K-Nearest Neighbor (KNN) Algorithm

A Report submitted for partial fulfillment of the requirements for the 3rd sem of

Masters of Computer Application

(MCA) of

JORHAT ENGINEERING COLLEGE

UNDER ASSAM SCIENCE AND TECHNOLOGY UNIVERSITY

Submitted by:

Madhurya Dutta (Roll Number: 210720043019)

MCA *3rd* Semester

Jorhat Engineering College, Assam

1. ABSTRACT:

Machine learning is a small part of artificial intelligence. Whatever we search on Google, Google takes our data and uses machine learning to show advertisement and search results accordingly. A machine learning system works on the principle that it has to take input data, learn something from it, and give output. This report aims to determine how the K-Nearest Neighbor (KNN) machine learning classification algorithm is applied to the model dataset and how the given data is predicted by the model to which class this given data will exist. K-Nearest Neighbor (KNN) is the simplest machine learning algorithm based on supervised learning. It is one of the top ten data mining algorithms, has been widely applied in various fields. The K-NN algorithm is mostly used in solving the classification problem.

2. INTRODUCTION:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. It is a distance-based algorithm.

3. IMPLEMENTATION:

(a) ALGORITHM:

The K-NN working can be explained on the basis of the below algorithm:

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

Some points to remember while selecting the value of K in the K-NN algorithm:

1.There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
2.A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
Large values for K are good, but it may find some difficulties.

(b) DATASET:

(i) Source: I have downloaded the dataset from the website called **Kaggle** . Kaggle is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges.

(ii) About:

This is a list of over 1497 consumer reviews for Amazon Musical Product Database. The dataset includes basic product information, rating, review text, and more for each product.

Ecommerce websites get vast amount of feedback from the users. To go through all the feedback's can be a tedious job. We have to categorize opinions expressed in feedback forums. We can perform Classification of individual reviews and we also determining overall rating based on individual reviews. So that company can get a complete idea on feedback's provided by customers and can take care on those particular fields.

(iii) Information:

This file has reviewer ID , User ID, Reviewer Name, Reviewer text, Summary(obtained from Reviewer text),Overall Rating on a scale 5, Review time

Description of columns in the file:

1. reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B

2. asin - ID of the product, e.g. 0000013714

3. reviewerName - name of the reviewer

4. reviewText - text of the review

5. summary - summary of the review

6. unixReviewTime - time of the review (unix time)

7. overall - rating of the product

(c) ANALYSIS:

(i) Display of Dataset:

	A	B	C	D	E	F	G
1	reviewerID	asin	reviewerName	reviewText	summary	unixReviewTi	overall
2	A2A039TZMZHH9Y	B00004Y2UT	Bill Lewey blewey	So good that I bought another one. Love the heavy cor	The Best Cable	1356048000	Best
3	A1UPZM995ZAH90	B00004Y2UT	Brian	I have used monster cables for years and with good rea	Monster Standard 100 - 21	1390089600	Best
4	AJNFQI3YR6XJ5	B00004Y2UT	Fender Guy Rick	I now use this cable to run from the output of my pedal	Didnt fit my 1996 Fender St	1353024000	Average
5	A3M1PLEYNDEYO8	B00004Y2UT	G. Thomas Tom	Perfect for my Epiphone Sheraton II. Monster cables ar	Great cable	1215302400	Best
6	AMNTZU1YQN1TH	B00004Y2UT	Kurt Robair	Monster makes the best cables and a lifetime warranty	Best Instrument Cables On	1389139200	Best
7	A2NYK9KWFMJV4Y	B00004Y2UT	Mike Tarrani Jazz Drum	Monster makes a wide array of cables including some tl	One of the best instrument	1334793600	Best
8	A35QFQI0M46LWO	B00005ML71	Christopher C	I got it to have it if I needed it. I have found that i dont r	It works great but I hardly	1398124800	Good
9	A2NIT6BKW11XJQ	B00005ML71	Jai	If you are not use to using a large sustaining pedal while	HAS TO GET USE TO THE SI	1384646400	Average
10	A1C0O09LOLVI39	B00005ML71	Michael	I love it I used this for my Yamaha ypt-230 and it works	awesome	1371340800	Best
11	A17SLR18TUMULM	B00005ML71	Straydogger	I bought this to use in my home studio to control my mi	It works!	1356912000	Best
12	A2PD27UKAD3Q00	B00005ML71	Wilhelmina Zeitgeist co	I bought this to use with my keyboard. I wasnt really aw	Definitely Not For The Seas	1376697600	Bad
13	AKSFZ4G1AXYFC	B000068NSX	C.E. Frank	This Fender cable is the perfect length for me! Sometim	Durable Instrument Cable	1376352000	Good
14	A67OJZLHBBUQ9	B000068NSX	Charles F. Marks charli	wanted it just on looks alone...It is a nice looking cord..	fender 18 ft. Cali clear...	1373328000	Best
15	A2EZWZ8MBEDOLN	B000068NSX	Charlo	Ive been using these cables for more than 4 months anc	So far so good. Will revisit	1363564800	Best
16	A1CL807EOUPVP1	B000068NSX	GunHawk	Fender cords look great and work just as well. By adding	Add California to the name	1375833600	Best
17	A1GMWTGXW682GB	B000068NSX	MetalFan	This is a cool looking cheap cable which works well. I be	Cheap and cool looking go	1331856000	Good
18	A2G12DY50U700V	B000068NSX	Ricky Shows	The Fender 18 Feet California Clear Instrument Cable - I	Fender 18 Feet California C	1390953600	Best
19	A3E0CF25A7LD2	B000068NSX	WBowie	Very good cable. Well made and it looks great with my (	Guitar Cable	1354924800	Good
20	A2W3CLAYZLDPTV	B000068NTU	Amazon Customer =Ch	Got this cable to run a rockband keyboard controller to	Quality cable!	1341446400	Best
21	A398X9POBHK69N	B000068NTU	Ann Vande Zande	When I was searching for MIDI cables for my ART X-15 I	I Got Great Pricing But Still	1383177600	Best
22	AXWB93VKVML6K	B000068NTU	Michael Hassey	Cant go wrong. Great quality on a budget price - Hosa i	Its a Hosa	1372809600	Good
23	A2FZ4Z0UFA1OR8	B000068NTU	Pat	The ends of the midi cable look and feel like quality. Co	Quality and Secure	1327449600	Best
24	AXP9CF1UTFRSU	B000068NTU	tada	Just trying to find a midi to midi was a task and you have	Midi to Midi	1381795200	Best
25	A2CCGGDGZ694CT	B000068NVI	b carney	The Hosa XLR cables are affordable and very heavily ma	Very Heavy Cables At Affor	1341964800	Good
26	A27DR1VO079F1V	B000068NVI	Dan Edman	I bought these to go from my board to the amp. We use	Still going	1392768000	Best
27	A1LQC225SE8UNI	B000068NVI	David Burch	Sturdy cord and plugs inexpensive good value. I dont rec	Does what its supposed to	1337990400	Best
28	AU9BPT3Y3K6J4	B000068NVI	G. L. Beebe	Use it every week at gigs. Solid no problems with the sc	Good cable	1376092800	Best
29	A14Z9LAETO21KL	B000068NVI	Gutjammer	Hosa products are a good bang for the buck. I havent lo	Good Enough	1394496000	Good
30	A2074KEJGRYJV4	B000068NVI	hcross	This was exactly what I was after. I have a voice touch	Great little cord	1379289600	Best

(ii) CLASS LABEL INFORMATION:

Label - Overall Rating

Best – 5 Star

Good – 4 Star

Average – 3 Star

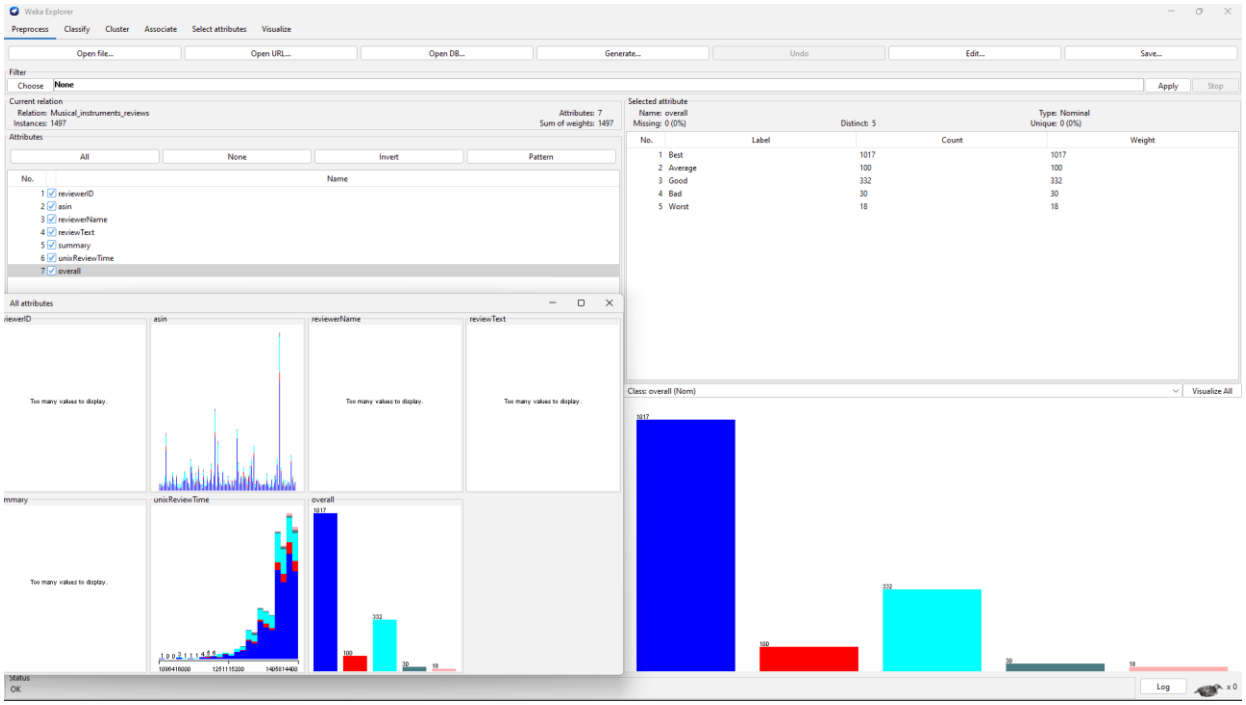
Bad – 2 Star

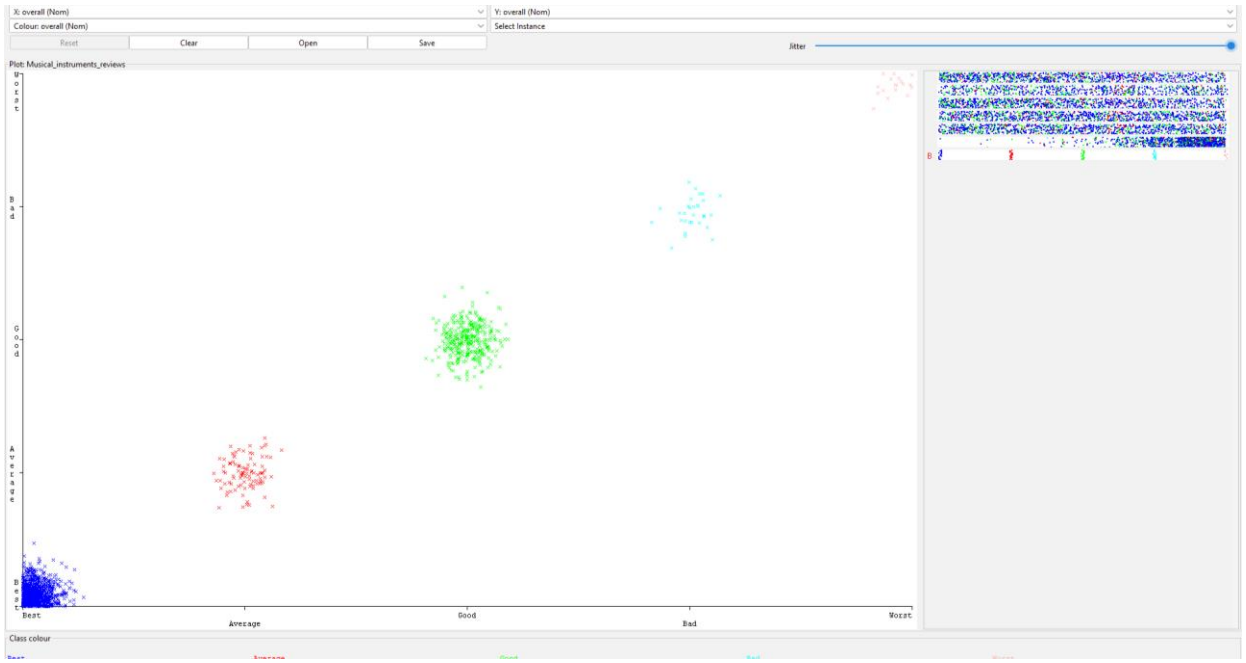
Worst – 1 Star

(iii) PREDICTION STATEMENT: classify the customer reviews. This would be helpful for the organization to understand Customer feedbacks.

(d) WEKA:

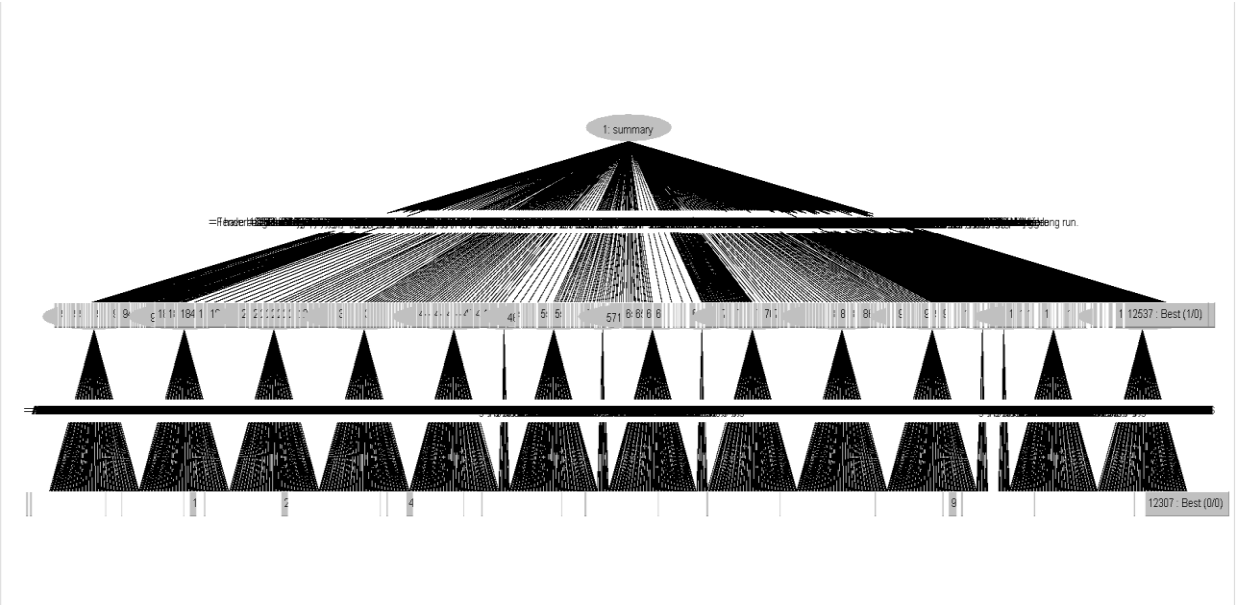
(i) ANALYSIS:



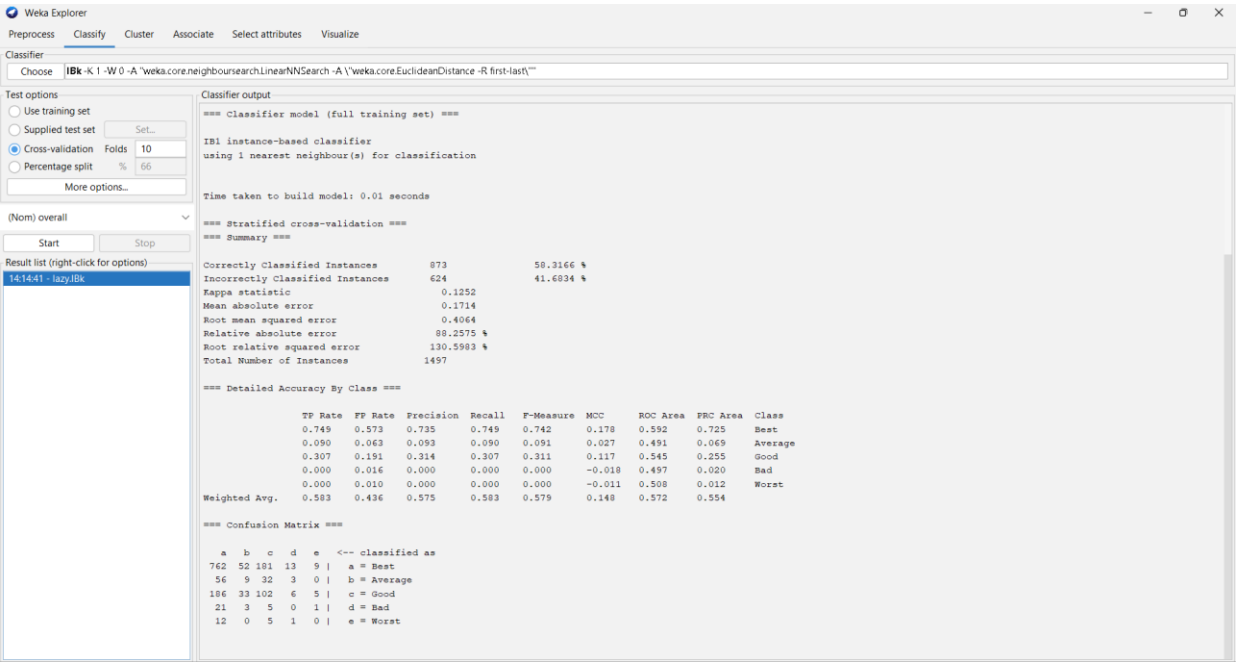


(ii)VISUALIZATION:

- **TREE STRUCTURE**



- **PREDICTION**



=== Summary ===

Correctly Classified Instances	873	58.3166 %
Incorrectly Classified Instances	624	41.6834 %
Kappa statistic	0.1252	
Mean absolute error	0.1714	
Root mean squared error	0.4064	
Relative absolute error	88.2575 %	
Root relative squared error	130.5983 %	
Total Number of Instances	1497	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.749	0.573	0.735	0.749	0.742	0.178	0.592	0.725	Best
	0.090	0.063	0.093	0.090	0.091	0.027	0.491	0.069	Average
	0.307	0.191	0.314	0.307	0.311	0.117	0.545	0.255	Good
	0.000	0.016	0.000	0.000	0.000	-0.018	0.497	0.020	Bad
	0.000	0.010	0.000	0.000	0.000	-0.011	0.508	0.012	Worst
Weighted Avg.	0.583	0.436	0.575	0.583	0.579	0.148	0.572	0.554	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
762	52	181	13	9	a = Best
56	9	32	3	0	b = Average
186	33	102	6	5	c = Good
21	3	5	0	1	d = Bad
12	0	5	1	0	e = Worst

- **Error Visualization**

(iii)COMPARISON OF KNN & RANDOM TREE

Parameters	KNN	Random Tree
Time taken to build model:	0 seconds	0.09 seconds
Correctly Classified Instances	873 (58.3166 %)	1010 (67.4683 %)
Incorrectly Classified Instances	624 (41.6834 %)	487(32.5317 %)
Kappa statistic	0.1252	0.0452
Mean absolute error	0.1714	0.1857
Root mean squared error	0.4064	0.3213
Relative absolute error	88.2575 %	95.6544 %
Root relative squared error	130.5983 %	103.2327 %
Total Number of Instances	1497	1497
Size of the tree:	No Information	12537

From the comparison, we can conclude that Random Tree is more efficient than KNN for our dataset. Because the number of correctly classifies instances in Random Tree is higher than the instances in KNN. So we can say that using Random Tree algorithm will provide more accurate result from our dataset

(e) CONCLUSION:

KNN is an effective machine learning algorithm that can be used in credit scoring, prediction of cancer cells, image recognition, and many other applications. The main importance of using KNN is that it's easy to implement and works well with small datasets.

However, KNN also has disadvantages. Specifically, it doesn't work well with large datasets because for every test data, distance between all training data points and the test data in question is computed, resulting in large space and a long-required timeframe.