AdaBoost algorithm, is a Boosting technique used as an Ensemble Method in Machine Learning. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lowe error is received.

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Data Binning is a type of data pre-processing, a mechanism which includes also dealing with missing values, formatting, normalization and standardization.

Bitmap Indexing is a special type of database indexing that uses bitmaps. It is a special type of indexing built on a single key. This technique is used for huge databases, when column is of low number of elements and these columns are most frequently used in the query.

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

DMQL stand for Data Mining Query Language. It specifies clauses and syntaxes for performing different types of data mining tasks for example data classification, data clustering and mining association rules. Also it uses SQL-like syntaxes to mine databases.

The Gini index (or coefficient) is a synthetic indicator that captures the level of inequality for a given variable and population. It varies between 0 (perfect equality) and 1 (extreme inequality).

When the data is linearly separable, and we don’t want to have any misclassifications, we use SVM with a hard margin.

The hinge loss is a specific type of cost function that incorporates a margin or distance from the classification boundary into the cost calculation.

A Kernel Trick is a simple method where a Non Linear data is projected onto a higher dimension space so as to make it easier to classify the data where it could be linearly divided by a plane.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process

1. Partition objects into k non-empty subsets
2. Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., mean point, of the cluster)
3. Assign each object to the cluster with the nearest seed point
4. Go back to Step 2, stop when no more new assignment

K-NN algorithm is the simplest Machine Learning algorithms based on Supervised Learning technique. can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It classifies the data point on how its neighbour is classified. KNN classifies the new data points based on the similarity measure of the earlier stored data points.

1. Step-1: Select the number K of the neighbors
2. Step-2: Calculate the Euclidean distance of K number of neighbors
3. Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
4. Step-4: Among these k neighbors, count the number of the data points in each category.
5. Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
6. Step-6: Our model is ready.

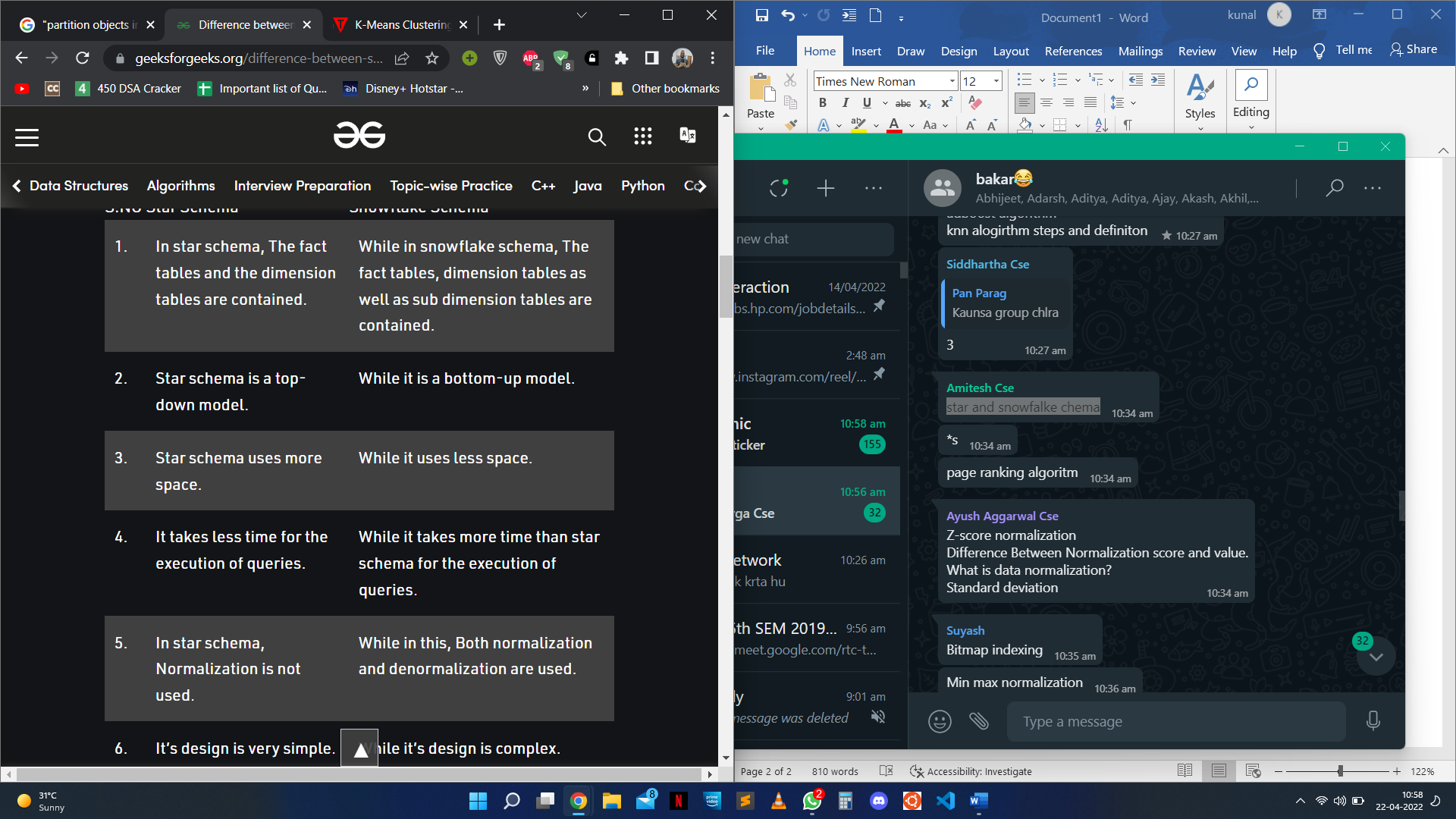
Normalization is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale.

Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.

Decimal Scaling Method normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data.

Z-score normalization In this technique, values are normalized based on mean and standard deviation of the data A. A high z -score means a very low probability of data above this z -score and a low z -score means a very low probability of data below this z -score. **z = (x-μ)/σ,** where x is the raw score, μ is the population mean, and σ is the population standard deviation.

Difference Between Normalization score and value. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane.

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

Web Log Mining, Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analysing these log files gives a neat idea about the user.

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. It uses automated methods to extract both structured and unstructured data from web pages, server logs and link structures.