*k-NN Machine Learning Algorithm*

**Introduction**

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning method employed to tackle classification and regression problems.

While the KNN algorithm can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

KNN is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

**How does it work?**

K is the number of nearest neighbors to use. For classification, a majority vote is used to determine which class a new observation should fall into.

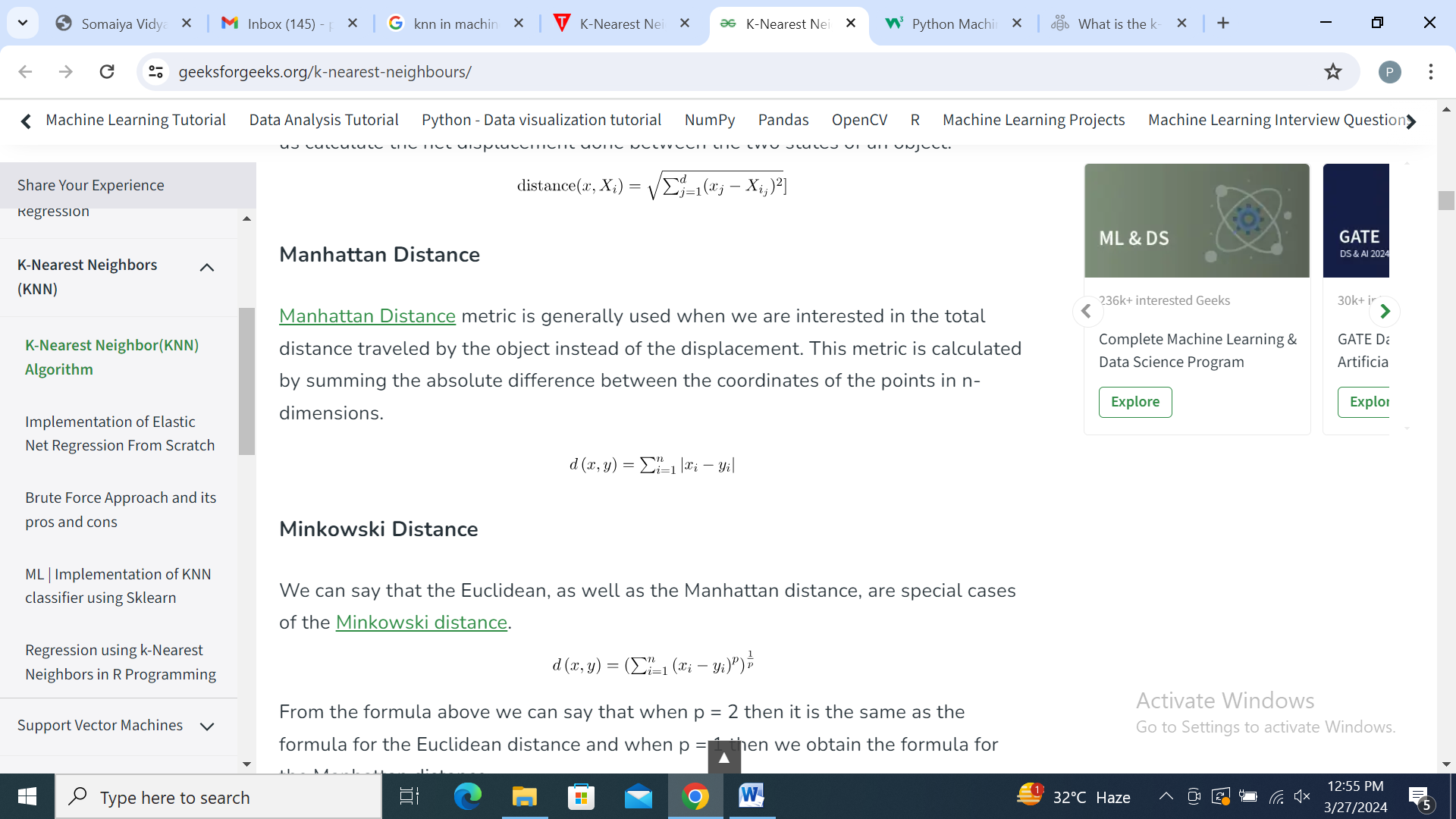
*Larger values of K are often more robust to outliers and produce more stable decision boundaries than very small values (K=3 would be better than K=1, which might produce undesirable results.*

**Distance Metrics Used in KNN Algorithm**

As we know that the KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. For this purpose, we use below distance metrics:

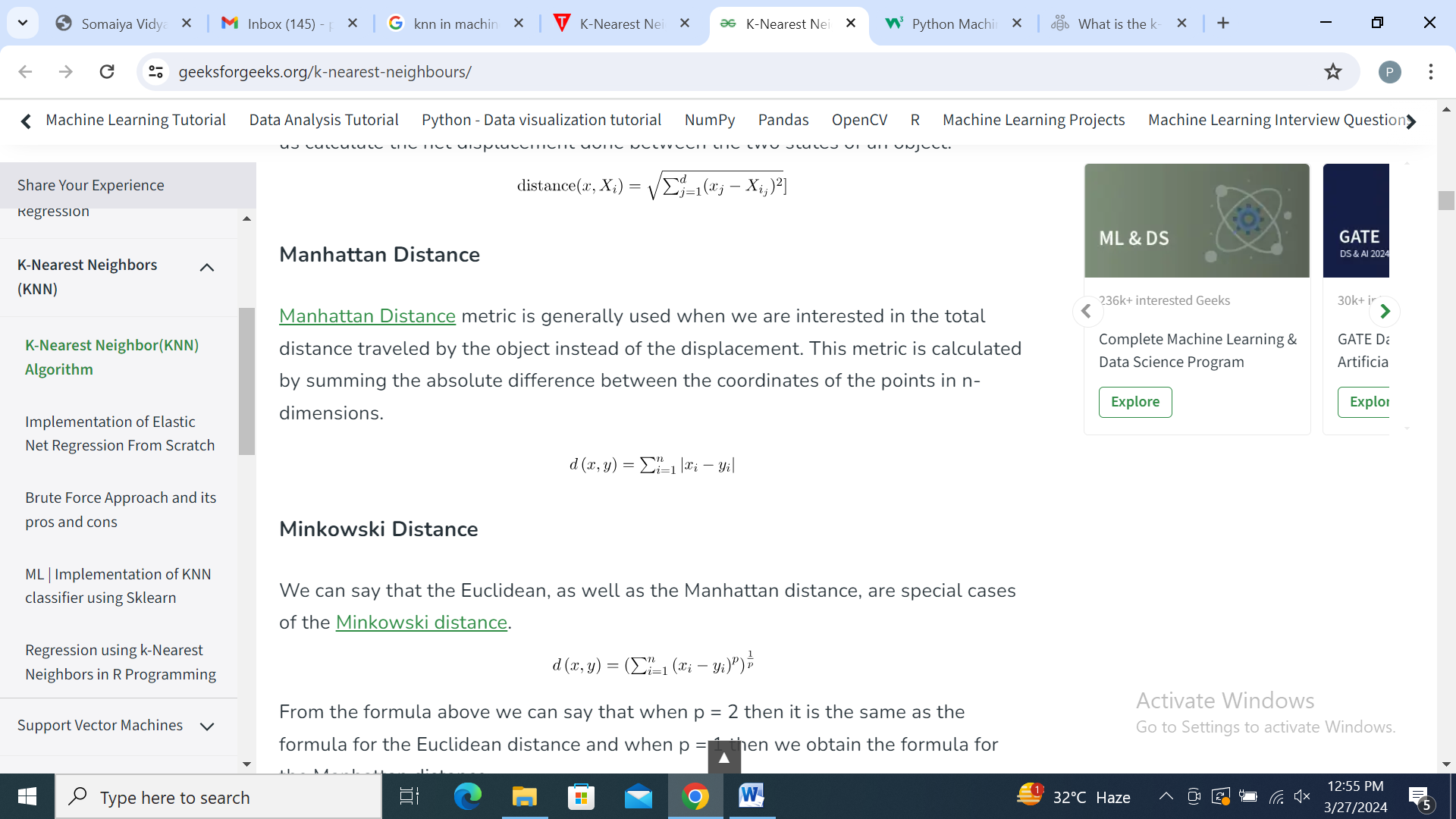
**Euclidean Distance**

This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean Distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.



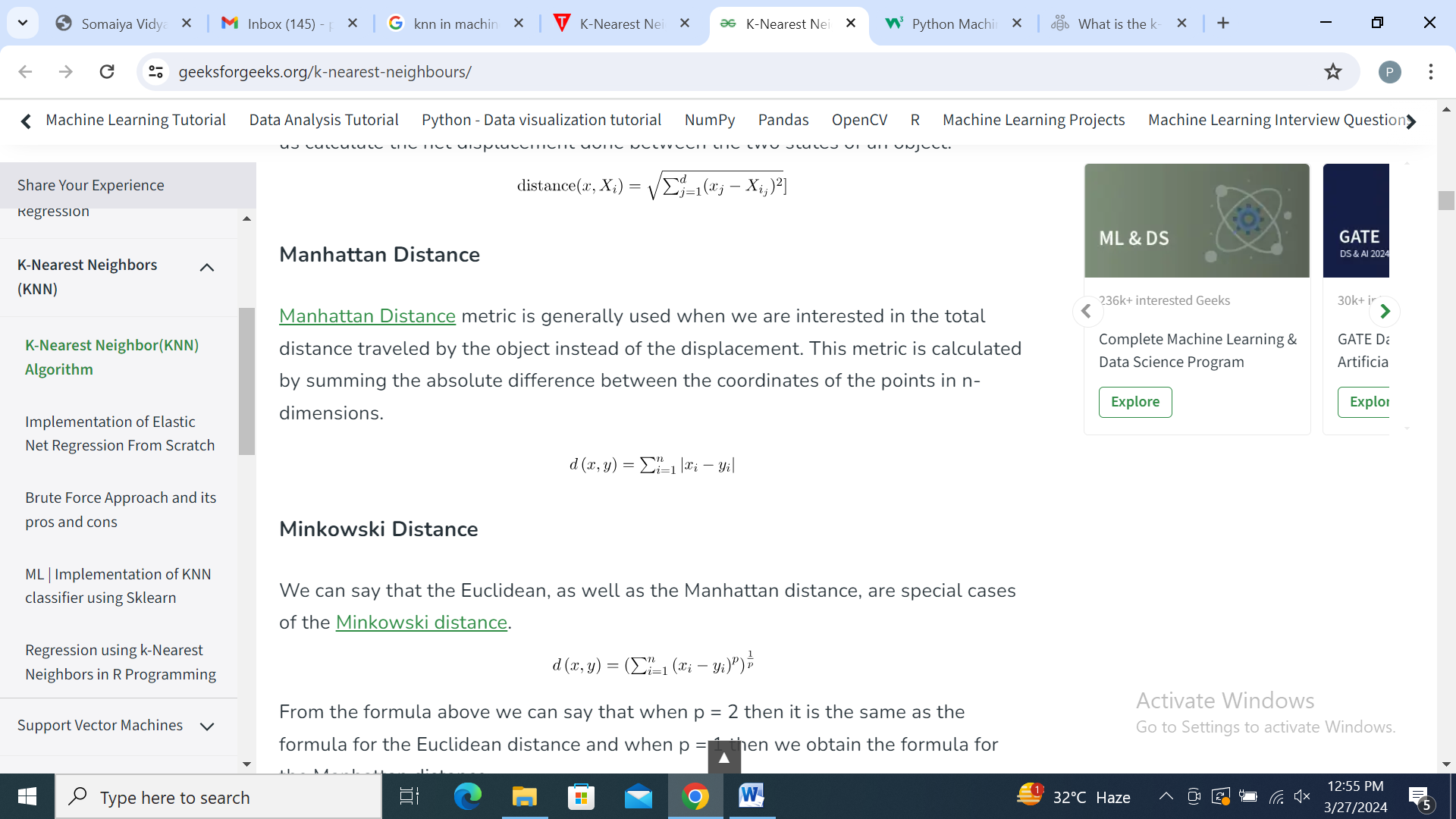
**Manhattan Distance**

Manhattan Distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions.



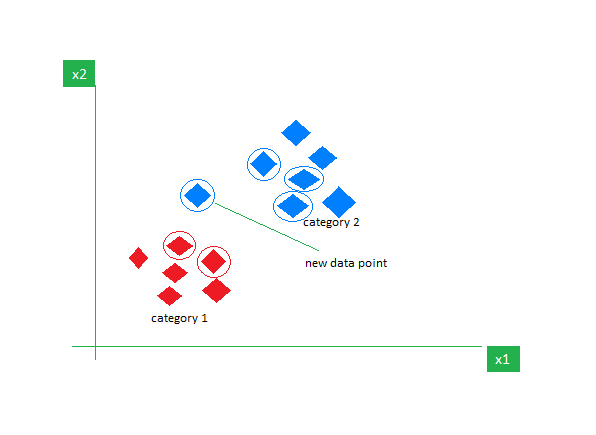
**Minkowski Distance**

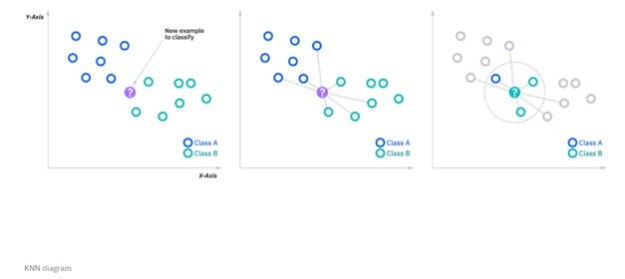
We can say that the Euclidean, as well as the Manhattan distance, are special cases of the  Minkowski Distance.



From the formula above we can say that when p = 2 then it is the same as the formula for the Euclidean distance and when p = 1 then we obtain the formula for the Manhattan distance.







**Applications of k-NN in machine learning**

The k-NN algorithm has been utilized within a variety of applications, largely within classification. Some of these use cases include:

- Data preprocessing: Datasets frequently have missing values, but the KNN algorithm can estimate for those values in a process known as missing data imputation.

- Recommendation Engines: Using clickstream data from websites, the KNN algorithm has been used to provide automatic recommendations to users on additional content. This research (link resides outside ibm.com) shows that the a user is assigned to a particular group, and based on that group’s user behavior, they are given a recommendation. However, given the scaling issues with KNN, this approach may not be optimal for larger datasets.

- Finance: It has also been used in a variety of finance and economic use cases. For example, one paper (link resides outside ibm.com) shows how using KNN on credit data can help banks assess risk of a loan to an organization or individual. It is used to determine the credit-worthiness of a loan applicant. Another journal (link resides outside ibm.com) highlights its use in stock market forecasting, currency exchange rates, trading futures, and money laundering analyses.

- Healthcare: KNN has also had application within the healthcare industry, making predictions on the risk of heart attacks and prostate cancer. The algorithm works by calculating the most likely gene expressions.

- Pattern Recognition: KNN has also assisted in identifying patterns, such as in text and digit classification (link resides outside ibm.com). This has been particularly helpful in identifying handwritten numbers that you might find on forms or mailing envelopes.