All Machine Learning Algorithms You Should Know in 2022

Intuitive explanations of the most popular machine learning models

Terence Shin Nov 4 · 7 min read ★

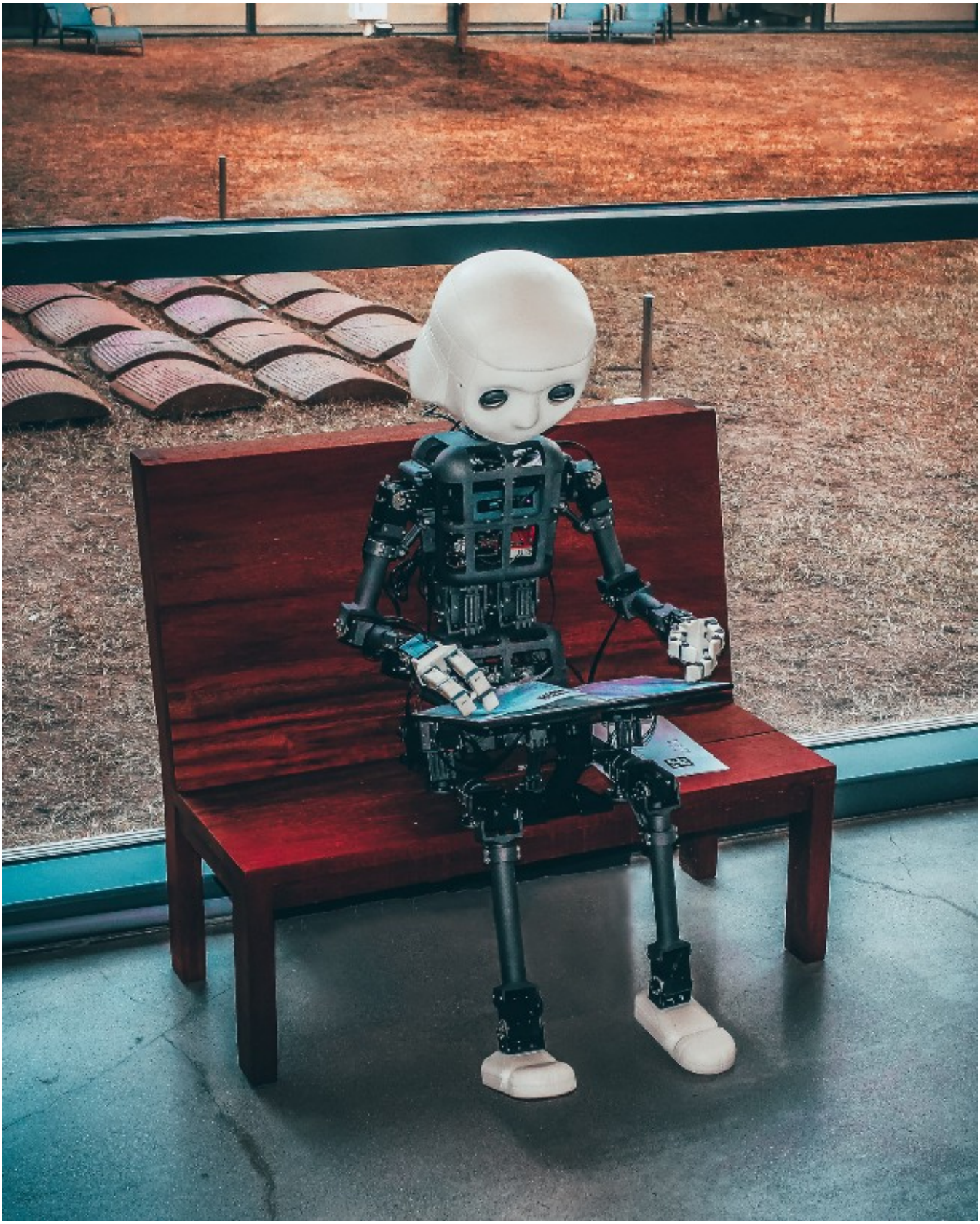


Photo by Andrea De Santis on Unsplash

> **Be sure to subscribe here or to my exclusive newsletter to never miss another article on data science guides, tricks and tips, life lessons, and more!**

Last year I wrote an article called *All Machine Learning Algorithms You Should Know in 2021*, so why am I writing another one for 2022? Are there that many new algorithms

that emerged in the past year?

Well, no.

But this year, I wanted to structure this article differently. Instead of listing every machine learning algorithm out there, I wanted to provide several **types** of machine learning models, and the most widely used models for each type.

Why am I doing this?

1. **Application**. Knowledge is effectively useless if it can't be applied to anything. By providing general categories of models, you'll have a better understanding of *what problems you can solve* rather than *what models are out there*.
2. **Relevancy**. The truth is that not all machine learning models are relevant anymore. You'll see immediately that traditional algorithms like Naive Bayes and SVMs are not included in this article, simply because they are outclassed by boosted algorithms.
3. **Digestibility**. I wanted to make this as easy as possible to digest. There are 1000s of resources online that can teach you how to implement the models that I'm going to talk about. And so, I'm going to focus more on WHEN to use each type of model.

With that said, let's dive into 5 of the most important types of machine learning models:

1. Ensemble learning algorithms
2. Explanatory Algorithms
3. Clustering Algorithms
4. Dimensionality Reduction Algorithms
5. Similarity Algorithms

> **Be sure to subscribe here or to my exclusive newsletter to never miss another article on data science guides, tricks and tips, life lessons, and more!**

1. Ensemble Learning Algorithms (Random ForestsXGBoost, LightGBM, CatBoost)

What are ensemble learning algorithms?

In order to understand what ensemble learning algorithms are, you first need to know what ensemble learning is. **Ensemble learning** is a method where multiple models are used at the same time to achieve better performance than a single model itself.

Conceptually, consider the following analogy:

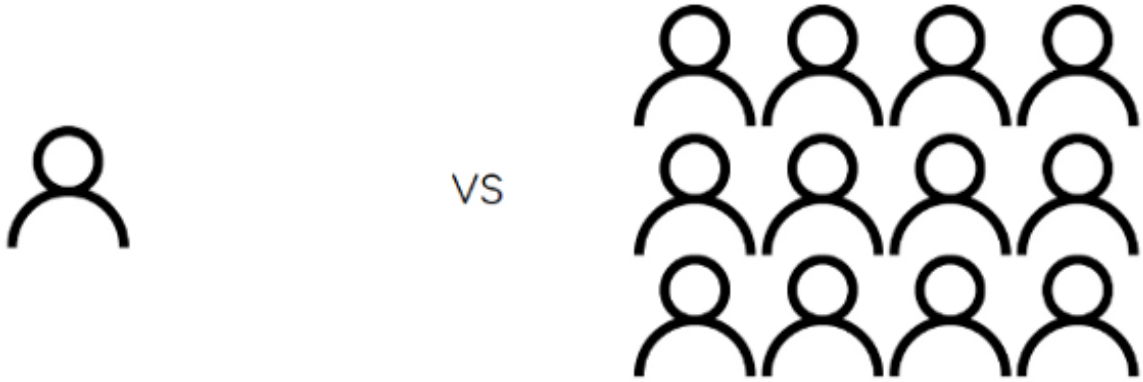


Image created by author

Imagine if one student had to solve a math problem versus an entire classroom. As a class, students can collaboratively solve the problem by checking each other's answers and unanimously decide on a single answer. On the other hand, the individual doesn't have this privilege — nobody else is there to validate his/her answer if it's wrong.

And so, the classroom with several students is similar to an ensemble learning algorithm with several smaller algorithms working together to formulate a final response.

If you want to learn more about ensemble learning, check out this article:

Ensemble Learning, Bagging, and Boosting Explained in 3 Minutes
Intuitive explanations and demystifying fundamental concepts
towardsdatascience.com

When are they useful?

Ensemble learning algorithms are most useful for regression and classification problems or supervised learning problems. Due to their inherent nature, they outclass all traditional machine learning algorithms like Naïve Bayes, support vector machines, and decision trees.

Algorithms

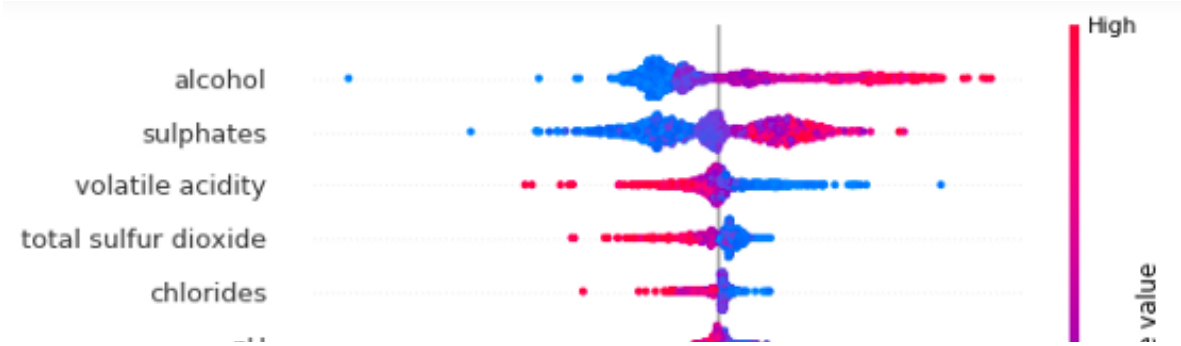
- Random Forests
- XGBoost
- LightGBM
- CatBoost

2. Explanatory Algorithms (Linear Regression, Logistic Regression, SHAP, LIME)

What are explanatory algorithms?

Explanatory algorithms allow us to identify and understand variables that have a statistically significant relationship with the outcome. So rather than creating a model to **predict** values of the response variable, we can create explanatory models to **understand** the relationships between the variables in the model.

From a regression standpoint, there's a lot of emphasis on **statistically significant** variables. Why? Almost always, you'll be working with a sample of data, which is a subset of the entire population. In order to make any conclusions about a population given a sample, it's important to ensure that there is enough **significance** to make a confident assumption.



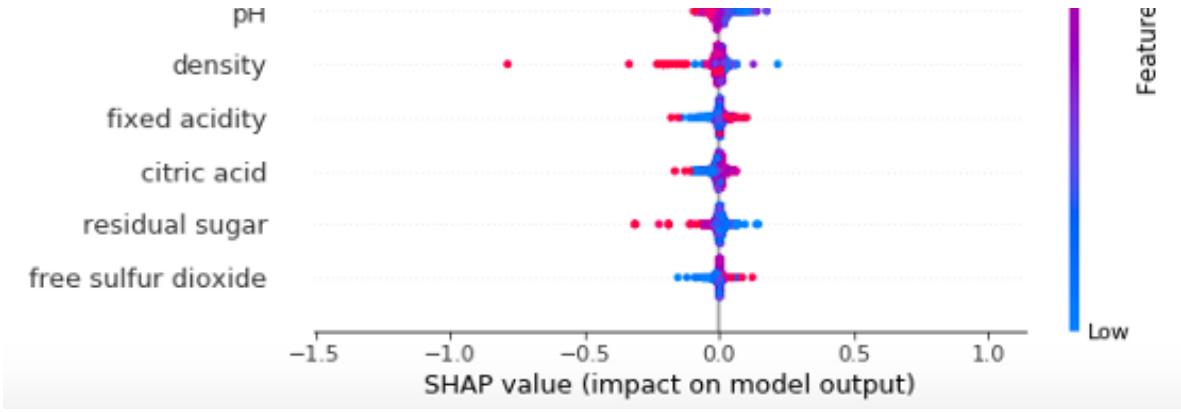


Image created by Author

Recently, there's also been the emergence of two popular techniques, SHAP and LIME, which are used to interpret machine learning models.

When are they useful?

Explanatory models are useful when you want to understand "why" a decision was made or when you want to understand "how" two or more variables are related to each other.

In practice, the ability to explain what you're machine learning model does is just as important as the performance of the machine learning model itself. If you can't explain *how* a model works, no one will trust it and no one will use it.

Algorithms

Traditional explanatory models based on hypothesis testing:

- Linear Regression
- Logistic Regression

Algorithms to explain machine learning models:

- SHAP
- LIME

> **Be sure to subscribe here or to my exclusive newsletter to never miss another article on data science guides, tricks and tips, life lessons, and more!**

3. Clustering Algorithms (k-Means, Hierarchical Clustering)

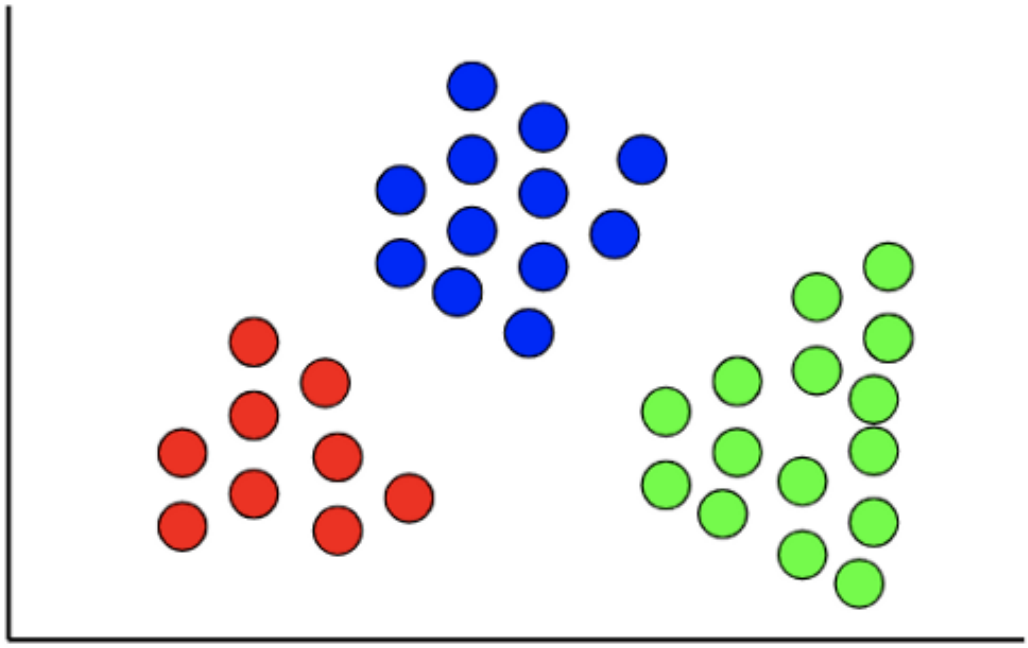


Image created by Author

What are clustering algorithms?

Clustering algorithms are used to conduct clustering analyses, which is an unsupervised learning task that involves grouping data into **clusters**. Unlike supervised learning where the target variable is known, there is no target variable in clustering analyses.

When are they useful?

Clustering is particularly useful when you want to discover natural patterns and trends in your data. It's very common for clustering analyses to be conducted in the EDA phase, to uncover more insights about the data.

Similarly, clustering allows you to identify different segments within a set of data based on different variables. One of the most common types of clustering segmentation is the segmentation of users/customers.

Algorithms

The two most common clustering algorithms are k-means clustering and hierarchical clustering, although many more exist:

- K-means clustering
- Hierarchical clustering

4. Dimensionality Reduction Algorithms (PCA, LDA)

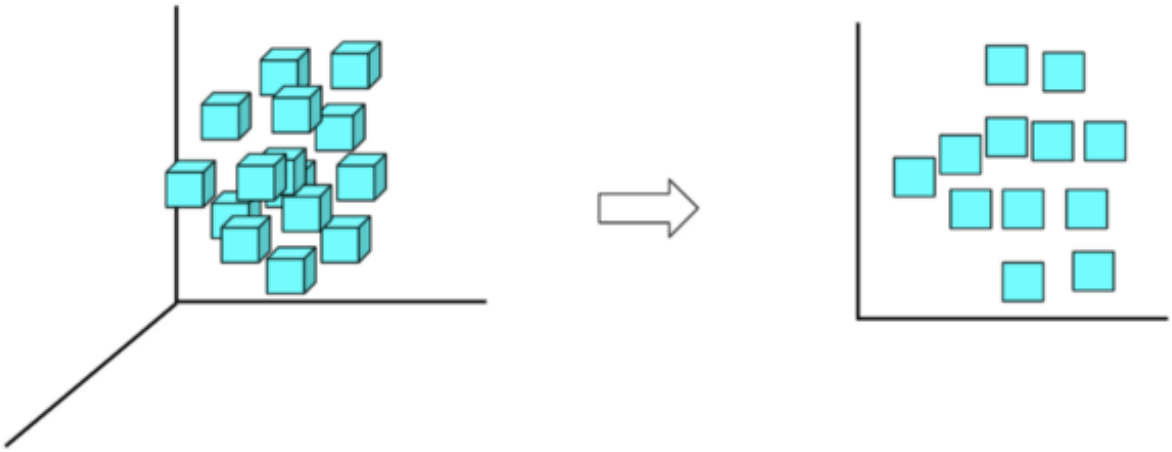


Image created by Author

What are dimensionality reduction algorithms?

Dimensionality reduction algorithms refer to techniques that reduce the number of input variables (or feature variables) in a dataset. Dimensionality reduction is essentially used to address the curse of dimensionality, a phenomenon that states, "as dimensionality (the number of input variables) increases, the volume of space grows exponentially resulting in sparse data.

When are they useful?

Dimensionality reduction techniques are useful in many cases:

1. They are extremely useful when you have hundreds, or even thousands, of features in a dataset and you need to select a handful.
2. They are useful when your ML models are overfitting the data, implying that you need to reduce the number of input features.

Algorithms

Below are the two most common dimensionality reduction algorithms:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

5. Similarity Algorithms (KNN, Euclidean Distance, Cosine, Levenshtein, Jaro-Winkler, SVD, etc...)

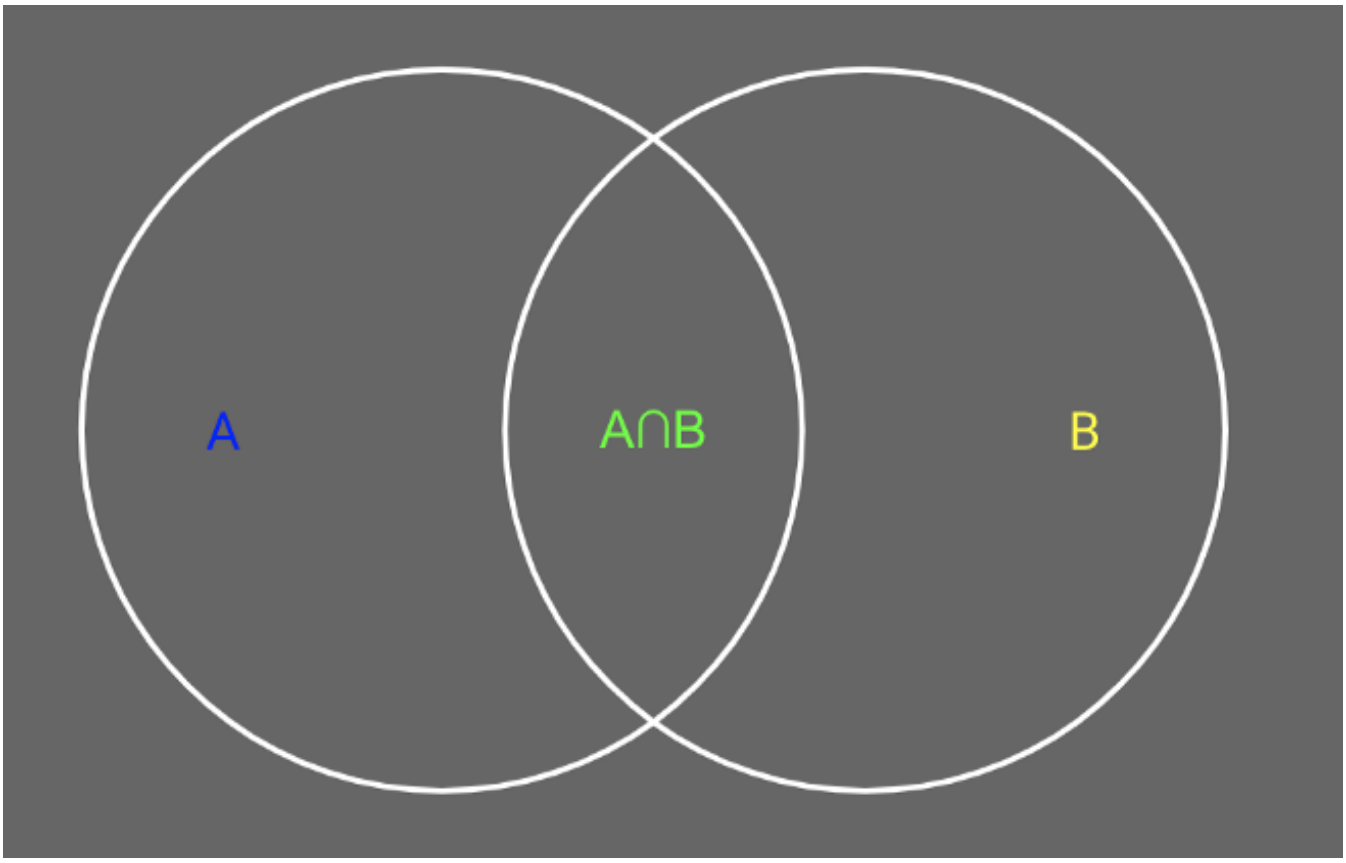


Image created by Author

What are similarity algorithms?

Similarity algorithms are those that compute the *similarity* of pairs of records/nodes/data points/text. There are similarity algorithms that compare the distance between two data points, like Euclidean distance, and there are also similarity algorithms that compute text similarity, like the Levenshtein Algorithm.

When are they useful?

Similarity algorithms can be used in a variety of applications, but they are particularly useful for **recommendation**.

- What articles should Medium recommend to you based on what you previously read?
- What ingredients can you use as a replacement for blueberries?
- What song should Spotify recommend based on what songs you've liked already?
- What products should Amazon recommend based on your order history?

These are just a few of the many examples where similarity algorithms and recommendation is used in our everyday lives.

Algorithms

Below is a non-exhaustive list of some similarity algorithms. If you want to read about more distance algorithms, check out this article. Likewise, if you want to read about more string similarity algorithms, check out this article.

- K nearest neighbors
- Euclidean Distance
- Cosine Similarity
- Levenshtein Algorithm
- Jaro-Winkler Algorithm
- Singular Value Decomposition (SVD) (not exactly a similarity algorithm, but indirectly relates to similarity)

Thanks for Reading!

> **If you enjoyed this, be sure to subscribe to never miss another article on data science guides, tricks and tips, life lessons, and more!**

After reading this, you should not only have a better idea of the various ML models out there, but you should also know **when** it's appropriate to use these models.

Now go out there and see what problems you can solve with ML!

As always, I wish you the best in your data science endeavors. If you liked this article, I'd appreciate it if you gave me a follow. :)

Not sure what to read next? I've picked another article for you:

The 10 Best Data Visualizations of 2021
Awesome visualizations on wealth distribution, the environment,

towardsdatascience.com

and another one:

10 Most Practical Data Science Skills You Should Know in 2022

Skills that will actually make you employable

towardsdatascience.com

Terence Shin

- *If you enjoyed this, SUBSCRIBE to my Medium for content!*
- *Likewise, you can also SUBSCRIBE to my exclusive newsletter*
- *Follow me on LinkedIn for other content*


Download on the
App Store


GET IT ON
Google Play