

SUBJECT: Machine learning and Data Science

NAME : YASH DATTATRAYA DESAI

CLASS : DIV 3

SEMESTER/YEAR : 7TH (BE)

ROLL NO : C43414

EXPERIMENT NO- 05

TITLE: KNN

AIM: Text classification for Sentiment Analysis using KNN on Twitter Data.

SOFTWARES:

SOFTWARE	VERSION
Jupyter Notebook	V5.1

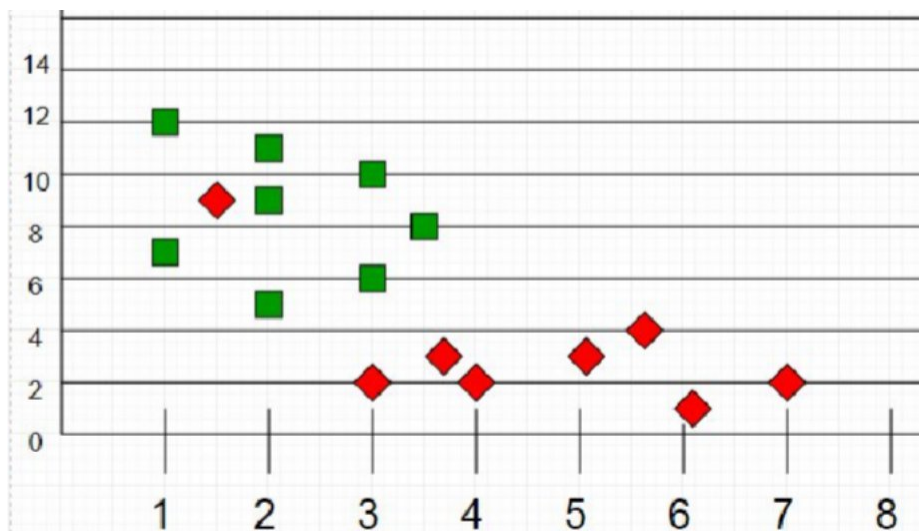
THEORY:

K-Nearest Neighbors 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.

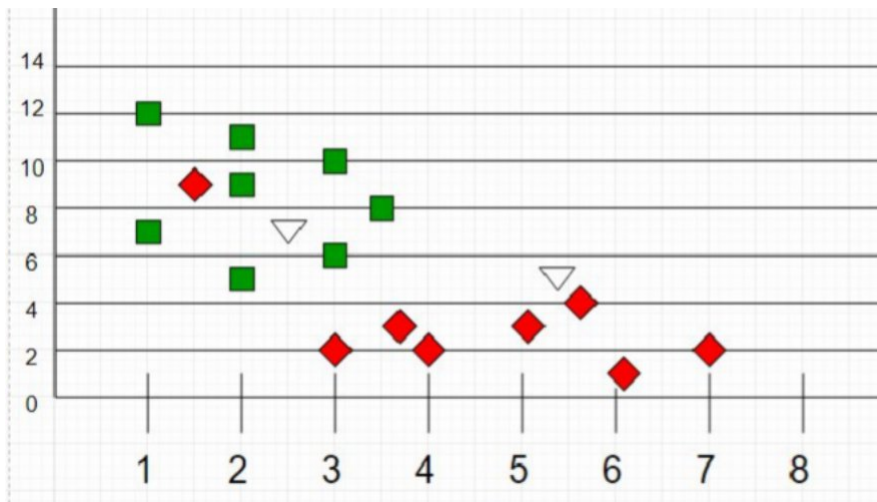
It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features:



Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as k-nearest-neighbours³.

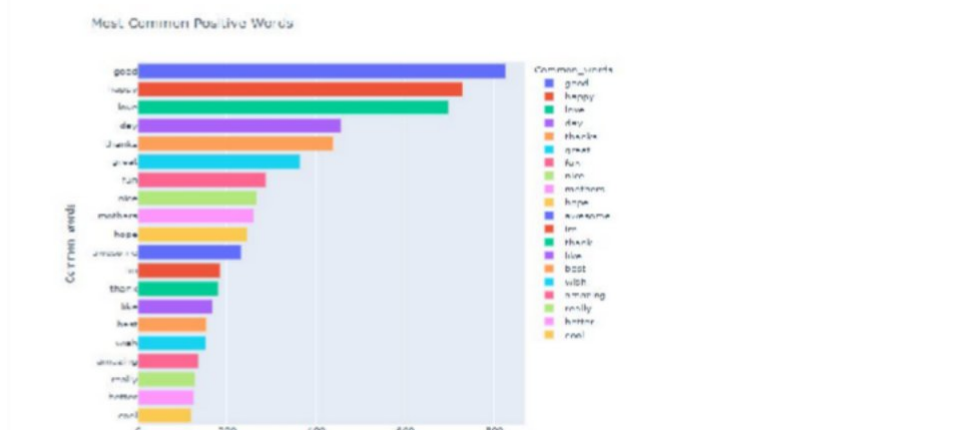


```
In [30]: fig = px.treemap(temp, path=['Common_words'], values='count', title='Tree of the most Common Words')
fig.show()
```

Tree of the most Common Words



```
In [31]: fig = px.bar(freq, x='Good', y='Common_words', title='Most Common Positive Words', orientation='h', width=700, height=700, n
nScore='Common_words')
fig.show()
```



CONCLUSION: We successfully classified the twitter data in various parts as positive twites, negative twites and neutral statements so that we come to know the thoughts of people.