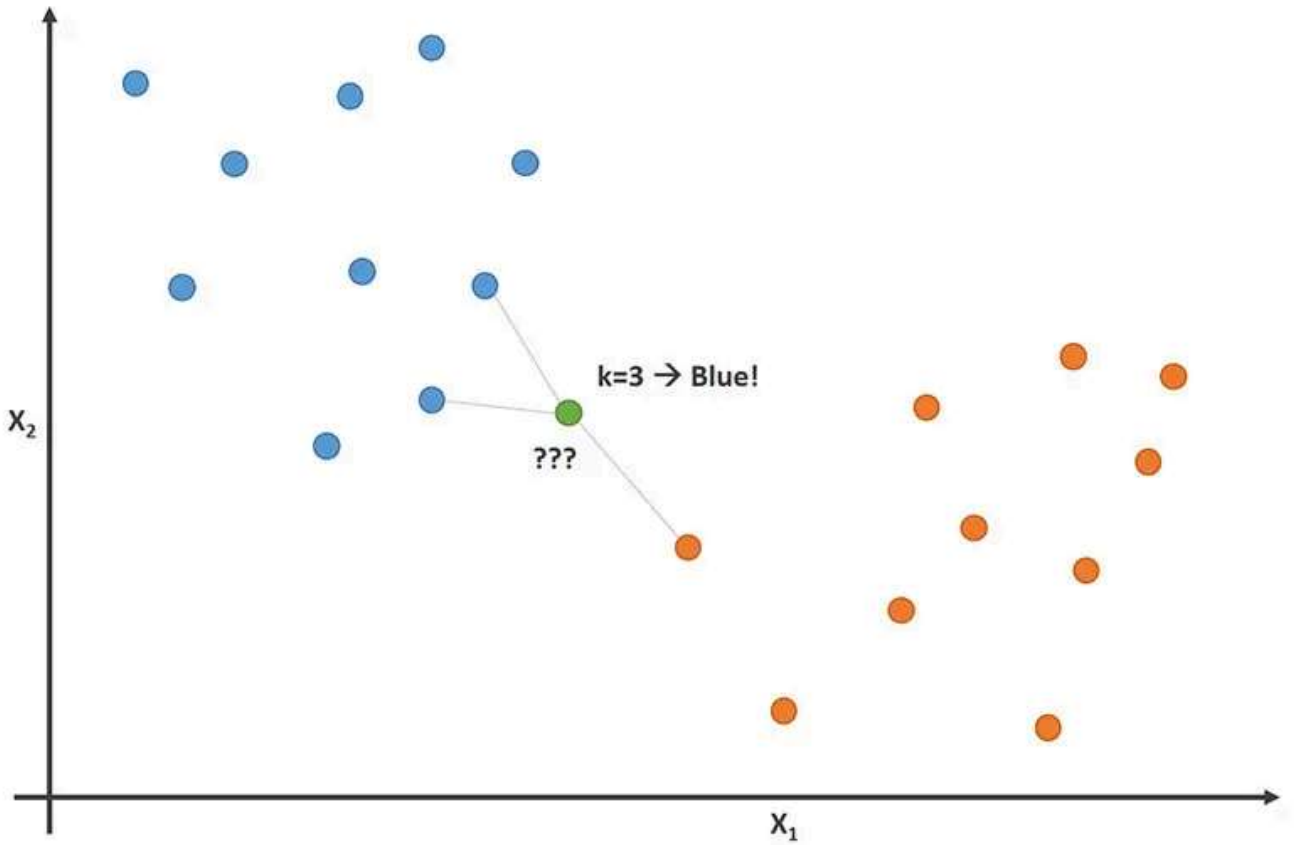
KNN Algorithm Visualization

Imagine you have a bunch of points, each belonging to a specific class. To keep it simple, let's say there are two classes, A and B, so each point belongs to either class A or B. All the points have specific feature values.

For example,

If class A is the class of coffee mugs and B is the class of tea cups, then they will have some distinguishable features. For example, height, weight, etc. Now, another drinking utensil has specific features and can only belong to either of the two classes.

But how will we know which class it belongs to? This is where KNN comes in!



K nearest neighbor graph visualization

Working of KNN Algorithm

Back to our point analogy, we can carry out the following steps to find out whether the new point belongs to class A or B.

1. Plot all the points on a graph where axes are the features. In our example, the axes will be height and weight, and all the points will be plotted based on their feature values.
2. Find the Euclidean distance of all points in classes A and B with the new point.
3. Now, combine all the distances in one array and sort all the distances in ascending order.
4. Select any value of k, say 3, and select the top 3 values from the distance array.
5. Analyze each value's category and decide based on majority votes. For example, if k is 3, out of the least 3 distance values, if 2 belongs to distance from class B points, then class B is the answer.

So, this is how we perform classification using K nearest neighbor.

Now, let's talk about the K means clustering algorithm. And then, we can see the contrast between the two.

What is K means clustering?

> *As per the name, we are looking for a "k" number of groups called clusters in this algorithm. It is an unsupervised learning algorithm, meaning that none of the points are labeled, and we try to group them to give each group a label.*

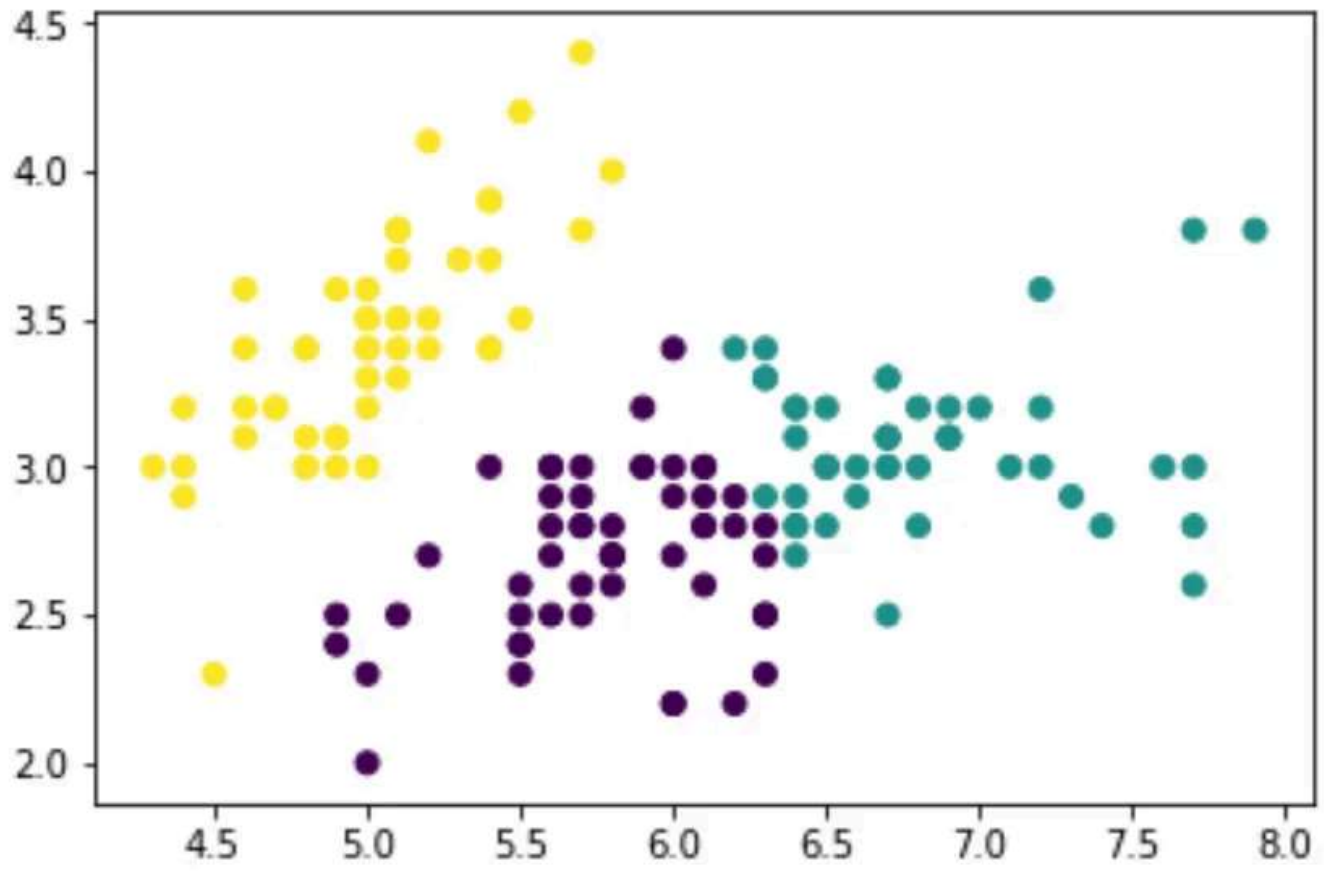
K means Algorithm Visualization

Imagine you have a bunch of points. Let us assume that there are two classes, A and B, so each point belongs to either A or B. All the points have specific feature values. But here, we need to find out which point belongs to which class.

For example,

Suppose some drinking utensils are of different weights and heights. No, we want to classify them into two classes, i.e., coffee mugs and tea cups.

But how will we know which utensil belongs to which class? This is where we need K means algorithm!



K means Clustering Graph Visualization

Working of K means Algorithm

1. Randomly select a value for K. This indicates how many clusters we want to make. For example, if I want to divide the points into 3 classes, we will take k as 3.
2. Pick any k number of points as centroids of the clusters.
3. Calculate the distance of each data point with the centroids and assign the data point to the closest cluster based on distance with the centroid.
4. Now, determine the mean of all the points in every cluster separately; these will be our new centroids.
5. Repeat steps 3 and 4. Iterate until there is a 1 to 2% error for the centroid value of two consecutive iterations.

Once convergence is achieved, the points are divided into k clusters.

So, back to our original question.

K means clustering vs K nearest neighbor: What is the difference?

The difference between k nearest neighbor and k means is obvious.

> **K nearest neighbor** *is a supervised machine learning algorithm. This algorithm is used to classify and determine which class the new point belongs to. It is a lazy learner algorithm, so we don't need a training phase; all the calculations are performed once the test point comes into the picture.*
>
> **K means clustering** *is an unsupervised machine learning algorithm. This algorithm is used to divide the existing data points into classes. It is an eager learner algorithm, so we can't skip the training phase.*

The Take Away

The question for choice between **K nearest neighbor vs. K means** will always be based on the type of input available and the intention for using the algorithm. Hope the article clarifies the difference and makes it easy to choose.

FAQs

What is "k" in k means?

In k means clustering k refers to the number of clusters to be made.

What is "k" in k nearest neighbor?

In k nearest neighbor, k refers to the number of neighbors based on which the decision of classification is to be made.

How do I select the value of k in k nearest neighbor?

K can be any value greater than one, but the value of k is preferred to be an odd number. The reason is to rule out the possibility of a tie between the classes. It is also a common practice to choose k between 3 and 10 based on the available data points.

How do I select the value of k?

K can be any value greater than one, but taking less value of k is preferred as it offers better clustering. It is common to take k anywhere between 2 to 10 based on the size of the dataset.

K Means | K Means Clustering | Knn | K Nearest Neighbours | Machine Learning

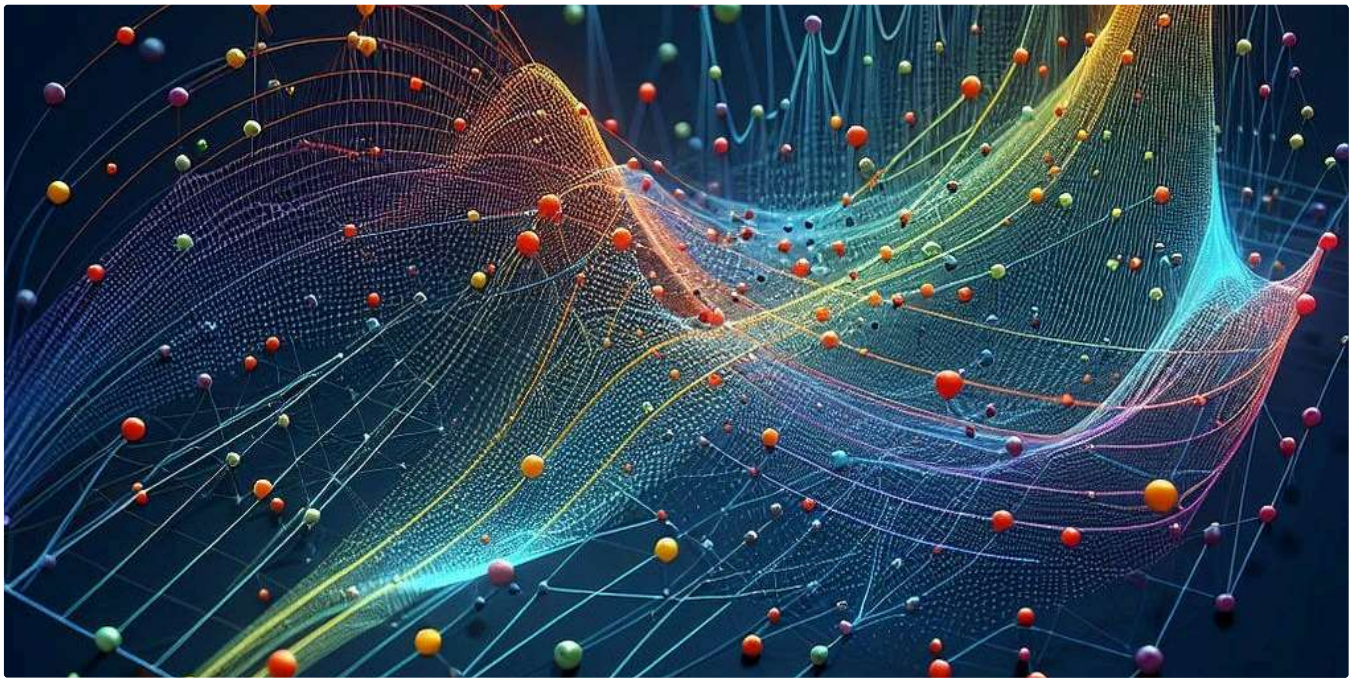
Follow

Written by Pakeeza Saleem

1 Follower

Technical Writer | Content Writer

Recommended from Medium



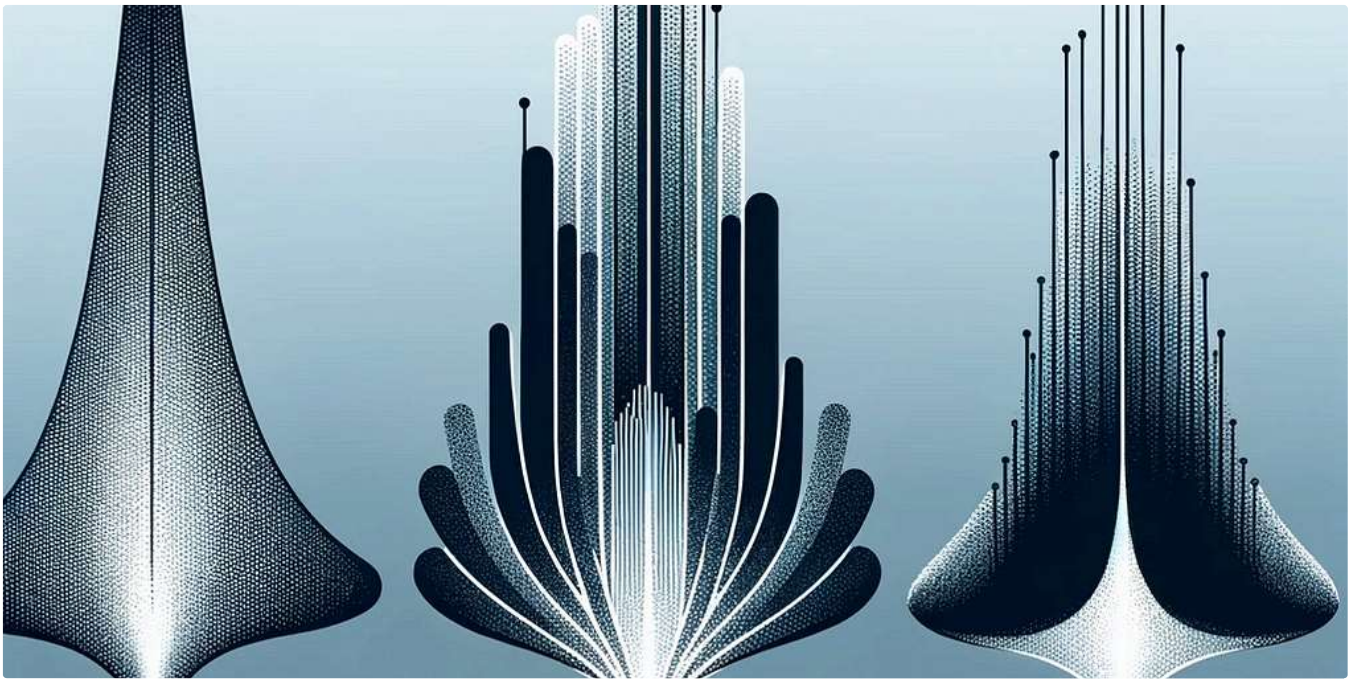
Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two variables just like correlation except possibly...

10 min read · Mar 31, 2024

2.3K 33



Bragadeesh Sundararajan

Balancing the Scale: Transforming Skewed Data into Normal Distributions

When diving into the world of statistics and machine learning, one fundamental concept you'll encounter is the idea of data distributions...

· 11 min read · Jan 2, 2024

Rohan Kumar Bohara

How to find the optimal value of K in KNN?

Introduction

7 min read · Oct 29, 2023

14 1

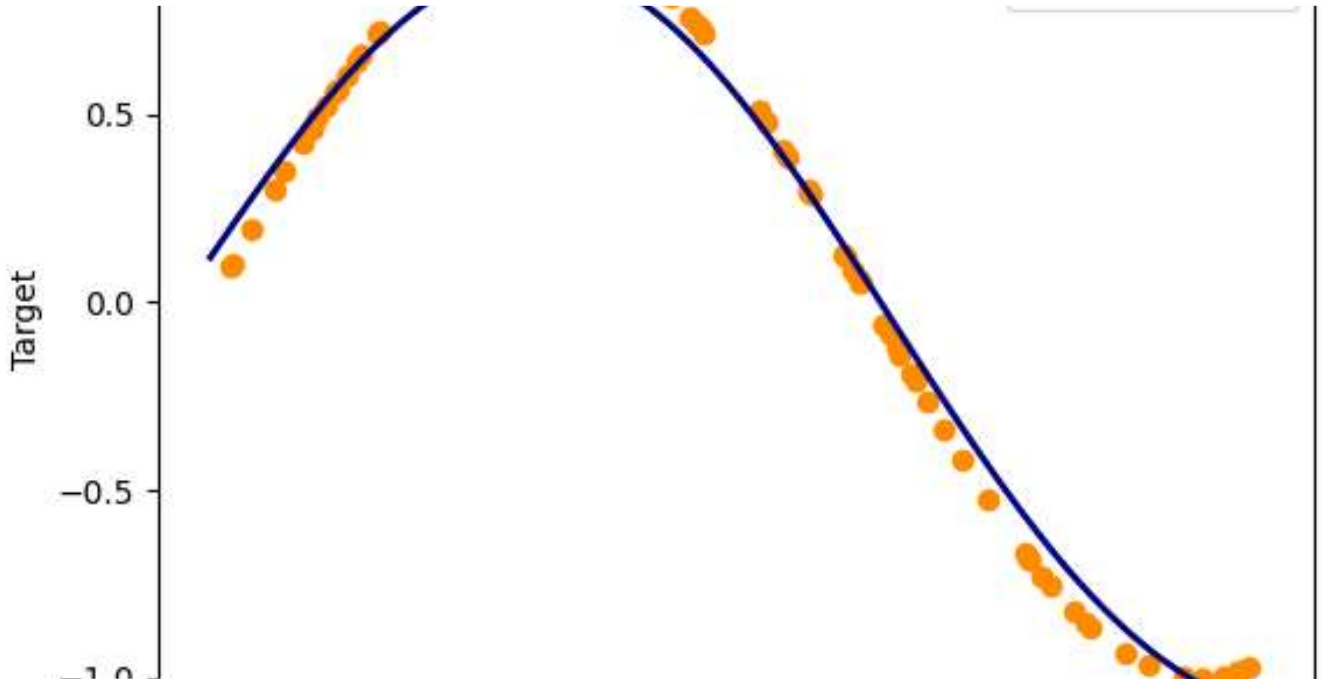
Hiraltalsaniya in Simform Engineering

Anomaly Detection with Unsupervised Machine Learning

Detecting Outliers and Unusual Data Patterns with Unsupervised Learning

9 min read · Dec 22, 2023

227



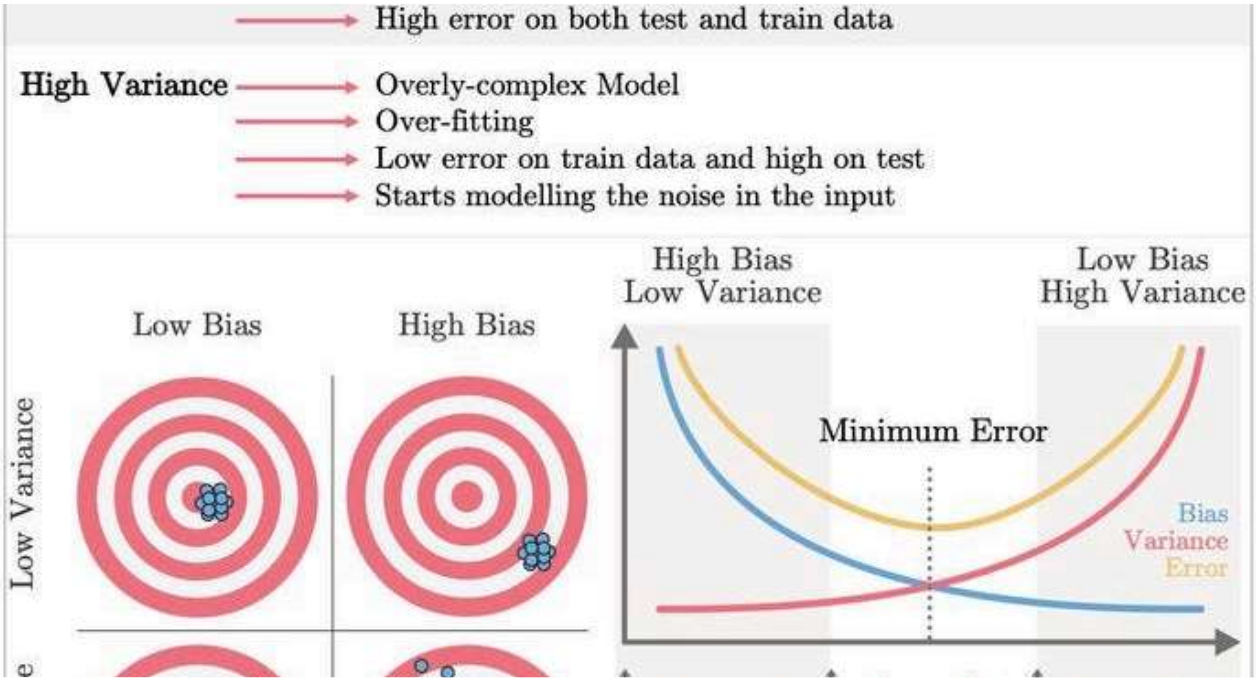
NANDINI VERMA

An Introduction to Support Vector Regression (SVR) in Machine Learning

Support Vector Regression (SVR) is a machine learning technique used for regression tasks. It is a variant of Support Vector Machines (SVM)...

3 min read · Nov 2, 2023

26



Sujatha Mudadla

Bias Variance tradeoff in Machine Learning.

The bias-variance tradeoff is a fundamental concept in machine learning that refers to the balance between two sources of error: bias and...

3 min read · Nov 28, 2023

108

See more recommendations